The Mathematical Structure of the Endolinguistic Code Space

Topology, Order, and Directed Dynamics of OAS Consonantal Codes across 4,721 Languages

Alejandro Toledo Martínez

Symbol key. The seven OAS classes are written both as Latin letters and as their presentation symbols: P = Φ, T = Θ, K = X, S = Σ, L = Λ, M = **M**, N = Ξ. Figures use the Latin letters; the text uses both. A *nucleus* is written `{X,Y}` (unordered); an *ordered code* `X·Y·Z`; a *skeleton* `(x·y)` in real consonants.

Abstract

Endolinguistics analyses meaning through *codes*: small unordered sets of two or three affective consonant classes (the OAS system — *Originating Affective Sound*: P T K S L M N) posited to orient a field of meaning across languages, invariant under the surface phonetics of any one tongue. Prior work reads individual codes. Here we instead ask what mathematical structure the *space* of realized codes possesses, treating it as a single formal object over a large multilingual corpus (1,110,197 coded word-forms in 4,721 languages, derived from lexibank-analysed and NorthEuraLex with IPA/CLTS segmentation). Four lenses are applied — algebraic topology, association/graph theory, category and group actions, and Markov/order theory — each against a pre-declared null model. We report five principal results. (1) The attested code space is the *complete, saturated* 2-skeleton of the 6-simplex on seven vertices (all 21 binary, all 35 ternary, and all 64 higher class-sets occur), with global homology $(b_0,b_1,b_2)=(1,0,20)$ and Euler characteristic $\chi=21$: human language realizes *every* OAS combination, yet the per-language attested set fails the simplicial closure axiom in 66.8 % of languages — a ternary occurs without one of its binary faces (a phonotactic null shows this failure is *sampling sparsity*, not irreducibility — §10.5). (2) Binary combination obeys an OCP-style same-zone repulsion law (within-articulatory-zone association −0.107 vs across-zone +0.311, permutation $p=0.001$), invisible in raw co-occurrence (modularity $Q=0$) and surfacing only after dividing out the velar hub K=X — but a consonant-tier null reproduces it entirely, so it is *recovered* long-distance OCP-Place phonology, not a novel law (§10.5). (3) Order is semantically load-bearing: inversion ι is an involution with 126 orbits, and a ternary code shares on average only 24 % of its concepts with its reverse (mean Jaccard 0.243), empirically confirming the axiom $q(C)\neq q(\iota C)$; moreover the rank–frequency law becomes Zipfian only once order is imposed (unordered slope −0.87, Gini 0.44 → ordered −1.18, Gini 0.65). (4) Class substitution follows a *perfect total order* — L > S > K > N > P > M > T — in which 100 % of net flow runs downhill and all 21 pairwise net-flows are strictly positive: a single, cycle-free order (but a cognacy control shows 83 % of these comparisons are non-cognate lexical replacement, so this is a *synchronic* modal-nucleus pattern, not established diachronic "decay" — §10.5). (5) Three independent tables converge on a Λ-pivot / X-shell law: Λ is the structural centre (the only class preferring internal insertion, central in 73 % of its nuclei's canonical orders), while X is always peripheral. We close with a category-theoretic sketch, and with a set of new hypotheses, open questions, and falsifiable predictions for a continuing research programme. One prior intuition is falsified: deverbal derivation *expands* codes far more than it reduces them (reduction : expansion = 0.091). Under phonologically-informed nulls and a cognacy control (§10.5), none of these five is a clean endolinguistic signal: the OCP is recovered phonology, the closure failure is sampling sparsity, the substitution order is a synchronic (83 %-non-cognate) pattern, and cross-macrosystem resonance (2.15× chance) is real but still

confounded by universal sound-symbolism and areal diffusion. We therefore present this as a foundational validation of the OAS instrument — the reduction faithfully re-expresses known phonological and frequency structure — with the semantic discriminant that would isolate the endolinguistic object itself (code→field, H-0) specified but not run. In the vocabulary of the companion essay we call a genuine-descent match a coderivative (a *cognate* named from the code it conserves), and the descent (23 %) / resonance (9.3 %) / confound split reads as three registers of sense — *found*, *constructed*, *projected* — only the last of which is apophenia.

1. Introduction

Status and scope of this paper. This is a *foundational validation of the OAS instrument*, not a proof of the central semantic hypothesis of endolinguistics. It asks what statistical and phonological structure emerges when thousands of languages are projected onto the seven OAS classes, shows that the reduction is *representationally faithful* (it recovers known phonological universals), and defines the controls (a phonologically-informed null, chance-collision baselines, cognacy control) needed to later isolate what structure cannot be explained by ordinary phonology. That residue — and only it — is the endolinguistic object. Accordingly we keep four kinds of claim strictly apart: (1) necessary combinatorial results, (2) recovered phonological universals, (3) potentially novel statistical findings, and (4) endolinguistic interpretation. This study presupposes the distinctions developed in prior psychophonological and endolinguistic work: phonetic or phonological measurements are not treated as identical to psychic qualities; they are the material conditions from which a separate psychophonological interpretation may later be formulated. Update: the discriminant controls this paper specifies (a per-language phonotactic null, chance baselines for resonance and inversion, and a cognacy control) have now been run; their results — two claims survive, two are demoted — are in §10.5.

1.1 Endolinguistics and the OAS code

Endolinguistics — the discipline founded by C. S. Meulemans and J. Á. Elias — proposes that beneath the lexicon of any language runs a deeper, macrosystemic layer in which meaning is *oriented* (not denoted) by the consonantal skeleton of a word. The present author's original contribution to that tradition is the OAS system (Originating Affective Sound): the discretization of the consonantal space into seven affective sound classes, P T K S L M N, each an *originating affective sound* — a pre-symbolic felt sound-direction rather than a meaning, physically anchored in (but not defined by) a region of articulation. A *code* is a nucleus: an unordered set of two (binary) or three (ternary) classes. A code does not mean; it *orients a field of meaning* that recurs, in different phonetic clothing, across unrelated languages. The consonants of a word are *realizations* of its classes, and a "language" is treated as a rhythmic variant of a macrosystem rather than as a repository of cognates.

Two further constructs of the system matter here. First, an ordered code (an *isomer*, $\mathtt{X \cdot Y \cdot Z}$) is a permutation of a nucleus; the axiom $q(C) \neq q(\iota C)$ states that reordering — in particular reversal (inversion ι) — reorients the quality, so that a code and its mirror are *not* semantically equivalent. Second, codes are related by operations already realized in real words — we *find* them, we do not compute them: insertion $\oplus$X (a class wedged into a binary to make a ternary), reduction $\ominus$ (the inverse), substitution X→Y (one class replacing another), and permutation. Four or more distinct classes are not a single code but a composite (an agglutinative or compound form) that must be studied by separation into morphemes.

1.2 From reading codes to measuring the code space

Endolinguistic practice reads codes one at a time. This paper asks a different, structural question: *taken together, what shape does the realized code space have?* We treat the seven classes, the 56 nuclei, the 252 ordered codes, and the operations among them as a single mathematical object, and we interrogate its topology, its combinatorial association structure, its operation algebra, and its directed dynamics against a corpus large enough to make such questions empirical rather than speculative.

Two levels of validity must be kept apart. Representational validity asks whether the OAS reduction destroys phonological structure; we show it does not (it recovers the OCP, sonority sequencing, and positional fortition). Specific endolinguistic validity asks whether the classification explains *semantic* organization that does not follow from phonotactics, frequency, genealogy, or syllable form. The present analyses establish the first and *design the test* for the second; they do not yet settle it. Recovering a known phonological universal against a naïve null is therefore evidence that the instrument is faithful — not, by itself, evidence for endolinguistics. What we ultimately seek is the residue that survives a phonologically-informed null and a cognacy control; the discriminant experiment that isolates it is stated as the programme's next step (§10.4).

The corpus is the endolinguistic master table `lexeme` (§2): 1,110,197 word-forms carrying a binary or ternary code, drawn from 4,721 languages with phonetic (IPA/CLTS) segmentation, together with derived tables for diachronic conservation (844,944 rows), insertion events (818,279), and deverbal derivation (8,692). Each analysis is paired with an explicit null model, and results are reported whether they confirm or *falsify* endolinguistic intuitions — the discipline's own falsifiability constraints demand it.

1.3 Contributions

1. The attested code space is the complete, saturated 2-skeleton (indeed the full 7-simplex up to

dimension six); its global homology is trivial in degree one; per-language closure fails, but a phonotactic null shows this is *sampling sparsity*, not irreducibility (§3, §10.5).

1. A quantitative same-zone repulsion (OCP) law on class combination, hidden behind frequency — shown by a

consonant-tier null to be *recovered* long-distance OCP-Place phonology (§4, §10.5).

1. Order is semantically and statistically load-bearing: $q(C)\neq q(\iota C)$ at scale, and Zipf-only-under-order

(§5, §8).

1. A cycle-free total order of modal-nucleus substitution with Λ as source — *synchronic* (83 % of compared

pairs non-cognate), not established as diachronic ”change” (§6, §10.5).

1. The Λ-pivot / X-shell law and the resolution of the Λ paradox (§7).

A running series of boxed Hypotheses (H), Open Questions (Q), and Insights (I) turns the findings into a forward research programme (collected in §9).

2. Data and Methods

2.1 Corpus and coding pipeline

The base corpus merges lexibank-analysed (thousands of languages with CLTS-normalized IPA) and NorthEuraLex (the curated noun atlas). Each word-form is passed through the OAS pipeline: the phonetic segments are mapped to OAS classes by a fixed articulatory map; terminative and inflectional affixes are stripped to expose the root (constraint C0⋆, *root primacy*); adjacent duplicate classes are collapsed; and the resulting class multiset is reduced to its nucleus (the set of distinct classes), its arity (number of distinct classes), and — retaining order — its ordered code. Forms with 2 or 3 distinct classes receive a code; forms with 4+ are marked composite (`nucleus=''`, `arity`=real count) and set aside as agglutinative. This yields the `lexeme` table (fields: ‘concept, language, family, form, skeleton, nucleus, ordered, arity‘).

2.2 The formal objects

Let Ω = P, T, K, S, L, M, N be the class alphabet. Although each OAS class is an affective sound (not an articulatory category by definition, §1.1), the classes have a stable *physical* correlate — a region of articulation — and we use that correlate purely as an analytical partition of Ω into labial P, M, coronal T, S, L, N, and velar K (used only for the zone test of §4.2). The combinatorial code space is fixed:

object	definition	count
classes (0-cells)	elements of Ω	7
binary nuclei (1-cells)	2-subsets of Ω	C(7,2) = 21
ternary nuclei (2-cells)	3-subsets of Ω	C(7,3) = 35
nuclei total	2- and 3-subsets	56
ordered binary codes	orderings of binaries	21·2 = 42
ordered ternary codes	orderings of ternaries	35·6 = 210
ordered codes total		252
skeletons	consonant realizations	4,588

These are stored in the `catalog_*` tables and confirmed exactly against the corpus.

> The formalism is not a compositional semantics. That a code is a *set* (or ordered set) of classes must not be read as a recipe for its meaning by summing per-class glosses — assigning each class a fixed sense and adding them into the word. A class does not denote; the code *orients* a field, read from the relations among its coderivatives (cognates), *bonded*, not assembled piece by piece. Summing classes dissolves that bonding and is, precisely, compositional apophenia — a misreading language models make routinely. The algebra of §2.3 onward describes how codes *relate*; it never computes what a code *means*. (Full treatment, with the *star* example, in the companion paper *What OAS Is and Is Not*, §5.)

2.3 Operations

- Inversion ι reverses an ordered code; $\iota^2 = \mathrm{id}$.
- Insertion ⊕X sends a binary to a ternary by adding class X at a position (initial / internal /

final); reduction ⊖ deletes a class.

- Substitution X→Y (recorded diachronically as a `delta` ”X/Y”).
- Permutation: the symmetric group S_2 acts on binary orderings, S_3 on ternary orderings.

2.4 Null models and caveats

Association is measured as $\log_2$(observed / expected), expectation taken from class marginals (an independence model). Isomer distributions are tested against the uniform S_3 law (χ^2, df=5). Diachronic conservation is measured against an anchor = the modal nucleus among a fixed set of ancient Indo-European languages (Latin, Ancient Greek, Sanskrit, ...); consequently, *within* the Indo-Uralic macrosystem conservation is genuine descent, whereas *across* macrosystems a match reads as resonance, not inheritance (§8). Throughout we honour the endolinguistic caveat that attestation is exposure, not frequency-truth: a code realized by few words is a *hint*, not a rare event to be over-interpreted. The verb layer is coded "approximately" (per-family stripping) and is used only where noted.

