Additive Structure of Phonological Correspondences: A Manual

Thesis, methodology, operations, and findings — with worked examples in real words

Alejandro Toledo Martínez — Independent researcher (ORCID: 0009-0000-1277-9697)

How to read this book. Most chapters follow the same teaching flow, signposted by bold headings: the motivating question, the intuition in plain language, the formal definition, a worked example in real words with IPA, the result in Indo-European and Austronesian, the limits (what *not* to conclude), and how to reproduce it. Start with Chapter 0, which walks the whole method by hand on real words. A standing epistemic rule, stated once and kept throughout. During the *discovery* phase this project uses no protoforms, no known sound laws, and no OAS classes as inputs. OAS is deliberately withheld: it would inject psychodynamic factors into a phase whose whole point is to find structure *before* interpretation. Like reconstructions and historical laws, OAS is reserved as an external partition for later contrast, never a coordinate of discovery. *Family-blind* comparison is likewise a future phase (aimed at the "universal" questions we are deliberately *not* asking yet), not a present result.

0.1 We start with words, not theories

Take one meaning — 'two' — and collect how a few Indo-European languages actually say it, each written in IPA (the phonetic alphabet). Nothing is reconstructed; these are attested, documented forms:

language	'two' (IPA)
Bulgarian	dva
Ancient Greek	duo
Serbo-Croatian	dva
Armenian (Eastern)	jerkʾu
Armenian (Western)	jergu
Yiddish	cvej

These words are coderivatives: forms with a shared, real history, worn differently by each language — recognizable because their sounds correspond systematically. *(We say* coderivative *rather than* cognate *deliberately; §3.1 explains the word and why it matters.)* A historical linguist would now reconstruct their common ancestor and derive each form from it. We will not — not because we doubt there was one, but because we want to study the forms' relationships *directly*, and bring reconstruction back only later, as an independent check. We will only compare the forms *to each other*.

0.2 Line two words up, letter by letter

Take the two Armenian varieties. Write their sounds in columns so that matching sounds sit together (this is called an alignment):

```
Eastern:   j   e   r   kʔ   u
Western:   j   e   r   g    u
```

Read down each column. Four columns are identical (`j`, `e`, `r`, `u`). Exactly one column differs: `kʔ` in the East, `g` in the West. That single difference is the raw material of everything that follows.

0.3 Say *what* differs, not *which becomes which*

`kʔ` and `g` are both sounds made at the back of the mouth (velars). The question a phonetician asks is: *in which properties — which "features" — do they differ?* Line up their features:

feature	`kʔ`	`g`	differ?
voiced (vocal cords vibrate)	no	yes	?
velar (back of mouth)	yes	yes	—
stop (full closure)	yes	yes	—
… (all others)	=	=	—

They differ in exactly one feature: voicing. So we write this correspondence as

$$\text{`k}^{\text{ʔ}}\text{`} \sim \text{`g`} \;=\; \{voi\}.$$

That object — the *set of features that differ* — is what we call an operator. Notice what it is not: it does not say the sound "became voiced" or "became voiceless." It is symmetric — it only says *these two sounds differ in voicing*. Direction (who came first) is a separate, later question. This single decision — describe a correspondence by *which features differ* — is the whole idea of the book.

0.4 A second example, with two features

Now 'tooth', and two more languages:

language	'tooth' (IPA)
Armenian (Eastern)	`atʔam`
Armenian (Western)	`adam`
Breton	`dant`
Modern Greek	`ðodi`

Align Eastern vs Western Armenian: `a–tʔ–a–m` against `a–d–a–m`. Again one column differs: `tʔ  d`, and again the only feature is voicing → `{voi}`. The same operator, **`{voi}`**, shows up in a different word. That recurrence is the point: operators are not one-offs; they *repeat* across the vocabulary.

Compare the Greek `ð` (the "th" in *this*) with a `d`: they differ in a single feature — a `d` is a full stop, a `ð` lets air flow (continuant) — so `d  ð = {cont}`, another atom (a spirantization). Now a two-feature operator: compare a plain voiced stop `d` with a `z`. They differ in two features — continuancy (`z` lets air flow) *and* stridency (`z` hisses, `d` does not). So

$$\text{`d`} \sim \text{`z`} \;=\; \{cont + strid\}.$$

An operator can bundle one, two, or three features. `{voi}` and `{cont}` are atoms (one feature); `{cont+strid}` is a molecule (several).

0.5 Do this everywhere → the *repertoire*

Repeat §0.2–0.4 for every pair of coderivative words in every meaning across the whole family. Collect all the operators you find, keeping the ones that recur often enough to be real (not alignment noise). The resulting set is the family's repertoire, written O. For Indo-European it has 67 distinct operators; for Austronesian, 22.

Here are Indo-European's atoms — its one-feature operators — each shown as the *real sound pairs* that realize it in the data:

atom	what it is	real correspondences
{voi}	voicing	d t, g k, b p, s z, f v
{cont}	stop ↔ fricative	t θ, d ð, k x, g ɣ, b β
{ant}	front ↔ back sibilant	s ʃ, z ʒ
{strid}	plain ↔ strident	s θ, z ð, v β
{hi}	palatalization / raising	l lʲ, n nʲ, k q
{back}	palatal ↔ velar	c k, cʰ k

(Two rarer atoms, {son} and {cor}, are left out of this table for brevity; Chapter 8 works with all eight.)

A reader who knows historical linguistics will recognize t θ, d ð, k x — this is the shape of what is traditionally called the Germanic consonant shift (Grimm's Law). We did not put it there. It *fell out* of comparing attested forms. Whether it matches Grimm's Law is a question we ask *afterward* (Part V), never an assumption we feed in.

0.6 Operators can be *added* — and the sum often lands back inside

Here is the first genuinely mathematical observation, and it is easy to see by hand. Suppose a family has the operator b p $= \{voi\}$ and the operator p f $= \{cont + strid\}$. What is the difference between b and f? You *combine* the two: whatever changes once stays changed, whatever changes twice cancels. Here nothing cancels, so

$$\{voi\} \ + \ \{cont + strid\} \ = \ \{voi + cont + strid\},$$

which is exactly the operator b f. Combining operators is just XOR of feature-sets (symmetric difference). The striking empirical fact — the spine of this book — is that when you combine two operators from a family's repertoire, the result is *another operator already in the repertoire* far more often than chance would predict. The repertoire is not a random grab-bag; it is additively organized. Measuring exactly how much, and against what "chance" means, is what Part III does.

0.7 What the rest of the book does

Everything from here is careful measurement of the object we just built by hand:

- How big is the space the repertoire lives in, and how much does it use? (Chapters 7, 9)
- Is the additive organization real, or an accident of the sounds involved? (Chapters 10, 17)
- Which operators are *missing* — combinations that could exist but don't? (Chapter 12)
- Do related branches of a family share their repertoires more than random groups of languages? (Chapter 15)

If you followed §0.1–0.6, you already understand the method. The rest is rigor.

PART I — The idea and where it stands

1. The problem: patterns traditional linguistics has not given us

The question. Historical linguistics is one of the most successful sciences of the humanities. It reconstructs ancestral languages, states sound laws, and draws family trees, all from patient comparison of documented forms. So what is *missing*? What could an explicitly *mathematical* description of sound correspondences add that the comparative method has not already given us?

The intuition. The comparative method answers questions of the form *"how did these forms descend from their common ancestor?"* It is a method of inference toward a hidden past. But there is a different question it was never designed to ask: *"treated purely as a set of differences, what shape does the space of attested correspondences have?"* Is that space full or sparse? Does it cluster? Are some differences combinations of others? Do two families occupy the same region of the space, or different regions? These are questions about the geometry and algebra of the differences themselves, independent of which form came first.

The reframing. Write each sound correspondence not as *"segment a becomes segment b"* but as the set of phonological features in which a and b differ. That object — call it $\Delta(a, b)$ — is a vector. A whole family, then, contributes a *set* of such vectors: its repertoire O. Once we have a set of vectors we can ask genuinely new questions: its dimension (rank), how much of the space it can generate it actually uses (occupancy), whether combining two of its elements tends to land back inside it (additive structure), and which generable differences it pointedly avoids (the *holes* $\langle O \rangle \setminus O$). None of these are questions the reconstructive order poses, because that order is busy with a different task.

The question genealogy cannot pose — *why this form, and not that one?* Sound laws tell us, with great precision, *that* a form changed and *what* it became. They do not tell us *why* one related system selected one outcome while another selected a different one — why, from a shared source, this lineage keeps one part of a word and that lineage keeps another. Genealogy names the divergence; it does not explain the *selection*. That "why" may depend on things a tree has no room for — substrates, contact, areal pressure — because a language is less a *branch* off one parent than a new formation, confluent from many sources (§3.1). The "why" only becomes a *measurable* question once we can see the whole space of what a system *could* have done and ask which parts it used. The holes $\langle O \rangle \setminus O$ — operators that were available yet went unchosen — are exactly that question, made quantitative. This book does not answer the "why"; it builds the instrument that lets it be asked. (Chapter 3.1 explains why, for the same reason, we speak of *coderivatives* rather than *cognates*.)

The limit, stated up front. This is not a replacement for the comparative method, and it is not a claim that "sound change is XOR." It is a change of *order of inference* (Chapter 3). The mathematics describes attested differences; whether any of it lines up with reconstructions, laws, or family trees is a *separate* question, asked afterward, with objects that never entered the computation.

What you will be able to do. By the end of this book you will be able to take any language family with IPA-segmented word lists, compute its correspondence repertoire without reconstructing anything, measure its geometry and additive structure against explicit null models, and compare families — all reproducibly, one `make` target at a time.

2. The thesis in plain language

The one-sentence thesis. *From documented forms, phonological correspondences produce sparse repertoires whose additive, geometric, and distributional organization can be measured without taking an*

ancestral reconstruction as the origin of coordinates, and only afterwards contrasted with the regularities historical linguistics proposes.

Unpacking it.

- *From documented forms* — the inputs are attested words with IPA segments, nothing reconstructed.
- *Sparse repertoires* — of everything the feature representation could express, a family uses very little… but, as we will see (Chapter 9), that sparsity is mostly an artifact of how vast the algebraic space is; measured against what the corpus actually makes available, the repertoire is fairly *dense*.
- *Additive organization* — combining two operators (XOR of their feature-differences) lands back inside the repertoire far more often than chance (Chapters 6, 10). The set behaves like an approximate subspace.
- *Measured without protoforms* — the computation uses only attested-to-attested alignments.
- *Contrasted afterwards* — reconstructions, laws, and family trees are brought in only to interpret and test what was found, never to build it.