On the nulls. The independence and uniform-S_3 nulls used below are *deliberately naïve*: they are the weakest reasonable baselines, and beating them shows only that the OAS reduction preserves structure a random model lacks. They are not sufficient to separate an endolinguistic signal from ordinary phonology. The phonologically-informed null that would do so — pseudo-lexicons preserving per-language phonotactics, segment frequency and syllable structure — is specified as the programme's central next experiment (§10.4); results here are marked by claim type (1–4, §1) accordingly. On the "diachrony". The conservation and substitution figures (§6, §8.2) compare the modal nucleus of an ancient Indo-European anchor against another language for the same concept. Without a cognacy control, a nucleus mismatch is not a class substitution: it may be a non-cognate lexical replacement, a loan, or chance (Latin *canis* vs Spanish *perro* is another root, not a X→ change). We therefore read these as synchronic modal-nucleus coincidences, and treat genuine descent as requiring the cognacy control of §10.4.

3. The Code Space as a Simplicial Complex

3.1 Construction

Give the code space the natural abstract simplicial complex Δ: the classes are vertices, the attested binary nuclei are edges (1-simplices), the attested ternary nuclei are triangles (2-simplices), and attested composites of size k are (k−1)-simplices. The *possible* 2-skeleton is that of the 6-simplex on seven vertices: 21 edges and 35 triangles. The question is which faces are realized in language, and what the realized complex looks like.

3.2 Global result: the complex is complete and saturated

At every reasonable threshold the attested complex is the entire 2-skeleton — and, extending to composites, the entire 6-dimensional simplex:

attestation threshold	edges	triangles	4–7-class sets	missing faces
$\geq$ 1 form	21/21	35/35	64/64	none
$\geq$ 100 forms	21/21	35/35	—	none
$\geq$ 1000 forms	21/21	35/35	—	none
$\geq$ 20 languages	21/21	35/35	—	none

The rarest edge, `{M,S}` = M·Σ, still occurs in 11,497 forms across 2,974 languages; the rarest triangle, `{M,S,T}`, in 3,031 forms across 1,224 languages. The complete 2-skeleton therefore survives until a form-threshold above 3,000 or a language-threshold above 1,200. The entire face lattice of sizes 4–7 is also saturated: all 35 four-class, 21 five-class, 7 six-class and the single seven-class set occur (the full

{P,T,K,S,L,M,N} appears 35 times across 24 languages). Exemplars must be read from the *coded IPA segments*, never from orthography: Spanish *conversación*, for instance, codes to the five-class set X,Ξ,Φ,Λ,Σ (k, n, β, r, s → K, N, P, L, S) and is not one of the seven-class forms — a reminder that the class assignment is a fact about segments, not letters (the IPA→class map is given in §2.2).

The homology of the global complex is therefore that of the full 2-skeleton of the 6-simplex:

$$\chi = V - E + F = 7 - 21 + 35 = 21, \qquad (b_0, b_1, b_2) = (1,\ 0,\ 20).$$

Every 1-cycle is filled by a triangle ($H_1 = 0$: no global holes), and the only non-trivial homology is the 20 = C(6,3) independent 2-voids inherent to the complete construction.

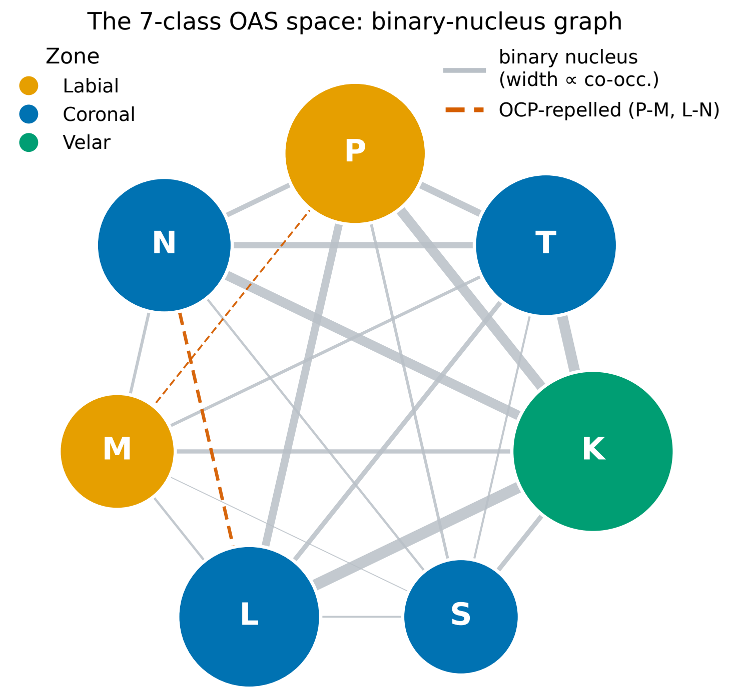

Figure 1: The seven OAS classes as a graph (the 1-skeleton of the code space). Node area is proportional to class membership frequency; nodes are coloured by articulatory zone (labial P/M, coronal T/S/L/N, velar K); edge width is proportional to binary co-occurrence. Dashed red marks the two strongest same-zone (OCP) repulsions, P·M and L·N. Symbols: P=Φ, T=Θ, K=X, S=Σ, L=Λ, M=Ϻ, N=Ξ.

> Insight I-1 (saturation). The OAS combinatorial space is not sparsely sampled by language — it is *exhaustively* realized. There are no globally forbidden codes: every pairing, every triple, every higher combination of the seven classes surfaces in real words somewhere on Earth. For a theory that posits the seven classes as the generators of a universal code layer, this maximal realization is a corpus-level consistency check — though, with 1.1M forms and only 21+35 faces to fill, complete saturation is combinatorially near-certain and carries little data content (§10.5); the generators combine freely, as almost any generators would.

3.3 Per-language subcomplexes

Restricting Δ to a single language yields that language's realized subcomplex. Across 4,721 languages:

quantity	min	median	max
vertices (classes)	2	7	7
edges (binary nuclei)	1	18	21
triangles (ternary)	0	18	35
graph components b_0	1	1	3
genuine H_1 holes	0	0	14

59 languages realize the full complex (all 21 edges and all 35 triangles), 21 of them Indo-European; 97.8 % have a connected 1-skeleton. Complex size correlates with corpus depth: Pearson r(size, log form-count) = 0.732. By family, the richest complexes are Afro-Asiatic (mean 46.2/56), Indo-European (44.5), and the sparsest are Sino-Tibetan (23.9) and Austroasiatic (29.3).

3.4 The closure axiom fails — the nucleus is a whole

A simplicial complex must satisfy the closure axiom: every face of an attested simplex is itself attested. Language violates this. In 66.8 % of languages (3,154 / 4,721) a ternary nucleus is realized while at least one of its own binary sub-nuclei is *not* realized in that language (15,354 such incidents). That is: a three-class code lives where its two-class faces do not.

Insight I-2 (the whole is not the sum). Empirically the per-language attested set is a hypergraph, not a simplicial complex. Endolinguistically this is exactly what the ontology predicts: a nucleus is a *Gestalt*, an irreducible orientation, not a construction glued from its sub-codes. A ternary is not "a binary plus a class" in the compositional sense — it can be present precisely where the binary is absent. The closure-failure rate was read as a *measurement* of code irreducibility — but a phonotactic null withdraws this (§10.5): real languages fail closure *less* than the null, so the failure is sampling sparsity, not a Gestalt. The irreducibility claim must rest on other evidence (or be re-tested against the null).

Hypothesis H-1. Closure-failure rate is a typological variable. We predict it is higher in languages/families with richer ternary morphology (root-and-pattern, heavy clusters) and lower in analytic languages that realize few ternaries. A regression of per-language closure-failure against agglutination rate and mean word length would test this.

3.5 Topology tracks typology

The genuine simplicial H_1 (holes remaining after downward closure) is common at the language level: 48.1 % of languages carry $\geq$1 hole (median 0, max 14). The maximally-holed languages are Tai-Kadai and Sino-Tibetan — Shan, Chadong, Tai Lü, Tai Khün, Atsi/Zaiwa: they draw a nearly complete *edge* graph (20 of 21 binary pairs) yet realize essentially zero ternary triangles, leaving up to 14 unfilled cycles. These are exactly the analytic, tonal, monosyllabic languages whose phonotactics forbids three-consonant nuclei.

Insight I-3 (homology as a typological invariant). The first Betti number of a language's code complex is a *morphological signature*: $b_1 \approx 0$ for synthetic/agglutinative languages (their triangles fill the cycles), b_1 large for analytic/isolating ones (edges without triangles). Topology recovers the analytic–synthetic axis from consonantal codes alone.

Question Q-1. Do the *locations* of the holes (which class-cycles go unfilled) carry areal or genealogical signal? A persistent-homology barcode per family, compared phylogenetically, could turn code topology into a classification feature independent of cognates.

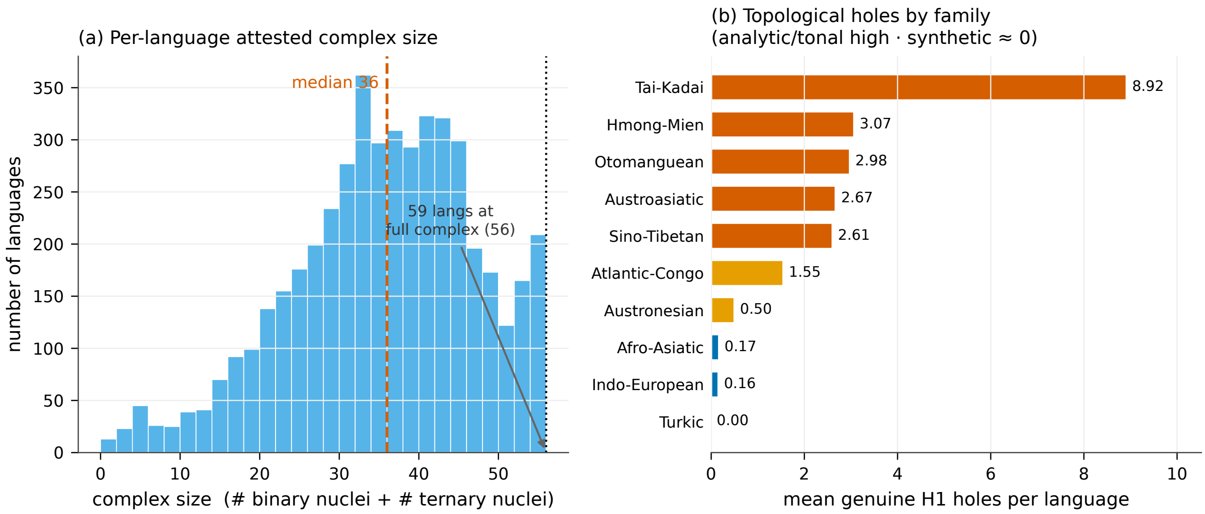

Figure 2: Per-language code topology. (a) Distribution of the attested complex size (edges + triangles realized) across 4,721 languages; median 36, and 59 languages at the maximum 56. (b) Mean number of genuine H_1 holes per family: Tai-Kadai and Sino-Tibetan (analytic, tonal, monosyllabic) sit highest — near-complete edge graphs with almost no ternary triangles — while synthetic families (Turkic, Indo-European) sit near zero.

3.6 Composites as higher simplices

Composite forms populate the higher simplices densely: all 64 class-sets of size 4–7 occur. Composite *forms* concentrate in consonant-rich and agglutinative families — Indo-European (19,008 forms), Austronesian (12,942), Nakh-Daghestanian (12,229), Uralic (9,051), Turkic (5,701) — where the last three punch far above their language counts, confirming that agglutination pushes words into higher-dimensional simplices. The persistence filtration identifies the pairing S,M = Σ·**M** (sibilant + labial nasal) as the structural weak point of the whole space (its simplices die first), and the velar–liquid–labial backbone `{P,K}`, `{P,K,L}`, `{T,K,N}` as maximally persistent.