A first taste in real words. Between Eastern and Western Armenian for 'two' (*erku*), `kʔ` corresponds to `g`: the single feature that differs is voicing, so the operator is `{voi}`. Between a Greek `d` and a Slavic `z` for 'tooth', two features differ — continuancy and stridency — so the operator is `{cont+strid}`. These are not directions of change; they are symmetric signatures of difference. The whole method is built on cataloguing such signatures and studying the set they form.

What "new" means here. The novelty is not a new sound law. It is a new *object of study* — the repertoire as a measurable region of a feature space — and a new *order of operations* that keeps the historical apparatus out of the measurement so it can serve as an honest test afterward.

3. Protoform-agnostic *by design*, not by omission

The question. Every earlier draft of this work apologized for "working without a proto-language," as if it were a limitation to be excused. Is it a limitation — or is it the point?

The intuition. Protoforms are *inferred* objects. They are excellent historical models, but they are not independent observations: they are reconstructed *from* the documented languages using assumptions about how sound change works. If you build your mathematical space *on* protoforms and then ask whether that space confirms the regularities of historical linguistics, you risk circularity — in machine-learning terms, leakage of the target into the inputs. You would be testing the theory against a space the theory helped build.

The inversion, formally. The traditional reconstructive order runs

$$X_{\text{documented}} \longrightarrow \widehat{Z}_{\text{proto}} \longrightarrow H_{\text{historical}},$$

where $\widehat{Z}_{\text{proto}}$ is a reconstructed ancestor and $H_{\text{historical}}$ the proposed changes. This project runs instead

$$X_{\text{documented}} \longrightarrow \mathcal{M}(X), \qquad \text{then, separately,} \qquad \mathcal{M}(X) \longleftrightarrow H_{\text{historical}},$$

where $\mathcal{M}(X)$ is a mathematical structure discovered with no protoforms and no precoded laws. The question shifts from *"how did these forms descend from the proto-form we reconstructed?"* to *"what regularities emerge from the documented forms, and how far do they coincide with, diverge from, or extend the regularities historical linguistics proposes?"* That is not a denial of reconstruction; it is a more demanding way to test it.

Not ”theory-free.” The method is not innocent of assumptions. Transcription, segmentation, concept assignment, the feature matrix, and the alignment algorithm are all measurement instruments, and their influence must be made visible and probed with sensitivity analysis (Chapter 5). The claim is not access to raw data; it is the separation of measurement decisions from the historical hypotheses we later wish to examine. And the discovery phase withholds three things on principle — protoforms, sound laws, and OAS classes — so that each can serve later as an *external* yardstick rather than a hidden input.

The consequence for the whole book. Everything in Parts II–IV is computed under this rule. Part V is where the reserved apparatus (history, and eventually OAS) is allowed back in — as contrast, never as coordinate.

3.1 Why we say ”coderivative,” not ”cognate” — a note to the comparative linguist

Throughout this book we call two related forms coderivatives, not cognates. The choice is deliberate, and we owe the comparative linguist a plain explanation, because it is not cosmetic relabelling but a statement of what we are trying to see. We ask for a careful reading here; disagreement is welcome — that is how science works — but we want at least to be understood.

What we are *not* doing. We are not claiming that proto-languages did not exist, and we are not questioning the comparative method or the reconstructions it has produced. Those are among the great achievements of the human sciences, and improving them is not our project; it is not our terrain to litigate. We take the historical reality of language relatedness as real: when the sounds of two forms correspond systematically across many words, that regularity is not an accident, and we honour it. On this we stand *with* the comparativist.

What the inherited vocabulary quietly assumes. The words the field inherited are *family* words: a cognate is, etymologically, a ”co-born” relative; languages are ”sisters,” with ”mother” and ”daughter” tongues descending from a single ”parent.” This metaphor encodes a specific and powerful model — a tree, with one ancestor at each node — and it has taken the field remarkably far. But a metaphor is also a fence around the questions one can ask. A tree has, by construction, a *single* source at each split; it has no natural place for substrate, contact, borrowing, or the coexistence of several lects. (Recall that Classical Latin was once taken for *the* Latin, until Vulgar Latin and the regional lects had to be let back in.) The family tree answers *from whom did this descend?* It is not built to answer *why did this system select this form, out of those available, while that system selected another?*

A deeper point: a language is not the linear descendant of one ancestor. Here is the claim that most moves us past the tree, and we make it carefully. A language is not simply the evolved continuation of a single predecessor. It is a new formation — not from nothing (there are always materials, antecedents, continuities; this is emphatically not *ex nihilo*), but a synthesis of many inputs at once: the lects that preceded it, the substrates left behind by populations who shifted into it, the neighbouring languages it lived in contact with, the cultural and social pressures that shaped it. To say that language B ”comes from” language A, as a single line of descent, mistakes a confluence for a branch. The substrate, in particular, is not a smudge on an otherwise clean inheritance — it is constitutive: a community that adopts a language re-makes it through the language it is leaving, so the new system is, in part, co-authored by what it replaced.

We are careful not to caricature the comparative method here: it is not unaware of any of this — borrowing, areal convergence, the wave model, *Sprachbünde* are all part of the field’s knowledge. The point is subtler. The tree’s *core inferential machinery* treats contact and substrate as perturbations to be filtered out so that a clean line of descent can be recovered; the tree stays the backbone, and confluence is the correction made around it. Our wager is the reverse — that the confluence is not

noise around a descent but part of what a language is — and that a mathematical, network-shaped description can hold that view *natively*, where a tree, by its very form, cannot.

We are not the first to distrust the tree — and this is where we go further. Doubting a purely arborescent history has a long and respectable lineage, and we lean on it gladly, *as support*. Johannes Schmidt's wave model (1872) already set continuous diffusion against Schleicher's family tree; Hugo Schuchardt, arguing against the Neogrammarians' exceptionless sound laws, held that mixture is everywhere — that there is no wholly unmixed language; Trubetzkoy's Sprachbund (1928) made convergence among unrelated languages a first-class object; and modern computational work has made the network explicit — François's linkages (2014), where subgroups intersect and no single tree fits, and phylogenetic-network studies that recover massive hidden borrowing within Indo-European itself (Nelson-Sathi et al. 2011; List et al. 2014). Our debt to this line is real. But notice what most of it still does: it *adds horizontal edges to a framework of descent* — borrowing detected as deviation from a tree, a network of who-transmitted-to-whom. We take one step further. We do not build descent at all during discovery; we study the geometry of coderivation directly, we treat confluence and substrate as constitutive of a language rather than as edges laid over an inheritance, and we turn the unanswered *why-this-form-and-not-that* into a measurable question about the region a system occupies. The wave, the mixture, the Sprachbund, and the network are our allies; the object we measure and the question we put to it are our own.

The gap this leaves. So sound laws are magnificently precise about *that* a change occurred and *what* forms resulted, and silent about *why* a system selected one outcome where a related system selected another. From a word for 'water' — say something shaped like *wodr* — why does one system keep the front (giving *wa-*) and another the coda (giving *-udr*)? "Independent evolution" *names* the fact; it does not *explain* the selection, which may turn on exactly the substrates and confluences a tree cannot represent. To pose that as a *measurable* question we need more dimensions than descent provides — and that is where a mathematical description earns its place.

Why "coderivative." A coderivative is a form co-derived with another — sharing a real, evidenced history of derivation — where the *evidence* is the regularity of correspondence itself, and where we deliberately do not commit, in the analysis, to a single reconstructed ancestor as the sole source of that shared history. It keeps everything the comparativist rightly insists on — relatedness is real, and it is evidenced by systematic correspondence — and it sets aside only the one part of the family metaphor that *pre-decides the shape of the answer*. A cognate, in this light, is the special case of a coderivative whose shared history is a clean descent from a single ancestor; the term "coderivative" simply also leaves room for the cases where it is not — borrowing, substrate influence, convergence, a network rather than a line.

Why this matters now. This is, frankly, an attempt to do for the study of sound history a little of what modern methods did for other historical sciences. Genetics did not discard heredity; it *re-described* it in a molecular, combinatorial, quantitative framework that could ask questions the older language of "blood" and "stock" could not even frame. The study of language may be near a similar threshold: the family-tree vocabulary, indispensable as it has been, can reach the limit of what it resolves — just as mathematical, computational, and representational tools have become powerful enough to see a little further. This book is an early step onto that path. It is not a refutation of what came before; it is the addition of a dimension to it. Renaming the cognate a coderivative is the smallest honest marker of that intent.

PART II — The instruments (so a reader can reproduce)

4. The data

The question. What exactly goes in, before any mathematics?

The sources. Three, all open and reconstruction-free at the point of use:

- Lexibank — a large aggregation of word lists across families, with every form already segmented into IPA (`Segments`). This is the primary discovery corpus. We pilot on Indo-European (304 languages available) and Austronesian (978), taking, per family, the languages with the most forms.
- IE-CoR (iecor) — Indo-European with expert cognacy and documented borrowings. We never use its cognacy to *build* the repertoire; we use it only to *validate* the statistical pipeline (Chapter 5) and to run the corpus-regime controls (Chapter 17).
- Glottolog — the genealogical classification, used only to label branches (Chapter 15), never as a computational coordinate.

The units. A concept is a comparison meaning (Concepticon's "TOOTH", "TWO"…). A form is one language's word for a concept, given as an IPA segment list. A segment is one IPA sound. That is all the discovery phase consumes.

The caveat about "family." In this pilot, "Indo-European" and "Austronesian" are *genealogical labels taken as first pilot boundaries*. They are a convenience, not a result; Chapter 18 discusses doing without them. Nothing in Parts II–III depends on the labels being "correct"; they only delimit which forms are pooled.

Reproducibility. Exact dataset versions, language counts, concept counts, and thresholds are logged in Appendix C; every table in this book names the `make` target that regenerates it.

5. From form to operator

The question. Given two words for the same concept in two languages, how do we extract *correspondences* without knowing in advance which sounds correspond?

The pipeline, in four steps.

1. Segmentation is already given by Lexibank (IPA `Segments`).
2. Statistical coderivation. We cluster forms of the same concept into likely coderivative sets — what the literature calls *cognate sets* — with LexStat (LingPy), which learns language-pair-specific sound-correspondence scores and clusters by them. No reconstruction, no expert judgment — just recurrent phonetic correspondence (§3.1, §5.1).
3. Alignment. Within a coderivative set, we align each pair of forms with Needleman–Wunsch, using a substitution cost equal to the fraction of panphon features in which two segments differ. Aligned positions are our candidate correspondences.
4. The feature matrix. Each IPA segment maps, via panphon, to a vector of phonological features (continuant, voice, nasal, coronal, labial, …). This is the representation on which every later measure is built.