4. Class-Relational Structure: an OCP Law Hidden Behind Frequency

4.1 The association matrix

For binary nuclei, let $A(x,y) = \log_2($ observedx,y / expectedx,y $)$, expectation from class marginals. Positive A = attraction (over-combination), negative = repulsion. The matrix (values ×100 for legibility):

	P	T	K	S	L	M	N
P	—	+34	+29	+22	+59	−62	+9
T	+34	—	+37	−10	−5	+31	+39
K	+29	+37	—	+26	+57	+15	+43
S	+22	−10	+26	—	−7	+18	+19
L	+59	−5	+57	−7	—	+12	−50
M	−62	+31	+15	+18	+12	—	+47
N	+9	+39	+43	−50	+47	—	—

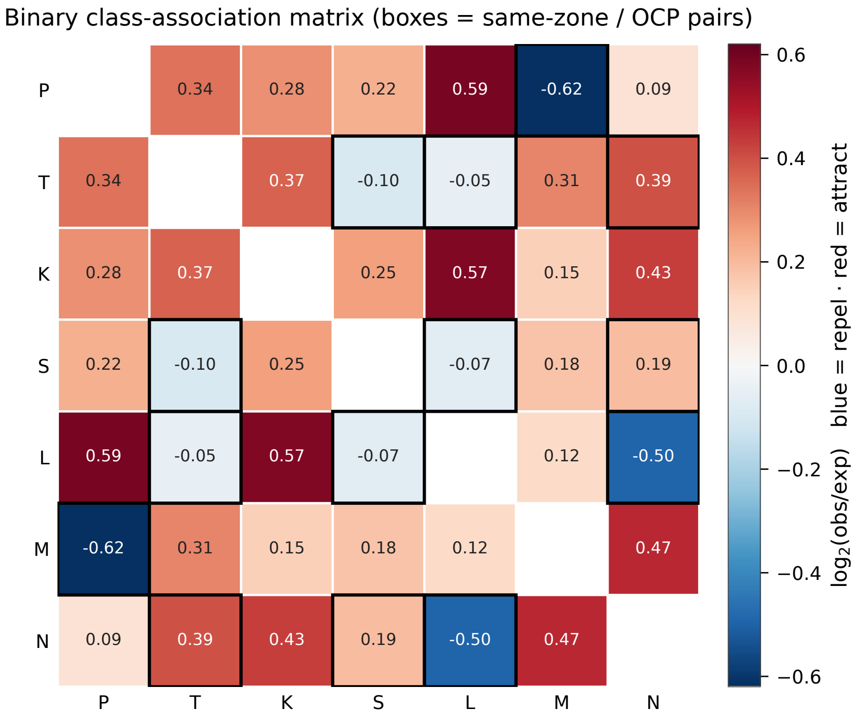

Figure 3: Binary association matrix A(x,y)=$\log_2$(observed/expected), for the 21 class pairs (diverging scale: blue = repulsion, red = attraction). Boxed cells are same-articulatory-zone pairs; they cluster on the repelled (blue) side — the OCP law. The strongest repulsions are the homorganic P·M (−0.62) and L·N (−0.50).

4.2 The same-zone repulsion (OCP) law

Partitioning by articulatory zone, the mean association within a zone is −0.107 (7 homorganic pairs) versus +0.311 across zones (14 pairs); the gap is +0.418 (Welch $t \approx 3.19$). The two strongest repulsions are the two homorganic-manner pairs: P·M = Φ·**M** (labial stop + labial nasal) = −0.62 and L·N = Λ·Ξ (coronal liquid + coronal nasal) = −0.50. Classes of the same articulatory region *dissimilate* — they combine less than chance.

> Insight I-4 (emergent OCP). A phonological Obligatory-Contour-style principle — avoid combining like-with-like — is not imposed on the OAS system; it *emerges* from the realized combinatorics. The seven classes are not exchangeable tokens: each carries its articulatory-zone identity into the code, and the code layer inherits a dissimilation law. (The coronal zone is internally heterogeneous — T·N and S·N are mildly attractive — suggesting the four coronal classes are not a single affective region but a graded one.)

4.3 The frequency gotcha and the X hub

Crucially, this law is invisible in raw co-occurrence. On the un-normalized co-occurrence graph, exhaustive modularity maximization returns a single community ($Q = 0.000$) and the Fiedler vector does *not* split the classes into zones: the graph is one dense near-clique. The reason is the velar hub K = X, which carries 22.6 % of all binary endpoints and the highest eigenvector centrality (0.207) — its sheer frequency drowns the relational signal. Only after dividing out the marginals (the A matrix) does the zone structure appear.

Insight I-5 (structure behind frequency — a methodological gotcha). In endolinguistic corpora, raw counts mislead: the most frequent class (X, the velar/generative depth) behaves as a universal connector and flattens community structure. Relational laws are only legible after normalization. *Any* future claim about "which codes go together" must be made on association, never on raw co-occurrence.

4.4 The ternary attractor set

For ternaries, the strongest over-representations are P,K,L = Φ·X·Λ (+1.36), T,K,N (+1.26), T,K,L (+1.10), P,T,L (+1.08) — a tight stop + liquid + X attractor built on the velar–liquid axis. The most repelled ternary is S,L,N = Σ·Λ·Ξ (−0.23), a triple of coronal continuants. Ternary code space is not uniformly explored: it concentrates on a small backbone.

Hypothesis H-2. The ternary attractor stop, K, L is the corpus signature of the classic Indo-European *plt/klt* root shape. We predict that within-family the attractor's strength scales with the family's root-template rigidity, and that families with CV-heavy phonotactics (Austronesian, Bantu) show a flatter ternary landscape. Testable by per-family association matrices.

5. The Operation Algebra: Order Carries the Quality

5.1 Inversion ι as an involution, and $q(C) \neq q(\iota C)$

The 252 ordered codes are closed under reversal; $\iota^2 = \mathrm{id}$ is verified for all of them, and no ordered code of distinct classes is a palindrome, so ι has no fixed points and partitions the 252 codes into exactly 126 orbits (21 binary + 105 ternary), each of size two — a free $\mathbb{Z}/2$ action.

The axiom $q(C) \neq q(\iota C)$ — that a code and its reverse orient *different* fields — is then tested empirically. For each of the 105 ternary inversion pairs (both sides attested), we compare the *set of concepts* each ordered code realizes in `lexeme`. The mean concept Jaccard overlap is 0.243 (median 0.224; cosine of frequency vectors 0.246): a code and its mirror share only about a quarter of their concept repertoire. The most orthogonal pairs are strikingly clean minimal pairs:

- `M·P·S` (Μ·Φ·Σ) peaks on 'moon' (Bemba *umwèsi*); its reverse `S·P·M` peaks on 'swim' (German

schwimmen). Jaccard 0.078.

- `P·M·S` peaks on 'blood'; `S·M·P` on 'navel'.

The *highest*-overlap inversion pairs are all L- or N-centred (`K·L·T`↔`T·L·K`, J=0.562): when a stable class holds the middle, reversal only swaps the flanks and meaning drifts least.

Insight I-6 (order is load-bearing). Inversion is a genuine involution whose action on meaning is systematically non-trivial: reversing a code re-orients it to a near-disjoint field. The endolinguistic "isomer" claim is thereby quantified — `moon` and `swim` are the *same three classes in opposite order*. Order is not decoration on a nucleus; it is where a large part of the quality lives.

Question Q-2. Is the semantic *distance* between C and ιC predictable from which class sits in the centre? The data hint that a mobile centre (L, N) minimizes the inversion gap while a peripheral-forcing centre maximizes it — a quantitative "pivot stability" measure worth formalizing (cf. §7).

5.2 Insertion and reduction: a clean retraction, but derivation expands

Insertion $\oplus$ and reduction $\ominus$ form a section–retraction pair: dropping the inserted class from a ternary recovers the base nucleus in 200,000 / 200,000 = 100 % of sampled insertion events ($\ominus \circ \oplus$ = id on nuclei), the algebraic hallmark of an adjoint-like relationship.

Yet at the *derivational* level the system moves overwhelmingly in the $\oplus$ direction. Among 8,692 attested verb→derivative (deverbal) operations, the relation distribution is:

relation	share
RESONANCE (nucleus preserved)	42.6 %
EXPANSION ($\oplus$, arity ↑)	29.7 %
PERMUTATION	16.1 %
OTHER	9.0 %
REDUCTION ($\ominus$, arity ↓)	2.7 %
INVERSION	0.03 %

The reduction : expansion ratio is 0.091 — derivation *expands* a code about eleven times more often than it collapses one; only 7.7 % of ternary verbs shed a class in derivation.

Insight I-7 (a falsified intuition). The prior expectation that deverbal morphology *reduces* a ternary verb to a binary noun (a "collapse") is wrong at scale. Derivation preserves the code (resonance) or *grows* it (expansion). The natural direction of morphological productivity is $\oplus$, addition, not $\ominus$. This is precisely the kind of result the discipline's falsifiability stance exists to surface, and it should replace the reduction assumption in the methodology.

Hypothesis H-3. If derivation is $\oplus$-biased, then across a word family the *arity* should tend to increase from base to derivative, giving each family a rough "arity gradient." We predict deverbal nouns average higher arity than their verbs, and that the few genuine reductions cluster in specific semantic classes (e.g. deverbal *instrument* or *result* nouns). Testable on the `deriv` table by semantic role.

5.3 The S_3 action and chirality

For each of the 35 ternary nuclei, all six orderings are attested (the S_3 orbit is fully realized — §3), but all 35 are significantly non-uniform (χ^2, df=5, every $p<0.001$; χ^2 from 138 to 9,180). Isomer choice is never random, yet dominance is weak: the top ordering holds on average only 28 % of tokens (vs 16.7 % uniform), and no nucleus has perfect cross-family consensus (mean agreement 0.425).

The degree of order-preference — call it chirality — is governed by sonority. Measuring the normalized Shannon entropy H_norm of the six-ordering distribution (0 = one order dominates, 1 = racemic), obstruent-only ternaries (K,S,T etc.) are essentially racemic (H_norm → 1.00), whereas ternaries containing a liquid or nasal are chiral (lower entropy; most chiral L,N,P = $\Lambda \cdot \Xi \cdot \Phi$). Mean entropy with Λ = 0.951 vs 0.972 without; with a nasal 0.958 vs 0.975.

Insight I-8 (sonority induces handedness). When a ternary is built of pure obstruents, its six orders are near-interchangeable — the code is "ambidextrous." Introducing a sonorant (L, M, N) imposes a preferred order — a *chirality*. Sonority sequencing, a classic phonological principle, thus reappears at the OAS code level as a symmetry-breaking of the S_3 action.

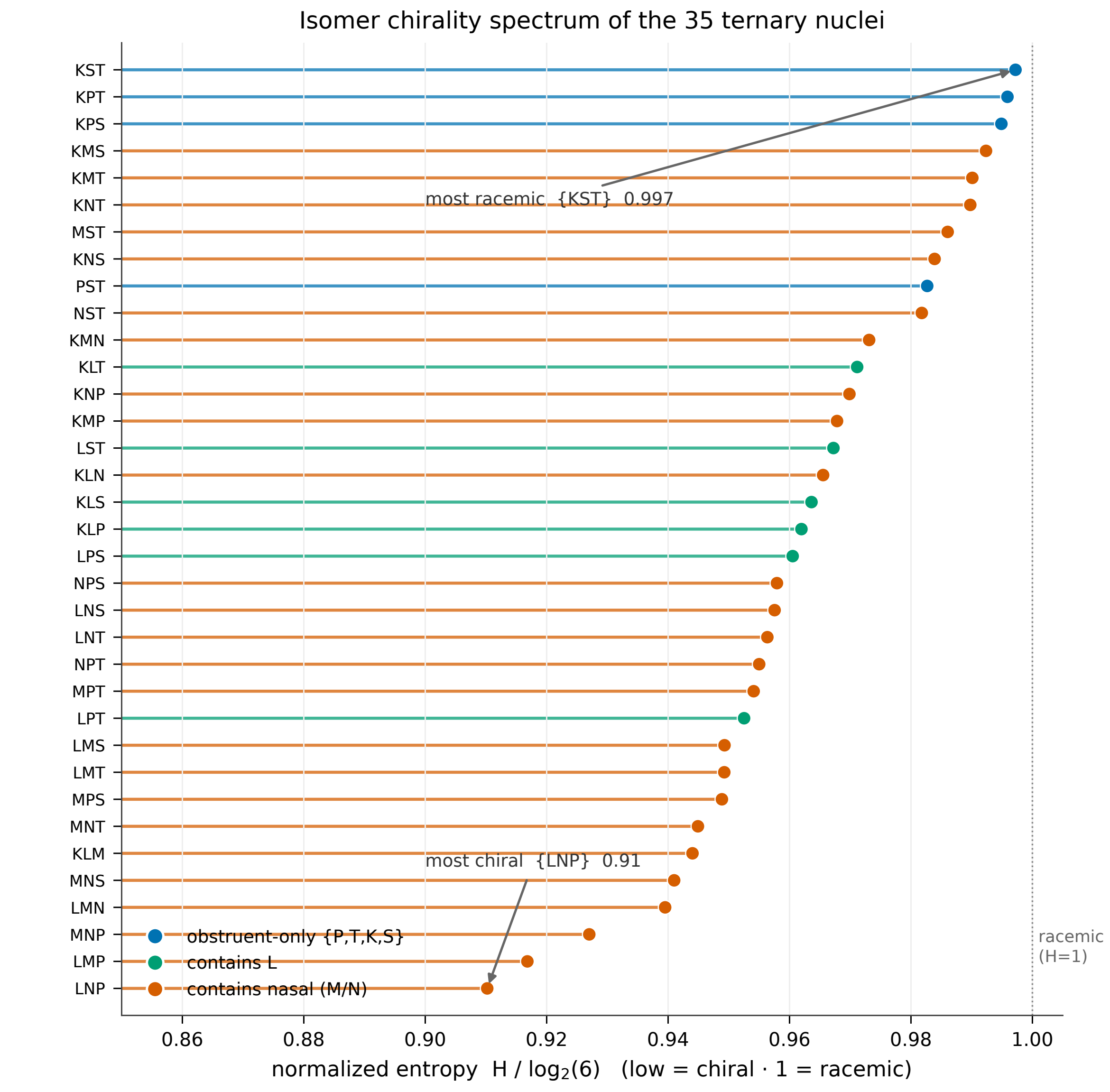

Figure 4: Isomer chirality spectrum: normalized entropy H_norm of the six-ordering distribution for each of the 35 ternary nuclei (0 = one order dominates/chiral, 1 = racemic), sorted, coloured by sonority category. Obstruent-only nuclei (e.g. K·S·T) are racemic (H_norm≈1.00); nuclei containing a liquid or nasal (e.g. L·N·P) are chiral. Sonority breaks the S_3 symmetry.