The instruments are not neutral — so we test them. Every step is a measurement choice. Two matter most, and both are audited later: the feature set (Chapter 8 shows results split into basis-dependent vs invariant; Chapter 13 handles panphon's *ternary* values) and the cognacy threshold (Chapter 5's validation, and the corpus regimes of Chapter 17, quantify how much statistical cognacy contaminates the result). We find (Chapter 17) that LexStat is *conservative but clean*: at the level of operator types its precision against expert cognacy is near 1.0, even though at the level of individual pairs it is only 0.73.

A worked alignment, by hand. Take 'three' in Ancient Greek `treis` and Romanian `trei`. The aligner lays them out to minimize total feature-distance:

```
Greek:     t   r   e   i   s
Romanian:  t   r   e   i   –
```

Four columns match (`t`,`r`,`e`,`i`); the last is a *gap* (Greek has a final `s`, Romanian does not). Gaps are not operators — they are insertions/losses, recorded separately. Now Bulgarian `tri` vs Yiddish `draj`:

```
Bulgarian:  t   r   i
Yiddish:    d   r   a   j
```

Column 1 gives `t d` = `{voi}` (our familiar atom); column 3 gives a vowel correspondence `i a`; the final `j` is again a gap. Every aligned, non-identical, non-gap column is one operator instance. Run this over all coderivative pairs and you have counted the whole repertoire — mechanically, with no reconstruction anywhere.

One thing this alignment cannot see: metathesis. Needleman–Wunsch is *monotonic* — it keeps order — so when two forms differ by a reordering of segments (the classic TR↔RT, or *ask*↔*aks*), it cannot represent that as a single move. It will instead fake it with substitutions and gaps, injecting noise. Such pairs align *badly* (high cost), so the cohesion screen filters the worst cases; but partial metathesis leaks. This is not a small patch to make — metathesis is a permutation of the skeleton's positions, a different *kind* of operator: it lives in the group of reorderings (S_k), orthogonal to the feature space $\mathbb{F}_2^n$ where our operators live. Treating it as a first-class, second dimension of change (feature-change *and* position-permutation) is a natural and important extension — it is exactly the mirror-inversion the skeleton view was built to catch — and we flag it as such rather than hide it inside the noise. (See the research programme.)

Reproduce. `make family FAMILY="Indo-European"` runs steps 1–4 and writes the correspondence table; `make cognate-eval` runs the validation against IE-CoR.

5.1 Coderivation without reconstruction — why the method is built this way

A comparative linguist will raise a sharp objection, and it deserves a straight answer: *a correspondence is only meaningful between cognates — forms descended from one ancestor — so aligning same-meaning forms that may not be cognate yields noise, not correspondences.* Everything in this book turns on how we answer it.

The answer is that the evidence that two forms are cognate has always been the regularity of their correspondences — that their sounds line up the same way across many words. The reconstructed ancestor is an *inference built on top of* that regularity; it is not the evidence itself. This method keeps the evidence and suspends the inference: it detects cognacy by recurrent, systematic correspondence (that is what LexStat measures), which is exactly the Neogrammarian criterion — and it declines to posit a reconstructed protoform. We do not abandon comparative rigour; we keep its foundation and set aside only its final, reconstructive step.

This is not "anything goes." A cluster earns its place by *showing* recurrent correspondence; a meaning-matched set with no such regularity — unrelated words for 'fire', say — fails our own criterion, not merely the reconstructionist's. So when we prefer one worked example over another, it is because one exhibits a legible, recurrent correspondence and the other does not — never because a reconstruction certifies it.

Why not simply filter the input down to expert-certified cognates and be safe? Because that would defeat the experiment. If we cleaned the data *with* reconstruction and then asked whether the resulting structure matches historical regularities, we would be testing history against a history-filtered dataset — the answer leaking into the inputs. Withholding reconstruction from the *input* is precisely what lets the later comparison with reconstruction be an *independent* test (this is the whole point of Chapter 3).

Finally, the residual risk — that statistical clustering still groups the occasional look-alike — is not assumed away; it is measured. Chapter 17 permutes concept labels to destroy all cognacy and meaning (the D_R control) and reports what survives; the two-level validation there shows that although pair-by-pair cognacy is noisy (precision ≈ 0.73 against expert judgments), the operator *inventory* is far more robust (≈ 1.0). Contamination is a controlled variable, not a hidden flaw. That is the shield: we keep the comparative method's foundation, set aside only its reconstructive conclusion, and let explicit controls say how much the residue matters.

6. The operator $\Delta(a, b)$

The question. What, precisely, is an "operator" in this book — and what is it *not*?

The definition. For two aligned segments a, b, the operator is the set of primary features in which they differ:

$$\Delta(a, b) = \{\, k : \phi_k(a) \neq \phi_k(b) \,\},$$

where ϕ is the panphon feature map. Equivalently, over binary features, $\Delta(a, b) = \phi(a) \oplus \phi(b)$, the XOR of the two feature vectors. We keep operators of one to three features (single changes and small bundles), which are the phonologically interpretable ones.

Worked reading. `d` `z` differ in continuancy and stridency $\Rightarrow \{cont + strid\}$ (a spirantization signature). `t` `s` $\Rightarrow \{cont + strid\}$; but `t` `f` $\Rightarrow \{cont + cor + strid + lab\}$ — *not the same operator*, because `f` is a labiodental and differs from `t` in place as well. The feature representation makes the distinction automatically; lumping both as "lenition" would erase it.

What the operator is NOT. It is symmetric: $\Delta(a, b) = \Delta(b, a)$. It does not distinguish $t \to s$ from $s \to t$. It is a *differential signature*, a contrast — not a directed historical change. The frequently-repeated fact that "each operator is its own inverse" ($\Delta \oplus \Delta = \varnothing$) is a *consequence of this symmetry*, not a phonological discovery. Directed history, when we want it, needs a richer object,

$$T = (a, b, c, p, \ell_1, \ell_2, w) \quad \text{with signature} \quad \sigma(T) = \Delta(a, b),$$

carrying source, target, context c, position p, the two languages, and a weight — so that $T_{t\to s} \neq T_{s\to t}$ even though their signatures coincide. That object belongs to Part V; Parts II–IV study the symmetric signatures.

Reproduce. The operator of any segment pair is `delta(a,b)` in `src/algebra.py`; the family's operator set is what `make algebra FAMILY="X"` reads.

6.1 The same operators live *inside* a single language

(Illustrative interlude. The examples in this section are drawn from well-known grammatical descriptions, not computed from our reconstruction-free corpus. They are here to teach one thing: the operators the method finds between coderivatives across languages are the same objects that alternate inside a single grammar — in plurals, verb paradigms, and derived words. Sound-difference is sound-difference, whether it sits between Greek and Armenian or between two forms of one English word.)

Our operators compare two *languages*. But the very same feature-differences show up when a *single* language builds or inflects a word. Watch `{voi}` — the one-feature voicing operator from Chapter 0 — appear again and again, now *within* one grammar:

language	construction	alternation	operator
English	plural of *wife* → *wives*	`f` `v`	`{voi}`
English	noun *house* `haʊs` → verb *house* `haʊz`	`s` `z`	`{voi}`
German	*Tag* `ta:k` → plural *Tage* `ta:gə` (final devoicing)	`k` `g`	`{voi}`
Dutch	*huis* `hœys` → plural *huizen* `hœyzə`	`s` `z`	`{voi}`

The plural of *wife*, the verb *to house*, the German declension — three unrelated pieces of grammar, one and the same operator `{voi}`. This is why the repertoire is not an artifact of comparing languages: it catalogues feature-differences that grammar itself uses productively.

Two-feature operators appear inside grammars too, and so do the famous "corridors" of Chapter 12:

- A verb paradigm. English *was* *were* is a single verb's own past tense: `s` `r`. Historically this is Verner's Law followed by rhotacism (`s` → `z` → `r`); as an *operator corridor* it is `{voi}` composed with a sonorant step — the very `s` → `z` → `r` path the method surfaces from coderivatives in Chapter 12. A learner meets the corridor twice: across the family, and inside one irregular verb.
- Latin declension. *flōs* 'flower' → genitive *flōr-is*: `s` `r`, the same rhotacism corridor, now a case ending.
- Vowel ablaut in verbs. English *sing / sang / sung*, or Sanskrit *vid- / véd-a*: the vowel operators (`{hi}`, `{back}`, `{lo}`) that Chapter 13 treats as *graded* dimensions are exactly what conjugation exploits to mark tense.

The teaching point. An operator is a unit of *phonological difference*. The method reads it off coderivative pairs because that isolates it cleanly; but the same units are the working parts of every plural, tense, and derivation. When Part V speaks of adding a *directed, context-conditioned* layer (who changes into what, and when), morphology is exactly where those directed, conditioned rules become visible inside one grammar. *(None of the numbers in this book use these examples; they are the intuition, not the measurement.)*

PART III — The repertoire and its geometry

This is the heart of the book. Every number below was computed under the standing rule (no proto-forms, no laws, no OAS) and is reproducible with the named `make` target. We report Indo-European (IE) and Austronesian (AN) side by side.

7. Repertoire O vs generated subspace $\langle O \rangle$

The question. We have a family's set of operators, O. What is the *space* it lives in, and how much of that space does it actually use?

Two objects, kept separate. The repertoire O is the set of observed operators. Because operators are vectors over $\mathbb{F}_2$ (features under XOR), O *generates* a linear subspace

$$\langle O \rangle = \text{span}_{\mathbb{F}_2}(O),$$

of dimension $r = \dim\langle O \rangle$ (the rank, Chapter 8). Crucially, O is not a subspace: it is not closed under XOR and does not contain the zero vector. O is a *region*; $\langle O \rangle$ is the *capacity* it implies. The gap between them,

$$\langle O \rangle \setminus O,$$

is the set of operators the repertoire *could* generate but does not attest — the holes (Chapter 12).

The two families.

	$\lvert O \rvert$ (types)	rank r	$\lvert\langle O \rangle\rvert - 1 = 2^r - 1$
Indo-European	67	12	4095
Austronesian	22	10	1023

Why this separation matters. Almost every interesting question is really about the *relationship* between O and $\langle O \rangle$: how much of the capacity is used (Chapter 9), whether combining elements of O stays in O (Chapter 10), and which of the $\sim$4000 *span elements* $\langle O \rangle$ contains are the ones actually avoided — Chapter 12 makes this precise on the operator-shaped (weight-≤ 3) slice. Confusing the two — treating the repertoire as if it were the subspace — was the central error the earliest drafts made, and the reason the "do they form a group?" question was the wrong one to ask.