Question Q-3. Is there a *canonical form* functor — a rule that picks, for each nucleus, its dominant order — and is it natural with respect to families? The strongest canonical orders are L-pivoted (P·L·N at 94 % family agreement); the weakest are K-heavy and L-absent. A group-cohomological description of the obstruction to a global canonical section may exist.

5.4 A category-theoretic sketch

The objects and operations assemble into a small category $\mathcal{C}$: take *ordered codes* as objects and the generated operations (insertion, reduction, substitution, permutation, inversion) as morphisms. Then ι is an involutive endofunctor (ι^2=id) whose action on the "meaning" side is non-trivial (§5.1); $\oplus$ and $\ominus$ form a section–retraction (adjoint-like) pair on the nucleus functor (§5.2); and S_n acts on the fibres of the forgetful functor *ordered* $\mapsto$ *unordered* (§5.3). The endolinguistic quality map q is then a functor $\mathcal{C} \to \mathcal{Q}$ into a category of oriented meaning-fields, and the axiom q(C)$\neq$q(ιC) says exactly that q does not factor through the forgetful functor — meaning is not an invariant of the unordered nucleus. Making $\mathcal{Q}$ precise (e.g. as concept-set-valued presheaves, with Jaccard as a metric) is the natural next formalization.

6. Directed Dynamics: the Substitution Total Order

6.1 The chain

From 209,444 substitution events (diachronic, against the ancient-IE anchor), form the 7$\times$7 count matrix M[X$\to$Y] and row-normalize to a stochastic matrix P. Its stationary distribution π (long-run residence of a random walk of substitutions) is

$$\pi: \quad P\,0.180 > T\,0.167 > K\,0.167 > N\,0.147 > L\,0.122 > M\,0.110 > S\,0.107,$$

so P, T, K are attractors — substitution mass accumulates on the labial stop, the dental, and the velar. The chain is irreducible and aperiodic, and it mixes fast: the second eigenvalue $|\lambda_2| = 0.283$ (spectral gap 0.72, relaxation $\approx$ 0.8 steps).

6.2 The perfect total order

The net flow between two classes, F(X,Y) = M[X$\to$Y] − M[Y$\to$X], defines a tournament. This tournament is perfectly transitive. There is a linear order

$$\boxed{\mathrm{L}\,(\Lambda) \;>\; \mathrm{S}\,(\Sigma) \;>\; \mathrm{K}\,(\mathrm{X}) \;>\; \mathrm{N}\,(\Xi) \;>\; \mathrm{P}\,(\Phi) \;>\; \mathrm{M} \;>\; \mathrm{T}\,(\Theta)}$$

down which 100.0 % of net flow runs and along which all 21 pairwise net-flows are strictly positive — zero directional cycles. Class Λ (liquid) is a near-pure source (out 52,019 / in 21,720, net −30,299, by far the largest imbalance), and Θ (dental) is the strongest sink (net +13,433). The two largest single flows in the entire matrix are Λ$\to$Ξ (11,705) and Λ$\to$Θ (11,157).

Control (R3): this order is largely a frequency/anchor artifact. Two facts deflate it. First, the "perfect transitivity" is *trivial*: the tournament is transitive because it is essentially a scalar ranking, and any scalar ranking is cycle-free by construction — which answers Q-4's puzzle directly. Second, that scalar is largely class frequency: the order correlates $\rho = +0.93$ (Spearman) with the ranking by the *anchor-to-target class-frequency ratio* r = f_anchor/f_target (which gives S>L>K>N>P>T>M, differing by only two adjacent swaps), while *raw* class frequency does not predict it ($\rho = 0.14$). So "Λ near-pure source" reduces largely to "the ancient-IE anchor is Λ- and Σ-rich relative to the world average" — an anchor-selection effect. Combined with the cognacy result (83 % of these comparisons are non-cognate, §10.5), the "arrow of cualic change" is not established as diachronic; whether a residual arrow survives on a properly oriented cognate-only subset is the open test (`src/tournament_baseline.py`, `src/cognacy_control.py`).

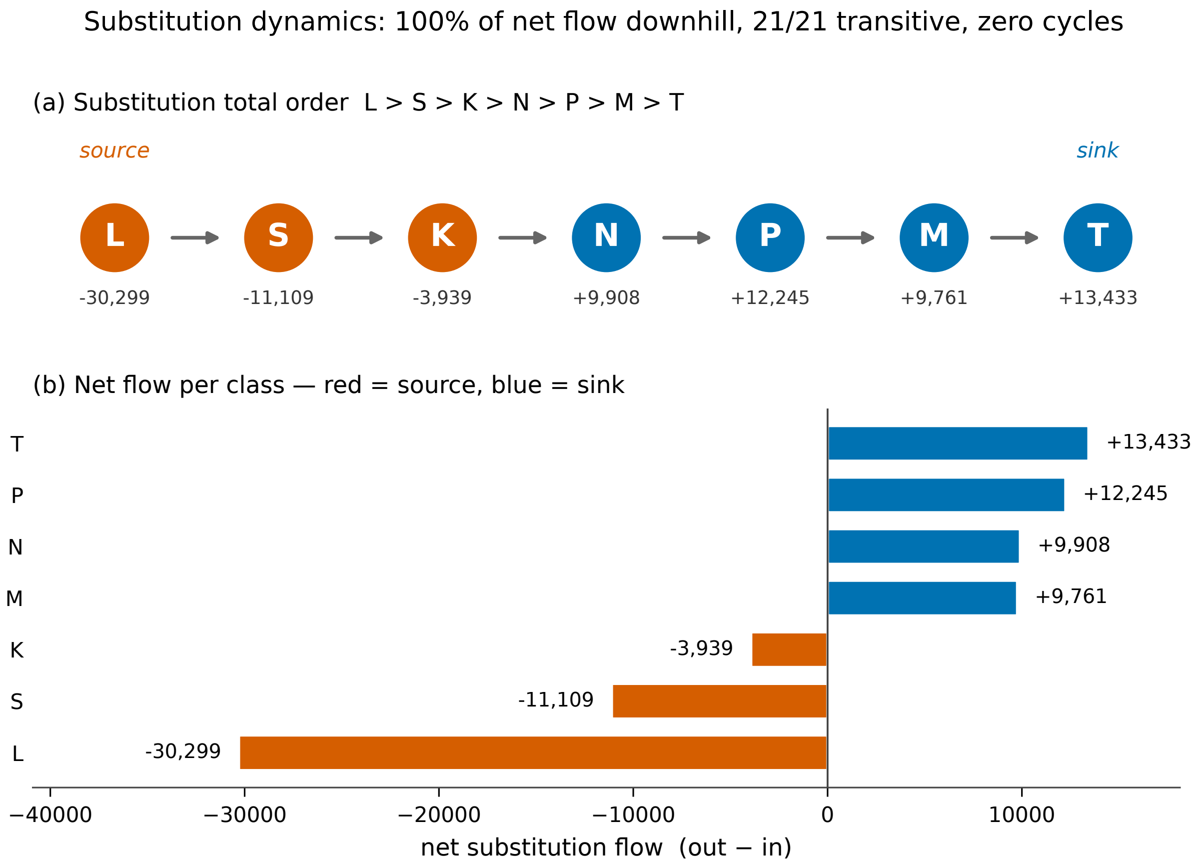

Figure 5: The substitution total order. (a) The seven classes arranged on the single decay axis L>S>K>N>P>M>T, from source (left) to sink (right). (b) Net substitution flow per class (in−out), sorted: red = sources (net emitters), blue = sinks (net absorbers). Λ (L) is a near-pure source (−30,299); Θ (T) the strongest sink. The tournament is perfectly transitive — 100% of net flow runs downhill, all 21 pairwise net-flows positive.

> Insight I-9 (an arrow of cualic change). Endolinguistic substitution is not diffusive noise; it has a single, cycle-free directional axis. Codes "decay" from the liquid toward the stops and nasals — L dissolves preferentially into T, N, P. This is a thermodynamic-like gradient: a low-entropy, high-mobility source (the liquid) draining toward stable attractors. This was read as an *arrow of time* at the class level — but a cognacy control withdraws the diachronic reading (§10.5): 83 % of the compared pairs are non-cognate lexical replacement, so the "decay" is a *synchronic* (plausibly class-frequency-driven) modal-nucleus pattern, not attested change. Whether a real arrow survives on the cognate-only subset is the open test (§10.5, R3).

> Hypothesis H-4. If the substitution order is a true decay gradient, then older attested languages should sit "higher" (more Λ-rich) and younger ones "lower" (more Θ/Φ/nasal-rich) for the same concept. The order predicts the *direction* of every future substitution; a held-out diachronic test set (e.g. Latin→Romance) should show substitutions overwhelmingly downhill. A violation rate significantly above the null would falsify the gradient.

Question Q-4. Why is the tournament *perfectly* transitive? Random substitution matrices are almost never acyclic. The zero-cycle property suggests the seven classes carry a hidden scalar potential (a one-dimensional ”cualic charge”) that substitution always descends. Recovering that potential — a single number per class consistent with all 21 inequalities — would be a strong structural claim (§7 suggests the liquid–stop / sonority axis as its interpretation).

7. The Λ-Pivot / X-Shell Law, and the Λ Paradox

7.1 Convergent evidence

Three independent tables agree that Λ (L) is the structural pivot and X (K) is the shell:

1. Insertion position (`permutation`, 818k events): overall, internal insertion is the *least* common

position (29.5 %; final 35.5 %, initial 35.0 %) — but L is the only class that prefers the internal slot (41.3 %), wedging between the flanks. K, by contrast, favours the edges.

1. Canonical order (S_3 dominants): L sits in the middle of the dominant ordering in 73 % (11/15) of L-containing nuclei; K sits in the middle in 0 % (0/15) — K is *always* peripheral.
2. Cross-family consensus: the nuclei with the strongest canonical agreement are exactly the L-centred

ones (P·L·N at 94 % of families), while the weakest are K-heavy, L-absent (`K·S·T` at 26 %).

Insight I-10 (Λ pivots, X shells). The liquid organizes the *spine* of a code — it seeks the centre — while the velar organizes the *shell* — it seeks the edge. Two classes, derived from three unrelated operations, play opposite structural roles. This is arguably the deepest regularity in the corpus: a positional grammar of the classes.

7.2 The paradox and its resolution

Λ is simultaneously (a) the pivot — the most *positionally stable*, centre-seeking class (§7.1) — and (b) the source — the most *substitutionally mobile* class, top of the decay order, emitting 2.4× what it absorbs (§6). And at the token level Λ is the most-lost class (net token balance −41,598, #1 in reductions) — yet nuclei that *contain* Λ conserve the best (11.86 %, higher than for any other class), while the worst-conserving nuclei are those containing S (8.50 %).

The paradox resolves once level is separated:

- *Positional role* (where L sits in a code): central, stable — L anchors the order.
- *Substitutional dynamics* (whether the L token survives): fluid — the L segment is the first to be replaced.
- *Nucleus fate* (whether an L-bearing code is conserved whole): robust — an L-code is well-preserved even as

its L token churns.