Limit. $\langle O \rangle$ is an *abstraction*: many of its $2^r - 1$ elements are not differences that could occur between two real segments. Chapter 9 replaces this too-large denominator with realizable universes.

Reproduce. `make universes FAMILY="X"`.

8. How many independent "knobs" does a family really use?

(A note on the two families used throughout. Indo-European, the familiar family that includes English, Spanish, Russian, Greek, Armenian and Hindi; and Austronesian, a very large family of roughly a thousand languages spread across the Pacific and island Southeast Asia — Malagasy, Malay, Tagalog, Fijian, Māori, Hawaiian. We compare the two throughout.)

The puzzle this chapter answers

We will meet a statement that sounds contradictory: *"Indo-European changes along 8 atoms, yet its space has rank 12."* How can there be more independent directions (12) than single-feature changes (8)? Clearing that up teaches the two ideas we need — rank and corank — from scratch.

Atoms and molecules

Recall an operator is *the set of features that differ* between two sounds. An atom changes exactly one feature; a molecule changes several at once.

- Atom: `{voi}` (voicing alone) — seen in `d t`, `g k`, `b p`, `s z`.
- Molecule: `{cont+strid}` (two features) — seen in `d z`, `t s`.

Indo-European uses 8 kinds of atom; Austronesian only 3. So far, so simple.

Rank: how many *independent directions* of change

Think of each phonological feature as a knob you can flip. A family's operators are the *combinations of knob-flips* it actually performs. The rank asks: how many knobs would you minimally need — as independent controls — to reproduce every combination the family uses?

Here is the subtlety, and the resolution of the puzzle. A feature can be a genuine independent direction *even if the family never flips it on its own*. Nasality (`nas`), labiality (`lab`) and rounding (`round`) never appear as Indo-European atoms — they only ever show up *inside* molecules like `{nas+son+voi}`. But they are still independent knobs: no combination of the other features reproduces them. So they count toward the rank even though they are not atoms.

That is the whole answer: atoms are the changes the family makes with a single knob; the rank counts every independent knob that moves at all, alone or in company. Indo-European makes single-knob changes on 8 features but *moves* 12 independent knobs in total — hence 8 atoms, rank 12.

A tiny worked instance. Suppose a family used just three operators: `{voi}`, `{cont}`, and `{voi+cont}`. It has 2 atoms (`{voi}`, `{cont}`) and rank 2 — because the third operator is just the first two flipped together (`{voi}` $\oplus$ `{cont} = {voi+cont}`), no new direction. Now add `{nas+voi}`: still only `{voi}` and `{cont}` as atoms (2 atoms), but rank rises to 3, because `nas` is a new independent knob even though it never appeared alone. That is exactly how IE reaches rank 12 with 8 atoms.

Corank: how many knobs turned out to be *redundant*

Count the features that move at all; call it n. Count the independent directions; that is the rank r. The difference $n - r$ is the corank — the number of knobs that are not independent, i.e. that always move as a fixed combination of others.

family	features that ever change (n)	rank r	corank $n - r$
Indo-European	12	12	0 — every moving knob is independent
Austronesian	11	10	1 — one knob is redundant

Indo-European has corank 0: all 12 knobs it moves are genuinely independent. Austronesian has corank 1: of the 11 knobs it moves, only 10 are independent; one is pinned to the others.

Which knob is redundant in Austronesian — and does it mean anything?

We can name the redundancy exactly. Across Austronesian's operators, whenever you know how two knobs — `cor` (coronal, made with the tongue tip/blade) and `lab` (labial, made with the lips) — are

set, the third knob `ant` (anterior, front-of-mouth) is already determined: in this sample, `ant = cor` $\oplus$ `lab`. It never varies independently of those two.

Be careful what this does and does not say. It does not say anteriority "is really" coronal-plus-labial in some deep sense. It says only that, *in the specific set of correspondences Austronesian happens to use*, the anterior knob never moves on its own — its pattern coincides with a combination of the other two. It is a fact about this family's *choices*, phrased in the language of linear algebra.

Is the redundancy real, or an accident of our feature chart?

This is the crucial control, and it is easy to state. Maybe `ant = cor` $\oplus$ `lab` is baked into the panphon feature chart itself — true of *any* set of sounds, not special to Austronesian. To check, we compute the rank not on the operators Austronesian *uses*, but on every difference that its sound inventory could possibly produce — all pairwise differences among the 88 consonants Austronesian actually has. Call that the *realizable universe* U_S. Result: U_S has rank 12, corank 0 — the feature chart *would allow* all 12 knobs to be independent. The redundancy is therefore not in the chart; it appears only in the operators Austronesian selects. So `ant = cor` $\oplus$ `lab` is a genuine property of the *system's choices*, not an artifact of how we wrote the features. (Run it: `make repr-control FAMILY="Austronesian"`.)

Which of these numbers are "real," and which depend on our bookkeeping

A last honest point. Some numbers change if we rename or recombine the features; others do not.

- Depends on the feature chart (bookkeeping): the *atom count* (8 vs 3), which features happen to appear alone, the size of each operator. An "atom" is only atomic *relative to the feature chart we chose*.
- Does not depend on it (intrinsic): the rank, the corank, the redundancy relations, and the circuit structure of Chapter 11. Rewrite the features in any invertible way and these stay the same.

So when we compare families, we lean on the *invariant* quantities (rank, corank, circuits) and treat the atom count as a readable-but-chart-dependent summary — never as an intrinsic dimension. It is meaningful only because our features carry real articulatory content, not arbitrary labels.

Reproduce. `make algebra FAMILY="X"` (atoms, rank, corank); `make repr-control FAMILY="X"` (the control).

8.1 The correspondence graph — the repertoire, drawn

The most direct picture of a repertoire is a graph whose nodes are sounds and whose edges are correspondences, one edge for each recurrent `a` `b`, coloured by the *kind* of change (voicing, fricativization, palatalization, place, prenasalization) and thickened by frequency. This is, strictly, a *segmental graph of correspondences labelled by operator* (Chapter 21 distinguishes it from the operator graph G_O and the language graph G_L). Read it as the "state space" of consonant correspondence for each family.

The Indo-European graph is denser and reaches more places of articulation; the Austronesian one is sparser and organized around voicing and its prenasalized bundle `t` ^{n}`d`, `k` ^{n}`g`, `p` ^{n}`b`. Neither picture required a single reconstruction.

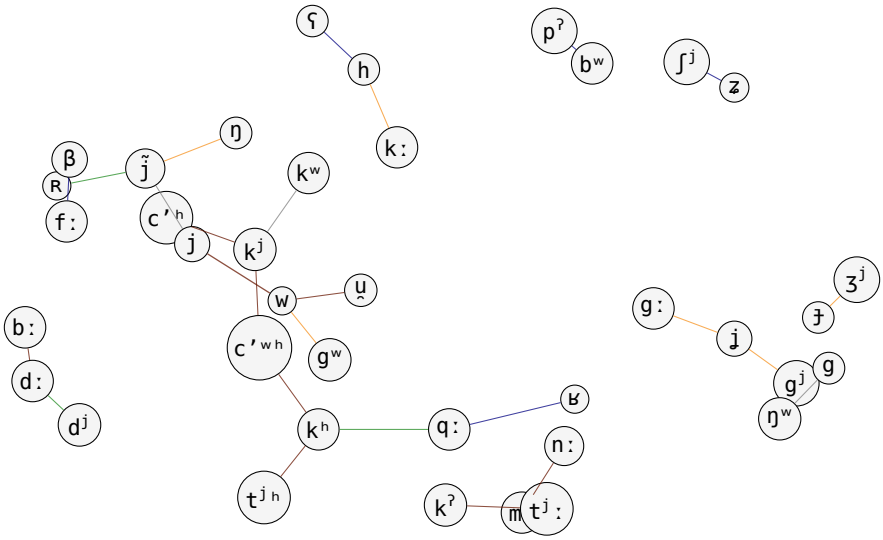

Figure 1: Segmental correspondence graph G_S of Indo-European. Nodes are segments; edges are recurrent correspondences, with thickness $\propto$ frequency and colour by change type: voicing, fricativization/lenition, palatalization/height, place, prenasalization.

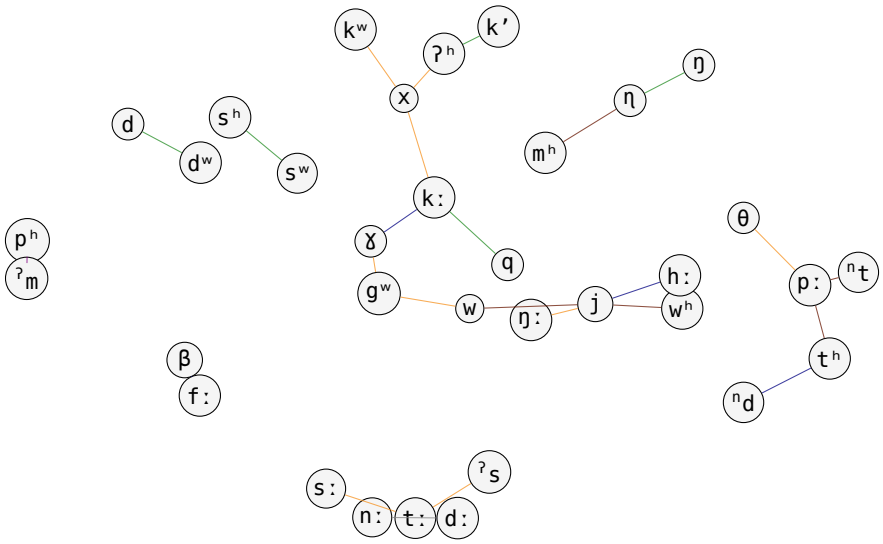

Figure 2: Segmental correspondence graph G_S of Austronesian. Nodes are segments; edges are recurrent correspondences, with thickness $\propto$ frequency and colour by change type: voicing, fricativization/lenition, palatalization/height, place, prenasalization.

9. Occupancy in three universes

The question. IE uses 67 of ∼4095 generable operators — 1.6%. Are phonological systems really that restrictive?

Worked example — 'three'. Take the forms the method groups for 'three' by recurrent correspondence (§5.1) — Bulgarian `tri`, Romanian `trei`, Ancient Greek `treis`, Breton `tri`, Yiddish `draj`. They align cleanly, and the clearest consonant correspondence is `t d` between `tri` and `draj` — our atom `{voi}`. Each meaning contributes a few such operators; pooled over the whole vocabulary they converge on just 67 distinct types. Sixty-seven out of an algebraic 4095 sounds like almost nothing — until you ask the right denominator (below).

The intuition. No — and this is one of the most important corrections in the book. The 1.6% compares the repertoire to the full *algebraic* span, most of which consists of feature-bundles that could never be a difference between two real segments, or that never had the chance to occur in the corpus. The honest question is "of what was actually available, how much is used?" — and that needs a smaller, realistic denominator.

Three nested universes.

$$O \subseteq \Omega_D \subseteq U_S \subseteq \langle O \rangle,$$

where U_S is realizable (differences between attested segments), and Ω_D is the opportunity universe (differences that actually occur in aligned positions of the corpus, at any support). Three occupancies follow:

$$\rho_{\text{alg}} = \frac{|O|}{2^r - 1}, \quad \rho_{\text{seg}} = \frac{|O|}{|U_S \cap \langle O \rangle|}, \quad \rho_{\text{opp}} = \frac{|O|}{|\Omega_D \cap \langle O \rangle|}.$$

The result.

	ρ_{alg} (span)	ρ_{seg} (inventory)	ρ_{opp} (aligner)	ρ_{ind} (alignment-free)
Indo-European	0.016	0.48	0.63	0.29
Austronesian	0.022	0.43	0.48	0.27

The "98% unused" was almost entirely the *algebraic* abstraction being enormous — not languages being restrictive. One honest wrinkle: ρ_{opp} uses Ω_D, which we built from our *own* alignments, so it could flatter itself. We therefore add ρ_{ind}, measured against an alignment-free opportunity Ω_{bag} (differences between consonants merely *co-present* in same-concept words). The aligner did inflate the number — ρ_{ind} is about half of ρ_{opp} — but the conclusion survives a denominator that never uses our aligner: at 0.27–0.29 (and $\rho_{\text{seg}} \approx 0.43$–$0.48$, also alignment-free) the repertoire is an order of magnitude denser than the span. Dense, not sparse. (`make independent-omega FAMILY="X"`.)

Limit. Ω_D depends on the corpus and the aligner; it is an *opportunity* estimate, not a phonological law. But it is far closer to the phenomenal reality than $2^r - 1$, and the three numbers *together* separate three kinds of constraint: algebraic, representational, and language-specific.

Reproduce. `make universes FAMILY="X"`.

10. Additive structure — the mathematical heart

The question. When we compose two operators — XOR their feature-differences — does the result tend to be another *observed* operator? Is the repertoire *additively organized*, or just an arbitrary set?

The measure and its conditioning. The composition-realization index

$$C(O) = \frac{|\{\{u, v\} \subseteq O : u \oplus v \in O\}|}{\binom{|O|}{2}}$$

(for $u \neq v$ in $\mathbb{F}_2$, $u \oplus v \neq \varnothing$ automatically, so the denominator is just $\binom{|O|}{2}$). Observed: IE 0.220, AN 0.234. Conditioned on realizability, $C_\Omega(O)$ asks: *of the compositions that land in the opportunity universe, how many are observed?* — IE 0.81, AN 0.66 (aligner Ω_D), and IE 0.57, AN 0.48 against the alignment-free Ω_{bag} (Chapter 9) — so the composition-closure survives a denominator that does not use our aligner.

A worked composition. `b p` $= \{voi\}$ and `p f` $= \{cont + strid\}$ compose to $\{voi + cont + strid\}$; the pair `b f`, when it occurs, necessarily carries that signature — the algebra *predicts* which operator it would be, and we can check whether the system realizes it.

Does it beat chance? A hierarchy of nulls. A single number is descriptive until compared to a null. We built six *increasingly constrained* nulls (not a strict nesting — the span null samples a different ambient set, so the Z column below is not monotone), and asked whether observed $C(O)$ exceeds them (Z-scores; empirical p_{MC} from 500 simulations — the resolution floor; a larger run and multiple-comparison control are future work):

null (preserves…)	IE Z	AN Z
0 size + weight ≤ 3	+15.7	+12.0
1 Hamming weights	+14.8	+9.0
2 exact margins (swap-MCMC)	+7.5	+3.6
3 rank / span $\langle O \rangle$	+45.6	+13.7
4 realizable U_S	+9.2	+7.3
5 opportunity Ω_D	+6.3	+5.9

All six survive at $p_{MC} \leq 0.004$. Even when we draw $|O|$ operators from *exactly the differences the corpus made available*, the observed repertoire is significantly more composition-closed. Additive structure is a robust, size-controlled property — but *how much* it is robust differs sharply between the two families, and the right test is a language-level bootstrap (resample the valid unit — the language — with replacement, rerun the whole pipeline, $B = 40$; $C(O)$ is a U-statistic over non-independent pairs, so the language, not the pair, is the unit). It gives $C(O)$ $\text{CI}_{95} = [0.18, 0.22]$ for IE — tight, comfortably above the null floor — but $[0.15, 0.30]$ for AN, wide and with its lower end near the null. So IE's additive structure is robust to which languages are sampled; Austronesian's is fragile — a caution the size-matched nulls alone did not surface, and a direct consequence of AN's smaller, more variable repertoire. (An earlier parametric support-bootstrap gave narrower intervals, $\approx [0.21, 0.23]/[0.20, 0.24]$; the language-level intervals supersede it as the statistically valid unit.)

Three converging signatures (each on unordered tuples, to match $C(O)$). Against the Ω_D null: triple density $\tau(O) = |\{\{u, v, w\} \subseteq O : u \oplus v = w\}|/\binom{|O|}{3}$ is high (IE $Z = +6.6$, AN $+6.5$); the doubling constant $\kappa = |O \oplus O|/|O|$ (how much the sumset $O \oplus O = \{u \oplus v\}$ blows up) is *low* (IE $Z = -5.1$, AN -6.0; small doubling = strong additive structure); and the additive energy $E(O) = |\{(a, b, c, d) \in O^4 : a \oplus b = c \oplus d\}|/|O|^3$ is high (IE $Z = +7.4$, AN $+7.3$). The repertoire behaves like an approximate subspace, possibly an affine one $v + W$ (a family of operators sharing an obligatory feature — a direct affineness test is future work).

The honest caveat — this is not (yet) a genealogical signal. At the level of operator *types*, this additive structure is largely a property of the phonological inventory and feature representation, not of history. The evidence is decisive: when we permute concept labels to destroy all real cognacy and semantics (regime D_R, Chapter 17), $C(O)$ and τ stay *as high or higher* than under expert cognacy D_G. So $C(O)$ exceeds *size-matched random* (the nulls above) — that is real — but it does not distinguish genealogy from inventory. Where genealogical signal lives is the subject of Chapter 14 and Part IV: in the *distributions*, not the type sets.

Reproduce. `make nulls FAMILY="X"` (the hierarchy); `make additive FAMILY="X"` (τ, κ, E).

11. Circuits and binary matroids

The question. Can we describe the repertoire's dependency structure in a way that does *not* depend on the arbitrary choice of feature basis?

The matroid. The operator×feature incidence matrix defines a binary matroid. *(In plain terms, a matroid is just the bookkeeping of what is independent versus redundant — the same idea as rank in Chapter 8, lifted from single features to whole operators. You do not need the theory; you need one object it hands us.)* Its most interpretable invariants are the circuits: minimal sets of operators whose XOR is zero. A size-3 circuit $u \oplus v \oplus w = 0$ means $w = u \oplus v$ — three operators locked in an additive triangle.

Worked example — a real circuit. Three Indo-European operators are locked in an additive triangle: `{ant}` (the sibilant shift `s ʃ`), `{strid}` (`s θ`), and `{ant+strid}` (`ʃ θ`). Check it: `{ant}` ⊕ `{strid}` = `{ant+strid}`, i.e. `s ʃ` composed with `s θ` gives `ʃ θ` — and all three are operators the family actually uses. That closed triangle is a size-3 circuit. It does not depend on how we named the features: rename or recombine them and the triangle is still there. Another, involving the hub: `{voi}` (`d t`) ⊕ `{ant}` (`s ʃ`) = `{ant+voi}` (`z ʃ`).

The result. IE has 162 size-3 circuits, AN 18. In *both* families the operator participating in the most circuits is **`{voi}`** — voicing is the additive hub of the repertoire. Because circuits are a matroid invariant, this is a basis-independent fact: it does not depend on how we named or coordinatized the features.

Why it matters. Two systems could use different feature labels yet share the same matroid — the same pattern of dependencies. The matroid is therefore the right object for asking *"do two systems have the same dependency structure?"*, a question much closer to the goal of finding patterns more abstract than any particular notation. (Part V returns to this comparison.)

Reproduce. `make additive FAMILY="X"`.

12. The span as a linear code

The question. We said the repertoire uses little of its span. *Which* generable operators are missing, and how far are they from what is present?

Worked example — 'mother'. Take the forms the method groups for 'mother' — Serbo-Croatian `mati`, Spanish `madre`, Armenian `majr`, Lower Sorbian `muterka`. Their medial consonant lines up recurrently: `mati`'s `t` against `madre`'s `ð` gives `t ð` = `{cont+voi}`, an *observed* operator (voicing alone, `t d` = `{voi}`, and continuancy alone, `{cont}`, are observed elsewhere in the family). Now ask about a nearby operator the family *could* generate but does not use here — say `{cont+voi+strid}` (a `t z`, adding stridency to the `t ð` bundle). If `{cont+voi}` is present and `{strid}` is present but `{cont+voi+strid}` is not, that missing combination is a hole at distance 1: available in the corpus's opportunities, one feature away from what is attested, yet unchosen. Cataloguing exactly these near-misses is the geometry below.

The code and the holes. Treat $\langle O \rangle$ as a binary linear code. *(A linear code is simply a set closed under XOR — here, every operator the observed ones can generate. The point of borrowing the term is that coding theory already has tools for exactly our question: which points are missing, and how far.)* Among the operator-shaped vectors (weight ≤ 3) that the code generates, some are observed (O) and the rest are holes. For each hole x we measure its depth $d(x, O) = \min_{o \in O} d_H(x, o)$ (Hamming distance to the nearest observed operator).

The geometry of holes.

	weight-≤ 3 generable	observed	holes	at $d = 1$	holes in Ω_D
Indo-European	298	67	231	160	40
Austronesian	119	22	97	53	24

Most holes sit at $d = 1$ — one feature away from completing a pattern. The linguistically loaded ones are the holes that lie in Ω_D (40 in IE, 24 in AN): operators that were *available* in the corpus yet went *unchosen*. These are the concrete realization of $\langle O \rangle \setminus O$ — the "generable-but-avoided" object — and the natural candidates for interpretation: absences that are choices, not impossibilities.