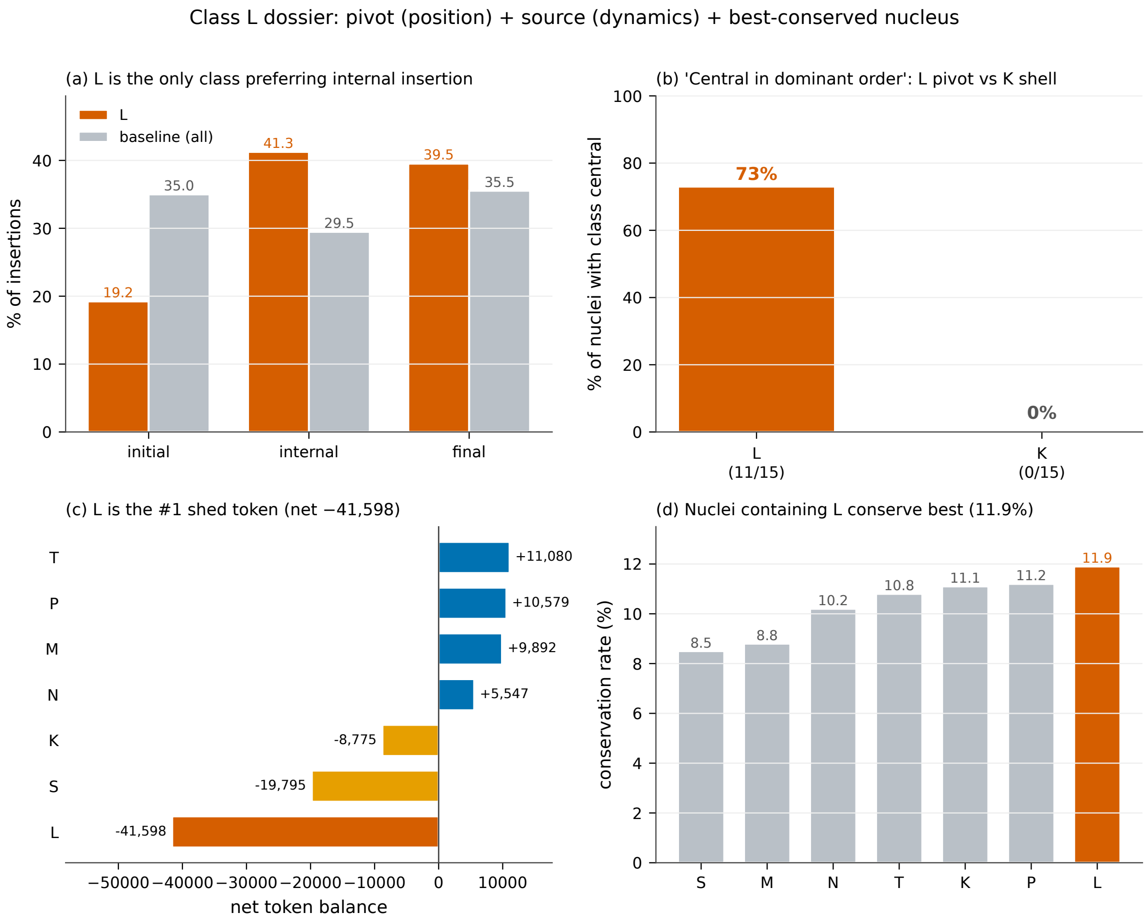

Figure 6: The Λ (L) dossier — convergent evidence for the pivot role and the paradox. (a) L uniquely prefers internal insertion (41.3% vs 29.5% baseline). (b) L is central in the dominant order of 73% of its nuclei; K in 0% (always peripheral). (c) L's net token balance (−41,598) is by far the most negative — the most-lost class. (d) Yet nuclei containing L conserve best (11.86%). Positional pivot and substitutional source at once.

Insight I-11 (Λ organizes by flowing — a structural characterization). The very mobility that makes Λ the substitution source is what lets it flow to the centre and anchor the code's geometry — Λ organizes *by* being fluid. Here "fluidity" names a corpus fact (substitutional mobility together with positional centrality), not physical liquidity and not a psychic category. Any relation between this structural fluidity and unconscious psychic processes is a *separate psychophonological hypothesis that this study does not test*; it must not be read as a demonstrated identity ($\Lambda \neq$ "the unconscious"). This is a type-(4) interpretive note, flagged as such.

Question Q-5. Is the pivot/shell dichotomy binary or graded across all seven classes? A single "centrality-seeking" score per class (from insertion position + order centrality) would rank Ω on a pivot$\leftrightarrow$shell axis; does that axis coincide with the sonority axis of §5.3 and the decay potential of §6? If the three independently-derived orderings of the seven classes coincide, endolinguistics has found a single latent scalar organizing the whole class system.

8. Scaling and Diachrony

8.1 Order, not arity, makes the inventory Zipfian

Ranked by frequency, the 56 unordered nuclei follow a *shallow*, near-flat law (log–log slope −0.87, R^2 0.86, Gini 0.44): the head nucleus `{K,L}` holds only 6.5 %, the top ten 45.8 %. Impose order and the 252 ordered codes follow a near-textbook Zipf law (slope −1.18, R^2 0.92, Gini 0.65). Within a fixed arity the slope is a constant ≈ -0.7 for both binaries and ternaries; the heavy tail lives entirely in the 210 ternary permutations, each carrying little mass.

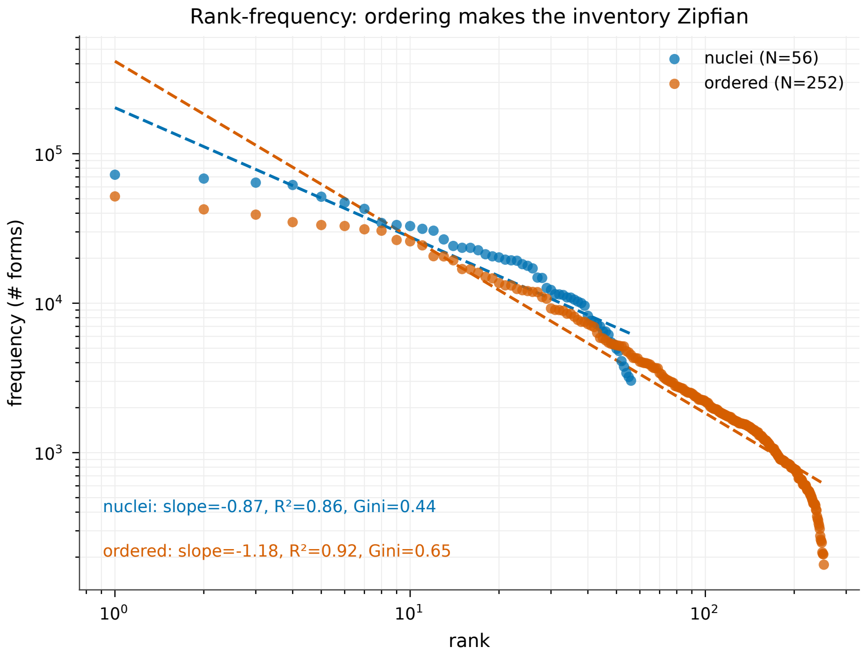

Figure 7: Rank–frequency (log–log). The 56 unordered nuclei follow a shallow, near-flat law (slope −0.87, Gini 0.44); the same data with class order imposed — the 252 ordered codes — follows a near-textbook Zipf law (slope −1.18, R^2=0.92, Gini 0.65). Ordering, not arity, creates the heavy tail.

Insight I-12 (information lives in the order). The *unordered* code inventory is egalitarian — the class-combinations are used with comparable weight. All the concentration, all the heavy-tailed structure, appears only when order is imposed. This is the distributional face of q(C)≠q(ιC) (§5.1): if order carried no quality, ordered and unordered inventories would obey the same law; they do not. Order is the information-bearing degree of freedom of the code system.

8.2 Diachrony versus resonance

Against the ancient-IE anchor, the Indo-Uralic macrosystem — where the anchor is genuine ancestry — conserves at 23.15 % (Indo-European alone 27.05 %), whereas *all other macrosystems pooled*, where a match is mere resonance, conserve at 9.34 % (N=290,748). Real descent conserves 2.5× more than cross-macro resonance — a clean, falsification-grade separation that validates the anchor method while confirming that resonance is a weaker-but-nonzero signal. In the vocabulary of the companion essay (*The Hidden Order*), a genuine-descent match is a coderivative — a *cognate* named from the code it conserves rather than from its etymology — and these three quantities separate three registers of sense: found in the coderivative's 23 % (descent guarantees it), constructed in whatever bounded resonance survives the controls (interpretable, not denoted), and projected in the confounds that must be removed (universal sound-symbolism, areal diffusion, chance) — apophenia is only the third. The cognacy control is, in these terms, a coderivative control: the discipline's *collective* judge between constructed sense and projection. Every single `pierde` (loss) event is a 3→2 reduction (100 %), and binary anchors conserve better than ternary (13.18 % vs 8.68 %): larger codes are harder to hold intact.

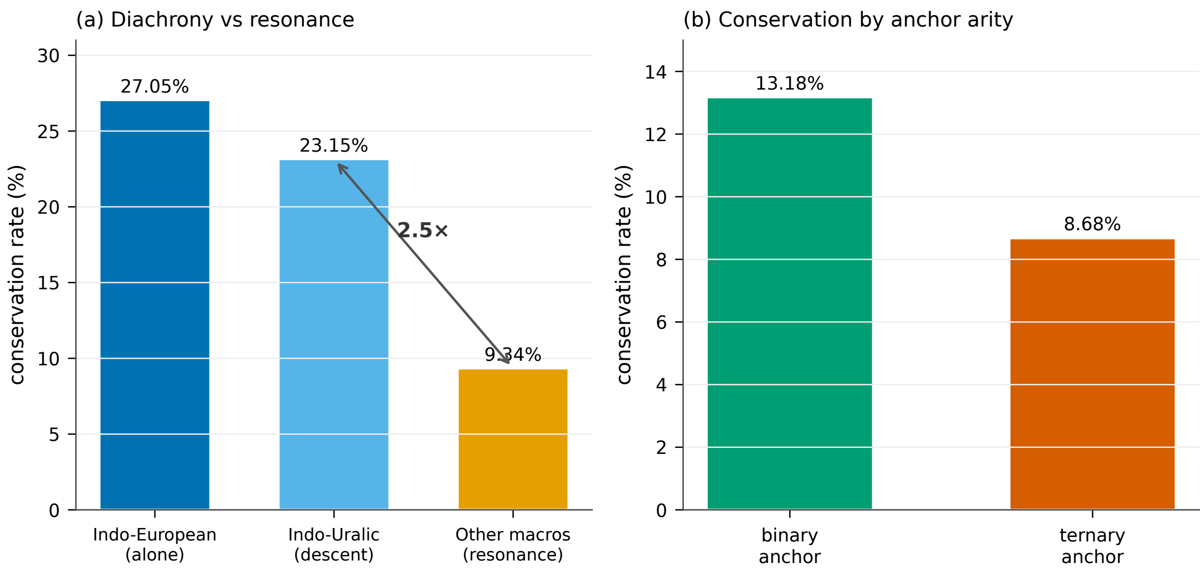

Figure 8: Conservation and diachrony. (a) Conservation rate by macrofamily: Indo-Uralic (genuine descent from the ancient-IE anchor) conserves 23.15% (Indo-European alone 27.05%) versus 9.34% for all other macrosystems pooled, where a match is resonance — a 2.5× separation. (b) Conservation by anchor arity: binary anchors (13.18%) hold better than ternary (8.68%).

Insight I-13 (resonance is real but faint). The 9.34 % cross-macro conservation is far below the 23 % of true descent, yet far above zero — establishing only that cross-macrosystem code-matching sits above chance-collision (2.15× the permutation baseline, §10.5). The earlier framing of this as "40 % of the diachronic signal" is withdrawn (R4): it compared a confound-laden upper bound — still containing universal sound-symbolism and areal diffusion — to genuine descent as if the two were commensurable. What survives is the modest statement: the match exceeds chance; *what fraction of that excess is endolinguistic remains unmeasured.*

8.3 Codes are semantically macrosystem-relative

The resonance rate has a sharp corollary that the corpus now lets us state directly: a code's *meaning* does not transfer across macrosystems. Taking the same widespread concept and reading off its modal nucleus in seven macrofamilies, the code matches the Indo-European one only 8 % of the time (5 of 60; e.g. 'two' is Φ·Θ in Indo-European but X·Λ in Austronesian and X·Φ in Mayan) — converging with the 9.3 % resonance figure of §8.2. Concretely, the nucleus Θ,Ξ orients *tooth / night / tongue* in Indo-European but *kinship / younger-sibling* terms in Mayan — arguably neighbouring "interiority/identity" fields, yet realized as different concepts. The seven classes generalize (they are the same everywhere), but the code→meaning projection is local to each macrosystem. This is exactly why the per-concept contextual anchor (§0, the German gloss in the noun atlas) is a *necessity*, not a convenience, and why macrosystems must never be pooled. Yet a faint class-level resonance survives the shift — 'name', for instance, keeps a nasal (M/Ξ) across Indo-European, Uralic, Turkic and Sino-Tibetan even as its full nucleus varies — which suggests the theory's abstraction gradient: the more abstract the level (class ⟶ nucleus ⟶ skeleton ⟶ IPA), the more it is shared; the more concrete, the more meaning localizes and diverges.