Limit. For IE the code is the whole space (rank 12 = 12 features), so "the code" is only a proper, informative object where the corank is positive (AN); the *hole geometry*, however, is informative for both, because it is measured on the operator-shaped (weight ≤ 3) slice, not the whole span.

Reproduce. `make additive FAMILY="X"`.

13. Non-binary dimensions (the "XOR-break")

The question. The composition law $\Delta(a, c) = \Delta(a, b) \oplus \Delta(b, c)$ is an identity for binary features. Yet on real data it holds "only" 98.1% of the time in IE. If it is an identity, how can it ever fail?

The diagnosis. It cannot fail for genuinely binary features. It fails precisely where a feature is not binary: panphon features are *ternary* $(+1, 0, -1)$. All five IE failures are triples where one feature takes three distinct values across a, b, c — every one of them in height (`hi`): `l` `lʲ` `r`, `l` `r` `ʎ`, ... So the "XOR-break" is not a violated law; it is a diagnostic that flags which dimensions of the system are *multivalued* and cannot be modeled with a single bit.

What it does and does not show. It localizes a *multivalued dimension*; it does not by itself prove a totally-ordered phonetic scale. `l, lʲ, r, ʎ` vary at once in laterality, roticity, and place; whether the right geometry is a total order, a partial order, or a state graph is a separate question. Austronesian shows no XOR-break in its 47 relevant triples — which means *none was detected in this sample* (with $n = 47$ a small true rate could be missed), not that AN "has no graded dimensions."

The fix, when wanted *(for the mathematically inclined; skippable on a first read)*. The takeaway in one line: *stop forcing every feature to be a yes/no switch — give a three-valued feature three states instead of one bit.* Concretely, for a ternary feature one can use a one-hot encoding $-1 \mapsto e_1, 0 \mapsto e_2, +1 \mapsto e_3$, restoring exact composition for *categorical* states; or, for *ordered* states, a path metric with $d(-1, +1) = 2$. The general system is then a product of typed spaces

$$\mathcal{F} = \prod_{j \in B} \mathbb{F}_2 \times \prod_{j \in C} \Delta_{m_j} \times \prod_{j \in O} \mathbb{Z},$$

binary $\times$ categorical $\times$ ordered — or, most generally, a box product of per-feature state graphs. The mathematics adapts to the feature instead of forcing every feature to be a bit.

Reproduce. `make algebra FAMILY="X"` (reports the break and the graded features).

14. Distributions, not only types

The question. The repertoire records *existence*. But how much of the structure is in the *frequencies* and in *what each operator co-occurs with*?

Worked example — 'nose' vs the voicing atom. Austronesian 'nose' is `ihu` in Hawaiian, Maori, Tongan, Tuamotuan, Rapa Nui, Marquesan — *identical everywhere.* Its correspondence set is empty: zero

operators, a word so conserved it contributes nothing to the repertoire. At the other extreme, {voi} (d t, b p, s z, ...) recurs tens of thousands of times across the vocabulary. The repertoire is not a flat list: a few operators carry almost all the mass, many carry almost none, and some words (like *ihu*) carry none at all. That imbalance is exactly what entropy measures.

Distribution and its concentration. With per-instance counts we form $P_L(o)$, its Shannon entropy H, and the effective number of operators $N_{\text{eff}} = 2^H$. *(Entropy just measures how* concentrated *a distribution is — low when one operator dominates, high when many share the load; N_{eff} turns it back into a count: "as if this many operators were used equally.")* IE: $N_{\text{eff}} = 24$ effective operators of 119 distinct types; AN: 18 of 63. *(These 119 / 63 are the full instance-level inventory —* every *operator type that occurs at any support; the smaller repertoire O of 67 / 22 counts only those above the support threshold. Three legitimately different counts, one per question: instance types (119), thresholded repertoire O (67), and branch-union O_F (97, Chapter 15).)* A handful of operators dominate each family.

Three geometries. *Unweighted* (what exists), *frequency-weighted* (how often), *confidence-weighted* (how securely inferred, using alignment cost as a proxy). The top-8 operators coincide 7/8 between frequency and confidence in both families — the repertoire's ranking is robust to confidence weighting. (Support, distribution, and confidence are nonetheless different robustnesses, and should be reported separately.)

Mutual information — where dependence lives. *(Mutual information asks a simple question: once I tell you the branch, concept, or position, how much less uncertain are you about which operator it is? Zero means the operator does not depend on that variable at all; higher means it does.)* $I(O;\cdot)$ (normalized) between the operator and four variables:

	branch R	concept C	position	context Γ
Indo-European	0.14	0.16	0.06	0.05
Austronesian	0.22	0.41	0.06	0.09

Dependence is strongest on concept (partly a confound: the concept fixes which segments are in play), then on branch — and here is the key: branch information is *higher in Austronesian* (0.22) than in Indo-European (0.14), consistent with the branch analysis of Part IV. Dependence on position and coarse context is *low*.

The synthesis of Part III. The additive/geometric structure of the *type set* is largely a property of the phonological representation, shared across the two families and even reproduced by concept-permuted artifacts. The genealogical signal — the thing historical linguistics cares about — is *second-order*: it lives in the distributions (this chapter's branch information), not in the type sets or their additive closure. That is why the fine, context-conditioned analysis (Part V's tensor) is the right next instrument; coarse context shows little here.

Reproduce. `make distributions FAMILY="X"`.

PART IV — Systems and comparison

Comparative linguistics compares *forms* and reconstructs *trees*. This part asks a different comparative question: do the repertoires of related systems overlap more than chance, and where does any genealogical signal actually reside? The answers are more sobering — and more interesting — than the enthusiasm of a first look.

15. The family as a union of branches

The question. Is the "algebra of Indo-European" a single thing, or the superposition of several branch algebras dominated by whichever branches are best sampled?

Worked example — 'two', within a branch vs across. Take the forms for 'two'. Inside the *Armenic* branch, Eastern `jerkʔu` and Western `jergu` differ by a single clean operator, `kʔ g = {voi}` — a tidy, intra-branch correspondence. Across *different* branches the forms still line up by a recurrent initial correspondence — Romanian `doi`, Serbo-Croatian `dva`, Ancient Greek `duo` all share a `d`-initial — so the intuition "related branches share structure" seems obvious. The surprise (below) is that, *measured against a proper null*, that overlap is no greater than among random language groupings in Indo-European — whereas in Austronesian the branches are genuinely more differentiated than chance.

The clean decomposition. Partition each family into its genealogical branches (Glottolog, cut where clades reach a manageable size), run the whole pipeline *within* each branch, and take the family repertoire as the union $O_F = \bigcup_r O_r$. For each operator define its persistence $p(o) = \frac{1}{R}\#\{r : o \in O_r\}$ and the histogram $H_k = \#\{o : o \text{ in exactly } k \text{ branches}\}$, which by construction sums to $|O_F|$. Three nuclei, named unambiguously: $K_\cap$ (in *all* branches), $K_{1/2}$ (in $\geq \lceil R/2 \rceil$), K_α (general threshold).

The decomposition. IE (7 branches): $|O_F| = 97$, $\Sigma H_k = 97$ ✓, $K_\cap = 12$ (twelve operators present in *every* branch — among them the atoms `{ant}` `{cont}` `{voi}`, the rest small molecules), $K_{1/2} = 38$, exclusives $= 26$. AN (11 branches): $|O_F| = 48$, $K_\cap = 4$, $K_{1/2} = 18$, exclusives $= 15$. *(This corrects a v3 table that mixed two universes and appeared to double-count; the fix was to use $O_F = \bigcup O_r$ consistently and let $\Sigma H_k = |O_F|$.)*

The decisive control. A raw Jaccard between branch repertoires (IE 0.42, AN 0.51) means *nothing* by itself. We compare it to a grouping null: partition the same languages into arbitrary groups of the same sizes and recompute. The result overturns the earlier "superposition" story:

	mean Jaccard (real)	arbitrary groupings	Z
Indo-European	0.42	0.45 ± 0.05	-0.7 (indistinguishable)
Austronesian	0.51	0.57 ± 0.01	-5.4 (real branches more differentiated)

So at the level of operator types, Indo-European branches overlap *no more than random language groupings* — there is no type-level genealogical signal in IE; the shared core is shared *inventory*. Austronesian branches, by contrast, are significantly more differentiated than chance (Formosan vs Oceanic have very different inventories) — a genuine signal.

Limits. Under rarefaction to equal branch sizes, the nucleus size is unstable (IE $K_{1/2}$ falls from 38 to $\sim$15), so branch *richness* counts are sampling artifacts and must not be interpreted. The robust facts are: (i) the union/histogram decomposition; (ii) the grouping-null result (no IE type-signal; AN signal). The genealogical signal, where it exists, is distributional (Chapter 14), not in the type sets.

Reproduce. `make superposition FAMILY="X"`; add `TF_RAREFY=3 TF_NULLBRANCH=3` for the controls.

16. Negative controls

The question. How do we know a ”system” pattern is not just general phonetic geometry?

The controls, and their status.

- Concept permutation (D_R, Chapter 17): destroys form↔concept links. *Run* — and it matches the real additive structure, showing much of it is inventory-driven.
- Branch-label permutation (Chapter 15): arbitrary groupings of the same sizes. *Run* — the grouping null above.
- Pseudo-families / inventory controls: random language groups matched on size, segment inventory, concept coverage, alignment count. *Partially run* via the grouping null; a fuller inventory-matched control is scheduled.
- Geographic control: compare genealogical groups with areal groups of the same size. *Planned.*
- Temporal control: the corpus mixes ancient and modern states; the absence of protoforms does *not* force us to drop time — we can use documented dates. Build a temporal network $G(t)$ or attach a date interval to each form and study repertoire change without reconstructing unattested states. *Planned.*

The lesson. A number is a finding only relative to the strictest control it survives. The additive structure survives size-matched random draws (Chapter 10) but not the concept-permutation artifact at the type level (Chapter 17) — both statements are true and must travel together.

Sensitivity — do the headlines depend on our knobs? No. Sweeping the support threshold (15–60), the operator weight cap ($\leq 2, 3, 4$), and the feature subset (the full 12, minus `lab`, minus `{round,lo}`, an 8-feature manner+voice+place set), the repertoire size $|O|$ and the rank vary *smoothly*, but $C(O)$ and the occupancy ρ_{opp} stay clearly above the null floor throughout — $C(O) \in 0.21$–0.37 (IE) and 0.14–0.28 (AN), against a null of ≈ 0.12–0.15. The one honest soft spot is Austronesian at weight cap ≤ 2 ($C(O) = 0.14$, only 12 operators), too small to resolve. So the headline signs are properties of the data, not of where we set a dial. *(The cognacy-threshold and aligner sweeps, which need the corpora, are scheduled.)*

Reproduce. `make sensitivity FAMILY="X"`; `make regimes` (concept permutation); `make superposition … TF_NULLBRANCH=3` (grouping).