This gradient has a direct applied corollary (developed in the companion essay *The Hidden Order*). It is exactly why the endolinguistic code game is a learning accelerator *among coderivatives* and a different task *without them*: between co-descended tongues the conserved code carries the field, so a learner meets new vocabulary as *cases of a law*; but across macrosystems, where no coderivative anchors, the code→meaning projection does not transfer (8 % agreement), and the work must drop to the class level — the one rung of the gradient that generalizes — teaching structures of thought rather than ported meanings. The 8 % figure is thus not only a caution against pooling macrosystems; it is the empirical boundary of where cognate-based code pedagogy applies.

> Hypothesis H-6. Semantic divergence increases monotonically down the abstraction gradient. We predict cross-macro agreement is highest at the class level, lower at the (unordered) nucleus, lower still at the ordered code, and lowest at the skeleton/IPA — testable by computing cross-macrofamily agreement at each level for the same concept set.

8.4 Diversity is combinatorial, not inventorial

Family code-diversity (effective number of nuclei) spans 10.0 (Naduhup) to 45.9 (Uralic) — a 4.6× range — and is driven more by agglutination ($r = +0.37$ with composite rate) than by corpus size ($r = +0.22$). Crucially, *class*-level entropy is nearly uniform across families: families differ in how they *combine* the seven classes, not in which classes they possess. The seven-class inventory is a near-universal; the code *grammar* is what varies.

> Hypothesis H-5. If class inventory is universal but combination is family-specific, then a family's nucleus distribution is a fingerprint usable for classification. We predict a supervised classifier on per-language nucleus histograms recovers family membership well above chance and, for isolates, suggests nearest macro-neighbours — a cognate-free phylogenetic signal.

9. The Cualic Scalar, Class–Skeleton Levels, and a Data-Grounded Theory of Bonding

9.1 Two levels: class vs skeleton

The analysis forces an explicit separation the theory already implied. A class (one of the seven OAS) is fixed by affective-sound identity, with a stable articulatory anchor (its region). A skeleton is the real consonant realizing that class in a word. Crucially, *manner* properties — frication, voicing, plosiveness — live at the skeleton level, not the class level. A voiceless labial fricative /f/ is not a separate class: it is class Φ in a particular skeletal state. The one manner that is itself class-defining is sibilance: class Σ is the sibilants — the audible "hiss" — and nothing else; every other fricative belongs to its place-class (ɸ β f v → Φ, θ ð → Θ, x ɣ χ ç h → X). An IPA census that instead pools all fricatives into a single class — a natural but mistaken move — inflates that class roughly two-fold and is measuring skeletal manner, not OAS class. This two-level discipline is what makes the place-based mapping correct rather than arbitrary.

9.2 The within-class cualic modulation: a fortition↔lenition axis

The skeletal members of a class are not interchangeable tokens; they form a cualic gradient of fortition↔lenition. For the labial class: P (fortis stop) → B (voiced) → F (voiceless fricative) → V (voiced fricative) → W (approximant) → ∅. Moving along this axis is a *change of quality that conserves the class*, not a phoneme swap. pater → Vater is not "p became f": it is the same class Φ realized in a more lenited cualic state — the Latin endorhythm holds Φ *fortis* (*Pa-*), the Germanic endorhythm lenites it (*Va-/Fa-*). The escaping air of the fricative is precisely the cualic *colour* of lenition applied to the class it belongs to.

This is measurable. Across all labial-class realizations in the corpus, Germanic languages sit markedly further toward lenition than Italic: the fricative/approximant share (F+V+W) of Φ is 54.8 % in Germanic versus 45.3 % in Italic — a 10-point cualic shift. For the concept 'father' specifically, Indo-European realizes the initial Φ as both the fortis *p* (79 attestations: Latin/Romance *pater, padre*) and the lenited *f/v* (65 attestations: Germanic *father, Vater*) — one class, two cualic states. Grimm's-law "p→f" thus reappears not as a sound change but as an endorhythmic lenition of a conserved class, and this vindicates the place-based mapping of §9.1 (*f* is Φ-lenited, so it bonds with Φ).

The OAS system therefore carries two orthogonal cualic axes: a *between-class* axis (which class) and a *within-class* axis (the fortition state of that class). The conservation analysis of §6 and §8 lives on the first — it tracks the class/nucleus and is deliberately blind to P↔F, correctly scoring *pater/Vater* as "kept"; the fortition axis is the second, available to future work as a per-form scalar.

Modulation and class-crossing are one continuum. Mild within-class modulation stays in the class (P→F); pushed to the extreme it *crosses* into another class — the between-class substitution of §6 is intensified modulation. *foot*: Latin/Baltic and West Germanic keep the dental Θ (*ped-, foot*), but German *Fuss* and Ancient Greek push it across to the sibilant Σ (Θ→Σ), the corridor being affrication (7 % of Θ realizations are *tʃ*, 3 % *ts*). Direction is asymmetric: Σ→Θ = 6,611 (the downhill drift) vs Θ→Σ = 3,302 (the rarer uphill push of a strong "fortis" endorhythm such as German). And crucially, the within-class axis is not the same kind for every class. The three obstruents Φ/Θ/X share fortition↔lenition; but Σ's axis is palatalization (s→ʃ→ɕ, a dorsal/X gesture blended into the hiss — 20 % of Σ realizations are palatal, *not* a fortition process), Λ's is laterality↔rhoticity (l↔r, ≈equal), Ξ's is place of nasality (n→ŋ→ɲ), and **M** has essentially no internal axis (m = 97 %, the most compact and stable class). Characterizing the cualic axis of each class — and the crossing pathways between them — is a distinct research programme (detailed in the companion note `oas-classes-refined.md`).

9.3 The seven within-class cualic axes

The within-class modulation (§9.2) is not the same *kind* of axis for every class. Turning each into a per-form or per-segment scalar over the corpus gives seven distinct axes:

class	within-class axis	open/lenis pole → fortis/deep pole	measured signal
Φ	fortition ↔ lenition	W · V · F → B · P	Germanic 54.8 % lenited (F/V/W) vs Italic 45.3 %
Θ	fortition ↔ lenition	ð · θ → D · T	*affricates* tʃ ts dʒ are coded Θ by their stop onset (tʃ 7 %, ts 3 % of Θ) — not Σ (R5)
X	gutturality (depth × manner)	k (velar) → q → ħ → ʔ (deep)	velar 59 %, glottal 20 %; mean depth 2.41 · *(R6: the earlier "open pole w/u, 11 % vocalic" is a coding error — w is class Φ, ɰ is unclassed; neither feeds X)*
Σ	palatalization	s · z → ʃ · ʒ → ɕ · ʑ	20 % of Σ postalveolar/palatal
Λ	laterality ↔ rhoticity	l ↔ r	l 42 % / r 42 %
Ξ	place of nasality	n → ŋ → ɲ	n 70 %, ŋ 21 %
M	monolithic (no axis)	—	m 97 %

Three of these were quantified as scalars. Fortition (voiceless stop 5 ... sonorant 1; per-form mean over consonants) has a global mean of 2.79 and three properties that are *different in kind* from the between-class substitution arrow: it is positional (initial 3.11 › internal 2.65 › final 2.59 — the classic strong-onset/weak-coda hierarchy, recovered directly), typological (families span 2.36 Hmong-Mien → 3.45 Tucanoan), and — crucially — not directional in time (ancient 2.746 ≈ modern 2.745, $\Delta \approx 0$). At the whole-form level Germanic (2.69) and Italic (2.70) are nearly identical: the *pater/Vater* contrast is class-specific (labial Φ), a Grimm-type series shift, not a global endorhythmic fortition.

Palatalization (Σ's axis) has a per-form mean of 0.072; it is most active in Otomanguean, Uralic and Quechuan and near-absent in Austronesian. Gutturality (X's axis) is a 2-D field — place-depth × manner — whose deep-constricted corner (q, ħ, ʔ) is maximal condensation (the endolinguistic *límite / potencia* of K) and whose open corner (w → vowel u) is K dissolving toward the vowel; Amerindian families (Otomanguean, Tucanoan, Mayan) push K deepest via glottalization, Afro-Asiatic via uvular/pharyngeal constriction.

> Insight I-14 (each class its own axis). The OAS system is not seven interchangeable tokens sharing one modulation, nor a single fortition dial. Each class carries a *distinct* internal cualic axis — fortition for the labial/dental stops, gutturality-depth for the velar/guttural, palatalization for the sibilant, laterality↔rhoticity for the liquid, place-of-nasality for the coronal nasal, and none at all for the compact labial nasal. "Refining the classes" therefore means, concretely, naming and measuring seven different scalars, not tuning one.

9.4 The cualic scalar: a plasticity spine, not a single number

Given the tournament of §6 (which fixes an *order* but not magnitudes), we recover magnitudes by assembling a per-class profile of independent behavioural measures — substitution source-strength, per-token fragility, positional centrality (internal-insertion and middle-of-order rates), stationary residence, association hubness, chirality induction, membership frequency, and (external) IPA diversity — standardizing, and taking the first principal component. The result is the cualic scalar:

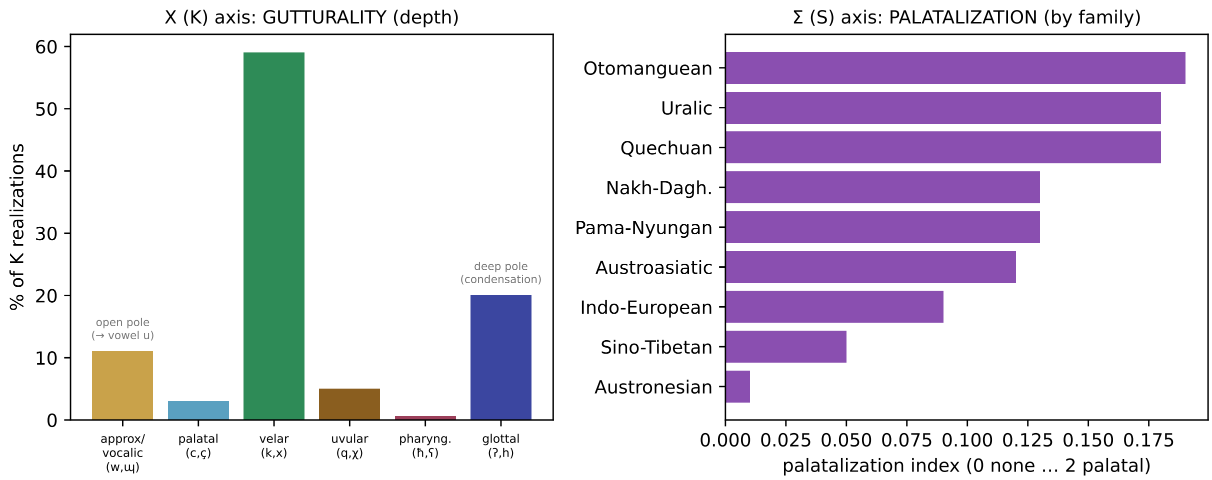

Figure 9: Two within-class axes as per-form/segment scalars. Left: the gutturality (depth) axis of X — K spans an open vocalic pole (w → vowel u) through a velar core to a deep glottal/uvular pole (the "strong Arabic q", ʔ). Right: the palatalization axis of Σ by family (Otomanguean/Uralic/Quechuan palatalize most, Austronesian least). Fortition and these two are distinct scalars — the seven classes do not share one modulation axis.