17. Four corpus regimes: D_G, D_L, D_C, D_R

The question. Rather than discard the concept-only corpus as ”dangerous,” can we use it — and its degradations — as a control that separates *sources* of structure?

Four regimes on IE-CoR. D_G expert cognacy · D_L statistical (LexStat) cognacy · D_C same-concept, no cognacy · D_R concepts permuted (the artifact floor).

The fingerprints.

regime	$\lvert O\rvert$	$C(O)$	τ	κ
D_G expert cognacy	65	0.169	0.166	14.4
D_L LexStat	44	0.162	0.158	12.3
D_C concept-only	108	0.204	0.202	14.6
D_R concept-permuted	104	0.215	0.212	14.6

The honest reading. The artifact D_R has additive structure *as high as or higher than* expert cognacy D_G. Therefore the additive organization of the operator type set is largely a property of the segment inventory and its frequencies — not of genealogy, and not even of conceptual equivalence ($D_C \approx D_R$). This is the type-level counterpart to Chapter 15's grouping-null result. It also tells us where the LexStat validation matters: at the type level LexStat's precision against D_G is ≈ 1.0 (it recovers real operators cleanly), even though at the pair level its precision is only 0.73 — the type inventory absorbs individual cognacy errors.

Loans as system-influence, not annotation (idea 2). The right way to ask "cognate vs loan, mathematically" is not to define it but to *observe* whether loans move the repertoire: compare O with and without loan-flagged forms — do loans add operators, or fall inside the existing O? The machinery is implemented; the current data cannot answer it (IE-CoR marks borrowings only in free-text comments; Lexibank's `Loan` column is empty for IE and AN). This needs a loan-annotated corpus such as WOLD, and is a clean next experiment.

Reproduce. `make regimes`.

18. Family-blind: when yes, when no

The question. Both "systems" here were defined by genealogical labels. Could we compare systems *without* presupposing families?

The intuition, and the caution. Yes — assign each language (or language pair) an operator distribution $P_{ij}(o)$, define a distance between distributions, cluster, and see what systems *emerge*. But this points at the universal/typological questions we are deliberately not asking in this phase. Family-blind analysis belongs to the *historical-holdout* programme (Chapter 20, phase C), where it is a powerful discovery tool — not to the present intra-system study, where imposing it would blur exactly the boundaries we are studying.

The right distance. Not Jaccard (which treats two near-identical operators as wholly different) but a Wasserstein distance over operator distributions, with ground cost $d(o_1, o_2)$ from Hamming or articulatory distance — so two systems using *neighboring* operators are close even if their type sets do not coincide.

Why not now. Family-blind clustering answers "what large-scale systems exist?"; this book answers "what is the structure *within* a given system?" Mixing them would trade a sharp question for a diffuse one. We flag the method and its tools, and reserve it.

Reproduce. Not yet a target — specified here as a Part-V experiment.

PART V — Toward history and toward AI

Everything so far kept history out of the computation. This part is where the reserved apparatus is allowed back — as *contrast*, never as coordinate — and where the programme's larger ambitions are laid out. These chapters are a research programme, not finished results.

19. Letting historical laws emerge without inserting them

A sound law is rarely a single operator; it is a *bundle* of correspondences, tied to particular languages and often conditioned by context. So we ascend in orders. Order 0 is this book: the aggregate repertoire O_L. Order 1 builds the tensor X_{ijo} = normalized frequency of operator o between languages i, j; a non-negative factorization $X \approx \sum_q \lambda_q \, a_q \otimes a_q \otimes b_q$ yields components where a_q selects a set of languages and b_q a *bundle* of operators — a recurrent transformation cluster the algorithm was never told the name of. Order 2 adds context X_{ijoc} (preceding/following segment, position, stress, morphology) and seeks $P(o \mid c, i, j)$. Here Minimum Description Length gives an emergence criterion: a regularity earns the name *rule* when a small set of context-conditioned operators compresses a large slice of the corpus, i.e. minimizes $L(\mathcal{R}) + L(D \mid \mathcal{R})$. "Grimm's law" would be *discovered* as a compressive bundle over Germanic, not supplied in advance.

20. A historically-blind validation protocol (historical holdout)

Phase A — discovery: no protoforms, laws, change classes, direction; in some runs, not even family labels. Phase B — mathematical stability: bootstrap over languages and concepts; robustness to the feature matrix, the cognacy threshold, and the aligner; superiority over the null hierarchy. Phase C — blind system discovery: cluster operator profiles with no genealogical labels; ask whether families, areas, or continua emerge, and which languages land in surprising places. Phase D — external contrast: only now compare with reconstructions, laws, chronologies, genealogies, and documented contact. A non-match is itself a result — it localizes either a corpus limit, a representational gap, or a place where traditional terminology lumped mathematically heterogeneous phenomena. This design is what protects the project from the charge of merely re-encoding what historical linguistics already handed it.

21. A general architecture

Four layers plus an external one. Phenomenal: $D = \{(\ell, t, g, c, w, \pi(w))\}$ (language, date, place, concept, form, phonetics). Relational: weighted alignments between forms. Operatory: the signatures $\Delta(a, b)$ and the distributions $P_{ij}(o)$. Emergent-structural: $\mathcal{A}_L = (O_L, V_L, G_S, G_O, G_L, P_L)$ — repertoire, span, and three graphs. External-historical: reconstructions, laws, genealogies, contacts — used only in Compare$(\mathcal{A}_L, H)$. The three graphs must be kept distinct: G_S segmental (nodes = segments, edges = correspondences — what the current figures actually show), G_O of operators (nodes = operators, edges = composition $u \oplus v \in O$), and G_L of languages (nodes = languages, edge weight = distance between operator distributions). History does not disappear in this architecture; it changes epistemic position, from initial coordinate to final contrast.

The segmental graph G_S was shown in §8.1. Here are the other two. First the operator graph G_O, whose edges are additive closure ($u \oplus v \in O$) — the visual face of the composition index $C(O)$ of Chapter 10:

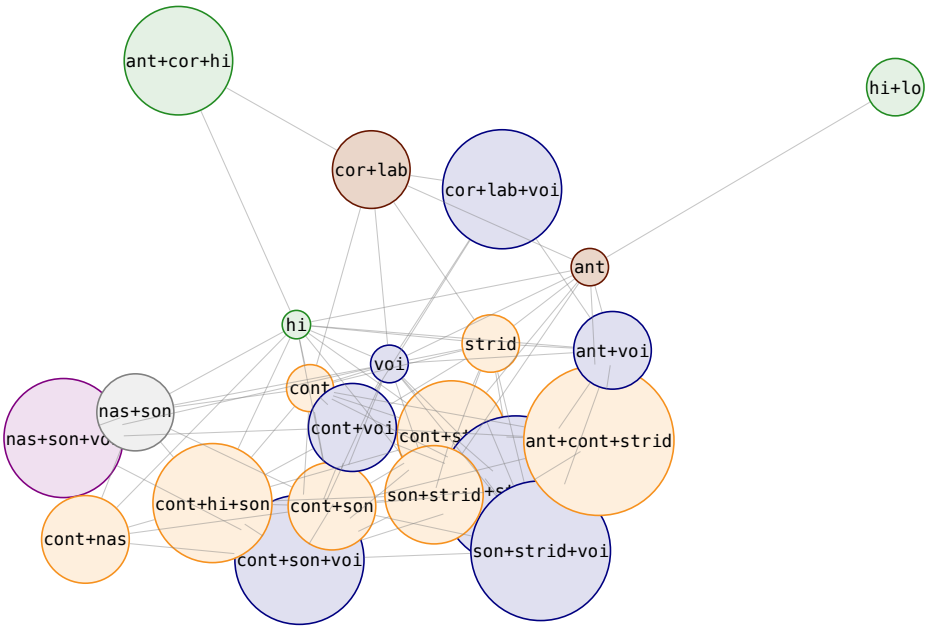

Figure 3: Operator graph G_O of Indo-European. Nodes are operators (the 22 with highest support), sized by frequency and coloured by change type (voicing, fricativization, palatalization/height, place, prenasalization). An edge joins u, v when their composition $u \oplus v$ is itself an operator of the repertoire — so this is the graph of *additive closure*. Dense connectivity is the visual face of the composition index $C(O)$.

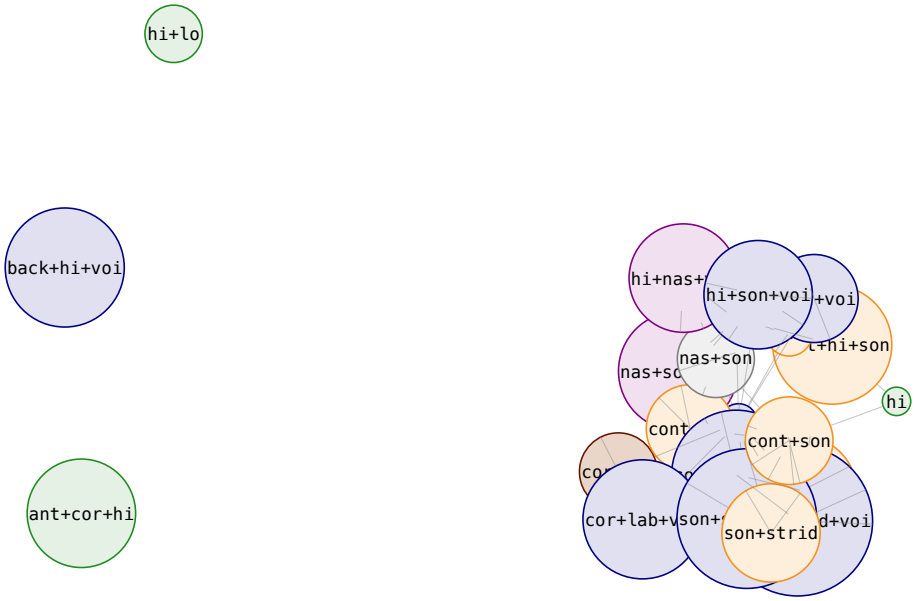

Figure 4: Operator graph G_O of Austronesian. Nodes are operators (the 22 with highest support), sized by frequency and coloured by change type (voicing, fricativization, palatalization/height, place, prenasalization). An edge joins u, v when their composition $u \oplus v$ is itself an operator of the repertoire — so this is the graph of *additive closure*. Dense connectivity is the visual face of the composition index $C(O)$.

Then the language graph G_L, built *only* from operator distributions (Jensen–Shannon distance), with nodes coloured — after the fact — by genealogical branch. Whether the colours cluster is a direct, visual version of the question in Chapters 14–15: how much genealogy do the distributions alone carry?

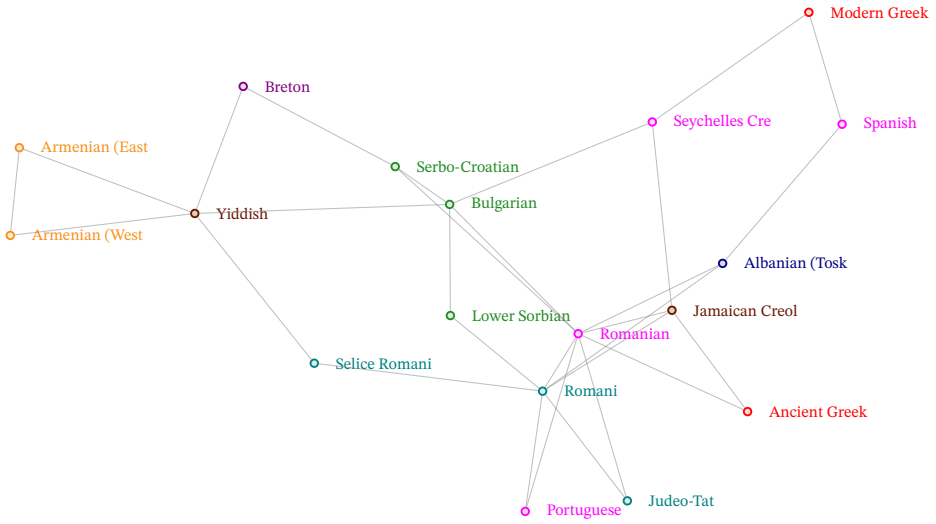

Figure 5: Language graph G_L of Indo-European. Each node is a language, coloured by its Glottolog branch; two languages are joined when one is among the other's two nearest neighbours in *operator-distribution* distance (Jensen–Shannon). Colour was *not* used to build the graph — only the operator frequencies were. Whether same-colour nodes cluster is a visual test of how much genealogy the distributions alone carry (cf. Chapters 14, 18). Branches: Albanian, Armenic, Balto-Slavic, Celtic, Germanic, Graeco-Phrygian, Indo-Iranian, Italic.

22. Implications for AI language models *(speculative)*

This chapter is openly speculative. Two threads. First, representation: if human phonological systems occupy sparse, additively-structured, few-effective-dimensional regions of a feature space, that is a statement about the *inductive biases* a good model of sound structure might want — not one-hot segment inventories but feature-additive codes with a strong low-rank/affine prior. Second, methodology: the project's core stance — *discover structure first, bring in the ground-truth labels only to test* — is exactly the discipline that guards machine-learning evaluation against leakage. A model that "recovers Grimm's law" is impressive only if the law was never in its inputs; the historical-holdout protocol (Chapter 20) is a template for such honest evaluation of language models on historical and typological tasks. We make no stronger claim than that these are worth trying.

23. A program, not a study

This is one study; the object it opens is large. Persistent topology (build a complex with a triangle whenever $u \oplus v = w$ and watch cycles and voids across a frequency threshold), spectral analysis of G_S/G_O, and multilayer graphs (a layer per branch, period, context) are all natural next instruments — none of which reduce the geometry to a single number. And two things stay reserved on principle, to be brought in only as external partitions once the structure is mapped: protoforms, and OAS. The question for OAS is precise and non-circular: *do the dimensions, communities, or latent factors discovered from IPA and documented correspondences spontaneously recover groupings compatible with OAS?*

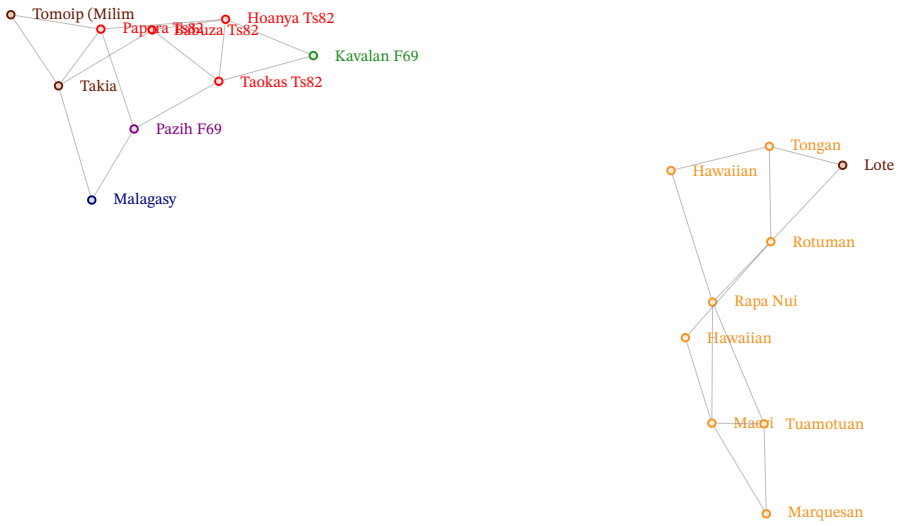

Figure 6: Language graph G_L of Austronesian. Each node is a language, coloured by its Glottolog branch; two languages are joined when one is among the other's two nearest neighbours in *operator-distribution* distance (Jensen–Shannon). Colour was *not* used to build the graph — only the operator frequencies were. Whether same-colour nodes cluster is a visual test of how much genealogy the distributions alone carry (cf. Chapters 14, 18). Branches: Basap-Greater Barito, Central Pacific linkage, East Formosan, Northwest Formosan, Western Oceanic linkage, Western Plains Austronesian.

If they do, that is independent evidence; if they do not, it lets the classes be revised without circularity. Either way, OAS earns its place by *surviving a blind test*, never by being assumed.

Appendices

Appendix A — Glossary of symbols

symbol	meaning
$\phi(a)$	panphon feature vector of segment a
coderivative	two forms with a shared, evidenced history (recurrent correspondence), *without* committing to a single reconstructed ancestor; a cognate is the special case of clean descent (§3.1)
$\Delta(a,b) = 1[\phi(a) \neq \phi(b)]$	operator = indicator of differing features (the signature); $= \phi(a) \oplus \phi(b)$ only on binary features
O (or O_L)	repertoire: the set of observed operator types
$\langle O \rangle$	subspace generated by O over $\mathbb{F}_2$; $r = \dim\langle O \rangle$ (rank)
$\langle O \rangle \setminus O$	the holes: generable but unattested operators
U_S	realizable universe: $\{\Delta(a,b) : a,b \in S\}$, S = segment inventory
Ω_D	opportunity universe: differences occurring in aligned corpus positions
$\rho_{\text{alg}}, \rho_{\text{seg}}, \rho_{\text{opp}}$	occupancy vs span / realizable / opportunity
$C(O), C_\Omega(O)$	composition-realization index (raw / conditioned on opportunity)
$\tau(O),\ \kappa(O),\ E(O)$	triple density, doubling constant, additive energy
$K_\cap, K_{1/2}, K_\alpha$	branch nuclei (all / majority / threshold); H_k, $p(o)$ persistence
$P_L(o), H, N_{\text{eff}}$	operator distribution, entropy, effective number of operators
$T = (a,b,c,p,\ell_1,\ell_2,w)$	the future *directed* transformation; $\sigma(T) = \Delta(a,b)$

Appendix B — Catalog of operations (**`make`** targets)

target	chapter	what it computes
`make family FAMILY="X"`	5	correspondences (LexStat + alignment) → repertoire
`make algebra FAMILY="X"`	8, 13	atoms, rank/corank, XOR-break, $C(O)$
`make repr-control FAMILY="X"`	8	representation control (O vs U_S rank/dependencies)
`make universes FAMILY="X"`	9	three universes, three occupancies, C_Ω
`make nulls FAMILY="X"`	10	six nested nulls for $C(O)$ with p_{MC} + bootstrap
`make additive FAMILY="X"`	10, 11, 12	τ, κ, E; matroid circuits; hole geometry
`make distributions FAMILY="X"`	14	$P_L(o)$, entropy, N_{eff}, mutual information
`make superposition FAMILY="X"`	15	branch decomposition, grouping null (+ `TF_RAREFY`, `TF_NULLBRANCH`)
`make cognate-eval`	5	LexStat vs expert cognacy (pair- and type-level)
`make regimes`	17	four corpus regimes $D_G/D_L/D_C/D_R$ + loans
`make sensitivity FAMILY="X"`	16	sweep support threshold, weight cap, feature subset (invariance)

target	chapter	what it computes
`make chains FAMILY="X"`	—	preferred change corridors (monotone dimensions)

Appendix C — Reproducibility

All results run from `src/` against `data/db/transf.db` (built by `make family`). Datasets: Lexibank (IE 304 / AN 978 languages available; per-family top languages by form count), IE-CoR (`iecor`, expert cognacy + segments), Glottolog classification (cached in `data/glottolog_classification.csv`, 26 748 paths). Feature matrix: panphon (ternary values); the 12-feature primary subset used is `cont, voi, nas, ant, cor, lab, back, round, strid, hi, lo, son` (a documented subselection of panphon's larger set). Cognacy: LexStat, threshold 0.55, scorer 100 runs. Alignment: Needleman–Wunsch with panphon feature-fraction cost. Thresholds: operator support $\geq$ 30 (family), $\geq$ 8 (branch); operators of 1–3 features. Exact per-family counts are printed by each script. Seeds are fixed in the null/bootstrap code for reproducibility.

Appendix D — Revision history

This manual descends from an internal report that went through three external review rounds. Round 1 corrected the conflation of representation with data (the "group under XOR" is a property of the encoding, not of languages). Round 2 added the nucleus/superposition analysis and the cognate-vs-conceptual validation. Round 3 demanded, and received, the realizable/opportunity universes, the nested nulls, the basis-dependence distinction, the two-level cognacy metrics, and the grouping null (which overturned the round-2 "IE more differentiated" reading). Retractions along the way — most notably "nasality is not an independent movement" (failed a null) and "branch richness differs" (a sampling artifact) — are kept here rather than in the body, so the argument can be read in its corrected form.

Appendix E — Data supplement

This appendix is a pointer, not a static supplement. The full operator tables, branch-by-branch repertoires, and per-chapter raw outputs are not reproduced here; they are regenerated on demand by the `make` targets of Appendix B against the bundled `data/db/transf.db`. Full-page and interactive versions of the graphs are planned as the manual matures.