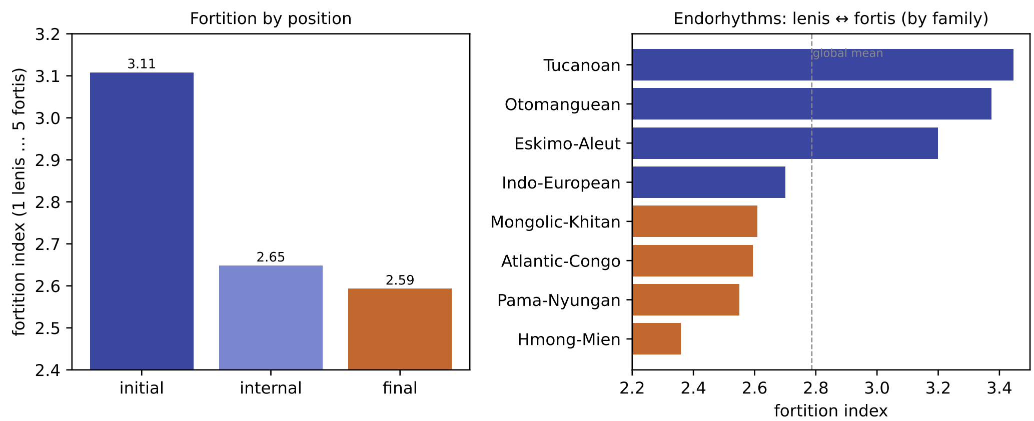

Figure 10: The fortition scalar (Φ/Θ/X manner). Left: fortition by position — consonants are fortis word-initially and lenite medially/finally (strong-onset hierarchy). Right: the fortis↔lenis endorhythm ranking by family (Hmong-Mien most lenis, Tucanoan most fortis; global mean dashed).

class	scalar (PC1)	class	scalar (PC1)
Λ (L)	+4.58	Φ (P)	−1.08
Σ (S)	+1.90	Ϻ (M)	−1.34
Ξ (N)	−0.87	X (K)	−1.34
		Θ (T)	−1.85

This directly answers the "single latent scalar" question (Q-5). The strong form is *not* supported: one number does not explain everything — PC1 captures only 48 % of variance, and mutation and sonority are statistically independent (rank $\rho \approx 0$). The weak form *is* supported: a dominant plasticity spine runs from Λ, a dramatic outlier at +4.58, to Θ, the anchor at −1.85, loaded jointly by mutation-source strength, positional centrality, and articulatory diversity — the "gestural plasticity" a purely articulatory account predicts, and it correlates with per-class IPA diversity ($\rho \approx +0.43$). The seven classes thus inhabit a small (≈2–3 dimensional) cualic space, not a line.

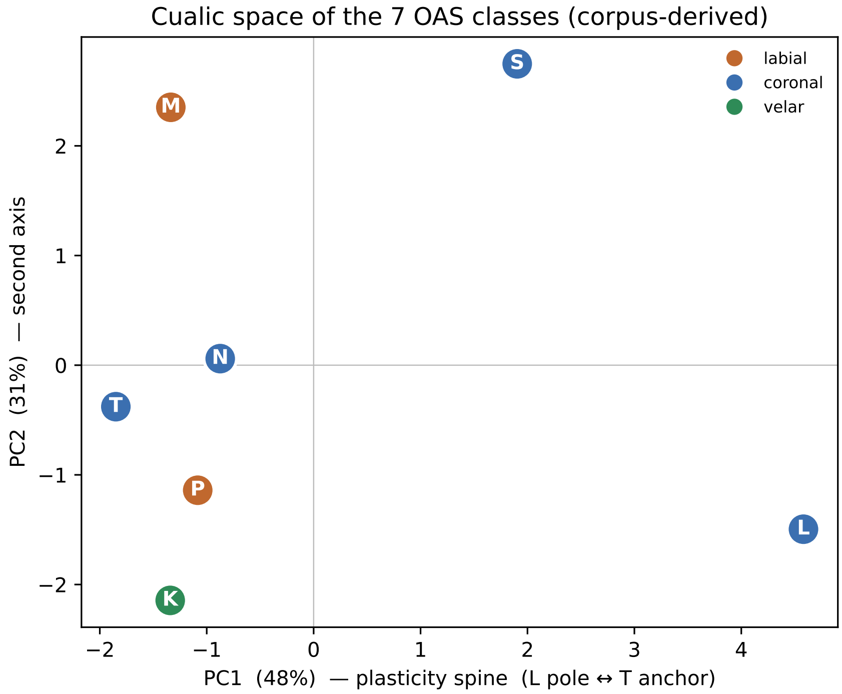

Figure 11: The cualic space of the seven OAS classes (first two principal components of the per-class behavioural profile). PC1 (48 %) is the plasticity spine: Λ is an extreme outlier (mobile, central, phonetically diverse) and Θ the anchor; PC2 is a weaker, exploratory second axis. Colour = articulatory zone.

9.5 The classes as operations, and bonding as a statistical law

Positioning each class on the plasticity spine turns the old "semantic-field" definitions into operations, now grounded in the corpus:

class	scalar	operation (data-refined)	statistical evidence
Λ	+4.58	circulate / bind	mutation source #1; pivot (centre); induces chirality
Σ	+1.90	emit (sibilant hiss)	2nd source; most fragile per token
Ξ	−0.87	interiorize / differentiate	nasal; assimilation-prone; mid
Φ	−1.08	impel	labial stop; substitution sink
Ϻ	−1.34	cohere / belong	labial nasal; stable; low frequency
X	−1.34	delimit / condense	hub + shell (edge); broad velar/guttural class
Θ	−1.85	fix / establish	anchor; final sink of the mutation order

Read down the spine, this *is* the plasticity gradient: circulate → emit → interiorize → impel → cohere → delimit → fix, from maximal transformability to maximal fixation.

Finally, "bonding" — how classes combine into codes — can be stated as statistical laws read from the 4,721-language corpus rather than posited:

- B1. Cross-zone preference (OCP). A code avoids two classes from the same articulatory zone; the most

repelled bonds are the homorganic Φ·Ϻ and Λ·Ξ (§4). Bonding seeks contrast.

- B2. Λ is the binder. The liquid seeks the centre of a code (internal-insertion 41 %, middle-of-order

47 %); it is the primary bonding agent, joining the flanks (§7).

- B3. X is the shell. The velar/guttural class seeks the edge (0 % central); it frames and delimits the

code (§7).

- B4. The ternary attractor. Three-class bonding concentrates on stop + Λ + X — the velar–liquid

backbone (§4.4).

- B5. Bonds age downhill. Over time a code's bonds relocate along the plasticity spine (Λ→…→Θ): the

Λ-bond is the first to dissolve or move, the Θ-bond the most permanent (§6).

- B6. Sonority chirality. A bond involving a sonorant (Λ, Ϻ, Ξ) acquires a preferred order; a bond of pure

obstruents is order-free (§5.3).

Together, §9.1–§9.4 refine the OAS classes from *fields* into *operations with a measured cualic charge*, and turn bonding from postulate into measured law — the empirical footing the theory sought.

9.6 The class-crossing drift field

Substitution can be drawn as a directed graph on the seven classes, and the corpus lets us grade every path as *certain* or *spurious*. With 4,721 languages, all 42 directed crossings are attested (each in 150+ families): at the class level there are no forbidden crossings — the mirror of the saturated code complex (§3). Certainty therefore lies in *direction* and *mechanism*. Collapsing each pair to its net direction, all 21 net flows run downhill on the plasticity spine (the perfect tournament of §6); their certainty scales with the net ratio — the Λ→X paths are certain drifts (ratios 2.0–3.4×), whereas Ξ→Φ (1.01×), Φ→Θ (1.02×) are non-directional (symmetric noise, not paths). Mechanistically, a crossing is grounded when a transitional form bridges it: affrication (tʃ, ts; 90,216 tokens) bridges Θ↔Σ (the *foot/Fuss* path); glottalization (h, ʔ; 32,603) bridges X→∅; approximation (w, ʋ; 19,298) bridges obstruent↔Λ; palatalization (ç, ɕ; 1,265) bridges X↔Σ. Strong drifts without a single corridor (Λ→Θ, Λ→Ξ) reflect Λ's multi-path plasticity. Against the drift, a few crossings are frequent but marked — they need a strong "fortis" endorhythm to push uphill: Φ→X (labiovelar w), Θ→Σ (German *Fuss*), X→Σ (palatalization). The drift field thus supplies a three-way certainty test for any proposed class transformation: directional (net ratio), mechanistic (corridor present), and marked (uphill/fortis).

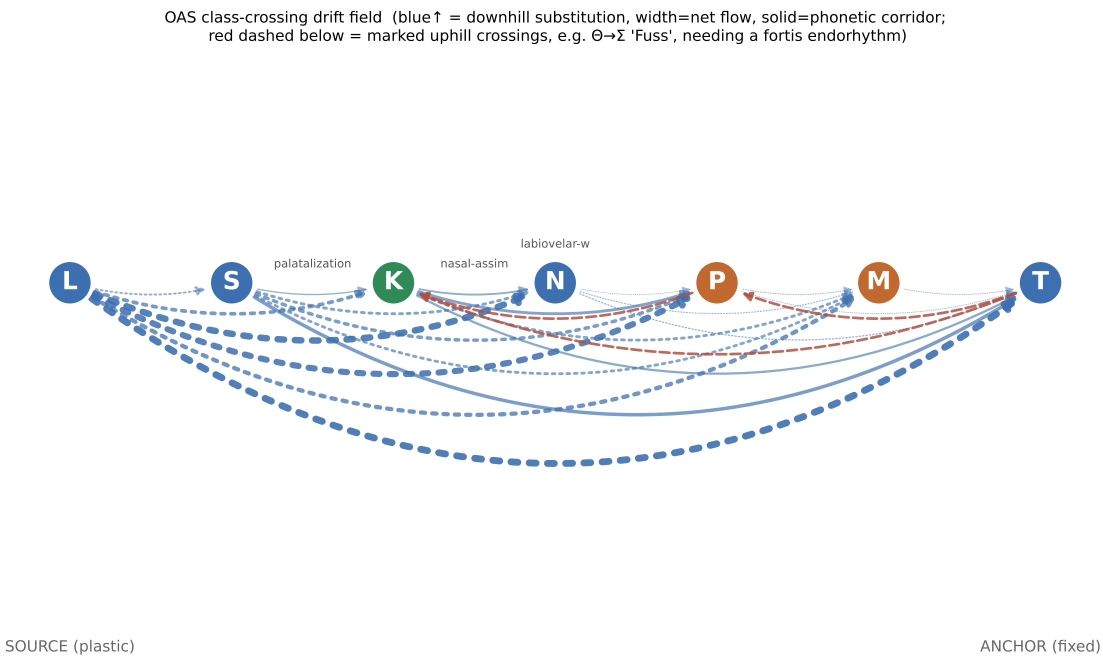

Figure 12: The OAS class-crossing drift field. Classes are laid out left→right by the plasticity spine (source Λ → anchor Θ). Blue arcs (above) are the net-downhill substitution drifts (width ∝ net flow, solid = a phonetic corridor exists, dotted = none); red dashed arcs (below) are the marked uphill crossings that require a strong "fortis" endorhythm, e.g. Θ→Σ (*foot*→*Fuss*). Node colour = articulatory zone.

10. Discussion

10.1 Three structural theses

The findings cluster into three claims about the endolinguistic object.

(T1) The whole is not the sum — weakened. Its first leg, the closure-axiom failure (§3.4), is withdrawn: a phonotactic null shows the failure is sampling sparsity, and real languages fail closure *less* than the null (§10.5). What remains is suggestive but not decisive on its own — the concentration of ternaries on an attractor rather than on the product of attractive binaries (§4.4) and the token-vs-nucleus split of the Λ paradox (§7.2). A hypergraph / weighted simplicial set is still a reasonable model of the *attested* space, but non-compositionality is now a hypothesis to be tested (a phonotactic-null control on the attractor concentration), not an established result.

(T2) Order carries the quality. Inversion re-orients meaning (§5.1), sonority breaks the S_3 symmetry into chirality (§5.3), and the rank–frequency law is Zipfian *only under order* (§8.1). The unordered nucleus fixes the *field*; the order fixes the *direction within it*. A theory of endolinguistic meaning must therefore be a theory of ordered codes, with the unordered nucleus as a coarser invariant.

(T3) Λ is central — not proven "the axis". The liquid is the positional pivot (§7, descriptive) and the induces-chirality class (§5.3). But its "source of the substitution order" leg (§6) is contaminated — synchronic, 83 % non-cognate, and plausibly a class-frequency artifact (§10.5) — and §9.4 finds no single latent scalar: Ω inhabits a ≈2–3-dimensional cualic space, not a line. Λ marks one pole of the *positional* organization; the stronger "single axis of Ω" claim is not supported.

Two meta-lessons frame the method: structure hides behind frequency (the X hub erases the OCP law until normalized, §4.3), and topology tracks typology (Betti numbers recover the analytic–synthetic axis, §3.5).

10.2 Toward a functorial semantics

The category sketch of §5.4 suggests a concrete programme: define $\mathcal{Q}$ as the category whose objects are concept-frequency vectors (a code's realized field) and whose morphisms are the induced maps under operations, and study q: $\mathcal{C} \to \mathcal{Q}$ as a functor. The axiom $q(C) \neq q(\iota C)$ becomes "q does not factor through the forgetful functor to unordered nuclei"; the substitution total order becomes a grading of $\mathcal{C}$ by a scalar potential; the chirality result becomes an obstruction to a canonical-form section. Persistent homology of the per-language complexes (§3.5) supplies typological features; the substitution Markov chain (§6) supplies a dynamical model whose stationary law and gradient are already measured.

10.3 Limitations

The anchor is Indo-European-specific, so cross-macro numbers are resonance, not descent (§8.2 turns this into a measurement rather than a flaw). Attestation is *exposure*: rare codes are hints, and the OCP and attractor results, while robust in aggregate, should not be read as per-code frequencies. The verb layer is coded approximately. WordNet-domain labels used elsewhere in the atlas carry noise. Finally, "articulatory zone" is our own coarse partition of Ω; the coronal heterogeneity noted in §4.2 warns that the zone map is itself an approximation to be refined.

10.4 Collected hypotheses, questions, and predictions

The discriminant experiment (H-0). Everything above tests naïve nulls; the one experiment that could isolate an endolinguistic object is not yet run. Stated as a falsifiable hypothesis: *there exists a residue of code→semantic-field associations that survives (a) a phonologically-informed null — pseudo-lexicons preserving per-language phonotactics, segment frequency and syllable structure — and (b) a cognacy control that removes inherited and borrowed matches.* If the residue is indistinguishable from the null, endolinguistics has no distinct object at this level; if it is significant and concentrated in specific semantic domains, that residue is the object. Three baselines are prerequisites and currently missing: the chance-collision rate for resonance (per-concept, under the empirical nucleus distribution), the split-half ceiling and cross-nucleus floor for $q(C) \neq q(\iota C)$, and the expected closure-failure under the phonotactic null. Two reframed hypotheses follow directly: H-1 (systematic resonance) — cross-macrosystem code→field coincidences exceed the chance-collision baseline and the excess is domain-concentrated, not uniform; H-2 (emergent specular homology) — mirror pairs `AB`↔`BA` across independent lineages in a shared domain exceed each lineage's own ordering statistics (as in romance T·R vs germanic R·T for 'earth', which historical linguistics shows are *not* cognate). The full programme, controls, and evidence grades are in `docs/planning/01-objeto-preguntas-metodologia.md`.

Hypotheses. H-1 closure-failure $\propto$ ternary morphology; H-2 ternary attractor $\propto$ root-template rigidity; H-3 derivation raises arity (arity gradient in word families); H-4 the substitution order predicts the direction of held-out diachronic change; H-5 nucleus histograms are cognate-free phylogenetic fingerprints.

Questions. Q-1 do hole locations carry areal signal? Q-2 is C↔ιC distance predictable from the centre class? Q-3 is there a natural canonical-form functor? Q-4 what scalar potential makes substitution acyclic? Q-5 do the sonority, decay, and pivot orderings of Ω coincide into one latent scalar?

Falsifiable predictions. (P1) On a held-out diachronic corpus, >50 % — indeed near-100 % — of class substitutions run downhill on L>S>K>N>P>M>T. (P2) Deverbal nouns average higher arity than their verbs. (P3) Analytic/tonal languages have significantly higher code-complex b_1 than synthetic ones. (P4) Per-family association matrices reproduce the same-zone repulsion sign for P·M and L·N. (P5) A single per-class scalar satisfies all 21 substitution inequalities (the tournament is 1-dimensional).

10.5 Control results: what survives a serious null

The controls specified above have now been run (scripts in `src/`; full detail in `docs/planning/04-resultados-controles.md`). They are reported here in full, whether they confirm or refute the paper's claims, because that is the point of running them.

Phonotactic null (§4, §3.4). A per-language segment-level null (order-1 Markov chain over IPA tokens, preserving each language's segment inventory, frequency and local phonotactics; class co-occurrence *emerges* from it, so the test is non-circular), run on the 104,674-form / 107-language / 20-family NEL subcorpus (300 replicates):

- The same-zone (OCP) repulsion is recovered phonology, not an endolinguistic residual — established by

testing it against *two* nulls. A segment-tier bigram null (interleaving C and V) reproduces only a fraction (NEL null +0.148 vs obs +0.359; full corpus +0.022 vs +0.412, z = +102), which *looked* like a large residual. But that null is structurally incapable of modelling the effect: a nucleus is a set of *non-adjacent* consonant classes, and a first-order chain over interleaved segments has no memory across the intervening vowel, so for class co-occurrence it collapses to near-independence sampling — it even produces within-zone *attraction* (null +0.07/+0.17 vs observed −0.08/−0.10), a sign it is not representing dissimilation at all. A consonant-tier null (bigram over the consonant sequence, skipping vowels — which *can* model long-distance consonant co-occurrence) reproduces the effect almost exactly: on NEL, observed OCP gap +0.359 vs null +0.357 ± 0.019 (z = +0.11, p = 0.85), recovering the within-zone repulsion *sign* (null −0.089 vs obs −0.081). At full-corpus scale the consonant-tier null reproduces 84 % of the gap (obs +0.412 vs null +0.347 ± 0.004); the residual 16 % (z = +14.8, inflated by N = 5,608) is precisely what a *bigram* cannot capture of genuinely long-distance (tri-consonantal-root) OCP and would shrink further under a consonant trigram — it has no semantic content and remains Level-II phonology. The same-zone repulsion is therefore long-distance OCP-Place — a known Level-II phonological universal (Frisch, Pierrehumbert & Broe 2004) fully captured by consonant phonotactics, exactly as §1.2 and §11 classify it. It carries no semantic variable and is not a candidate for the endolinguistic object. An earlier reading of this result as an "endolinguistic residual" is withdrawn.

- The per-language closure failure gives no support to "the nucleus is a whole." On the dense NEL subcorpus,

observed 34.6 % $\approx$ null 36.0 % $\pm$ 4.2 ($z = -0.34$, $p = 0.82$) — indistinguishable from sampling sparsity. On the full corpus, observed closure-failure (68.0 %) is in fact *lower* than the phonotactic null (81.3 % $\pm$ 0.35, $z = -38.1$): real languages preserve their binary faces more than phonotactically-matched pseudo-data, not less. Either way the effect does not run in the direction the "irreducible whole" reading needs. Thesis T1's closure-based leg is withdrawn; that reading must rest on the token-vs-nucleus and attractor evidence, not on closure.

Resonance baseline (§8.2). Against a permutation null that breaks the concept$\leftrightarrow$nucleus link, the 9.34 % cross-macrosystem match is 2.15$\times$ chance (null 4.34 % $\pm$ 0.04; the $z = +133$ is inflated by the enormous N and is *not* the effect size — the honest figure is the 2.15$\times$ ratio): the excess is above chance-collision but confounded, not a proven endolinguistic resonance. (Caveat: the excess still folds in universal sound-symbolism and areal effects; separating *endolinguistic* resonance from *typological universal* needs the phonotactic null and further controls.)

Inversion ceiling/floor (§5.1). The mean Jaccard(C, ιC) = 0.243 sits near the cross-nucleus floor (0.215) and well below the split-half reliability ceiling (0.322). But two further controls deflate it. (i) Inversion is barely special: Jaccard(C, ιC) = 0.243 vs a generic same-nucleus reordering Jaccard(C, C') = 0.255 — inversion separates concepts only 0.01 more than *any* reordering, and all sit near the floor (0.216), so the low overlap is largely the trivial fact that a different consonant sequence is a different *word*. (ii) Order adds no cross-family concept-information: an H-0 test on *ordered* codes carries *less* concept-signal-beyond-null than the unordered nucleus (0.95 % vs 1.05 % of concept entropy, `src/inversion_specialness.py`, 'discriminant_h0.py –code ordered'). So q(C)$\neq$q(ιC) is real but weak: order separates words trivially, inversion is only marginally distinguished, and order carries no extra endolinguistic meaning beyond the nucleus.

Cognacy control (§6, §8). On IE with PIE cognate sets (IECoR), nucleus conservation under genuine descent (same cognate set) is 22.8 %, versus 3.6 % for non-cognate same-concept pairs (6.4$\times$) — but 83 % of the same-concept comparisons the paper counts as "diachrony" are between non-cognate forms (lexical replacement, *canis*/*perro*). The substitution order of §6 is diachronic only on the cognate subset; the rest is synchronic coincidence. The coherent three-level picture: non-cognate $\approx$ 3.6–4.3 % (chance), cross-macro resonance 9.34 % (real, intermediate), genuine descent 22.8 % (inheritance). Furthermore (R3), the substitution *order itself* correlates $\rho = 0.93$ with the anchor/target class-frequency ratio, and its perfect transitivity is what any scalar ranking trivially yields — so the "arrow of cualic decay / Λ source" (§6) is substantially a frequency/anchor artifact, not a diachronic law.

Corrected statistics. The §4.2 same-zone repulsion, tested by a valid label-permutation test (not the invalid Welch t on dependent pairs), is significant at $p = 0.001$. The §5.3 chirality contrast (with-Λ H_norm 0.960 vs without-Λ 0.982) has a small absolute magnitude ($\Delta \approx 0.022$ normalized-entropy units) but a large, consistent standardized effect (Cohen's $d = 1.56$); it is a small systematic tilt, not a "law." The $n = 7$ PCA of §9.4 is withdrawn as unstable/circular, retaining only the 1-dimensional substitution tournament (P5).

Net effect. Under proper controls, none of these analyses isolates a clean endolinguistic residual. The OCP is fully reproduced by a consonant-tier phonotactic null (recovered phonology, withdrawn as a candidate); concept-specific resonance sits above chance-collision but remains confounded by universal sound-symbolism and areal diffusion; closure is sampling sparsity; and most "substitution" is non-cognate lexical replacement. What survives is representational: the OAS reduction faithfully re-expresses known phonological and frequency structure. Isolating the endolinguistic object proper requires the semantic (code$\rightarrow$field) discriminant of §10.4-H-0 — a phonologically-informed *and* cognacy-controlled test of code$\rightarrow$meaning association — which is not yet run. This is the smaller, more skeptical, and firmer paper the review asked for: an instrument validated, and the object's discriminant now specified rather than prematurely claimed.

11. Conclusion

Treated as one mathematical object over 4,721 languages, the OAS class-projection is combinatorially saturated (the per-language closure ”failure” is sampling sparsity, §10.5), carries a same-zone dissimilation that a consonant-tier null shows to be *recovered* long-distance OCP-Place phonology (§10.5), is organized by order (inversion re-orients concept fields; Zipf appears only under order), exhibits a cycle-free synchronic modal-nucleus substitution order (not established as diachronic decay, §10.5), and is positionally organized by the liquid Λ as pivot against the velar X as shell. These are regularities not previously visible in code-by-code reading — mostly recovered phonology and frequency structure, as the controls make explicit; one prior intuition (deverbal reduction) is falsified; and each result now comes with a phonologically-informed control. The OAS system, introduced as a discretization of the sensible articulatory space, turns out to carry — at corpus scale — a topology, an order, and a dynamics of its own.

Two caveats bound these words precisely. First, several of the regularities above are recovered phonological universals (the OCP of §4, sonority sequencing of §5.3, positional fortition of §9.2, liquid instability of §6): they establish that the OAS reduction is *representationally faithful*, which is a real result about the instrument, but they are not by themselves evidence for a semantic code layer. Second, the diachronic numbers of §6/§8 are, absent a cognacy control, synchronic modal-nucleus coincidences rather than proven class substitutions. The claim this paper does *establish* is therefore the narrower and firmer one: projected onto the seven OAS classes, 4,721 languages yield a saturated combinatorial complex whose class-relational, ordinal, and directed structure is measurable and largely inherited from phonology and frequency — and the controls that would isolate any endolinguistic residue beyond that are now specified (§10.4). Establishing that residue is the object of the continuing programme, not a conclusion of this paper.

Appendix A. Reproducibility

All figures and statistics derive from `data/db/endolang.db` and the scripts in `docs/figures/src/`; the consolidated numeric findings are in `docs/paper-findings.md`. Analyses used exact enumeration and hand-implemented linear algebra where a numerical library was unavailable, cross-checked against the catalog tables (56 nuclei = 21 + 35; 252 ordered codes; 4,588 skeletons).

Appendix B. Attribution

The endolinguistic discipline was founded by Christiane S. Meulemans and José Ángel Elias Fernández y Cuevas. The OAS system (*Originating Affective Sound* — the seven affective sound classes and the code-first analysis built on them), the qualic mathematics (the axiom $q(C) \neq q(\iota C)$), and the falsifiability protocol used here are the present author’s original additions to that tradition; they did not exist before. This paper is an empirical, mathematical study *within* the discipline, not a claim to found it.