Symmetry Hides History

A symmetric correspondence operator is blind to conditioned sound laws; direction, taken from an attested ancestor, restores them

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

Abstract

A companion study introduced an object: the repertoire O of *symmetric* feature-difference operators $\Delta(a, b)$ that a language family draws on, computed with no protoforms and no sound laws. It found that the additive structure of O behaves like a property of the phonological representation rather than of genealogy.

This paper isolates *why* the symmetric view cannot see conditioned sound laws, and shows what recovers them.

The reason is structural. The operator $\Delta(a, b) = \Delta(b, a)$ is symmetric by construction, so any signal that lives in the *orientation* of a change or in the *environment* that conditioned it is discarded before any number is computed. A conditioned sound law is a statement about orientation and environment. So it is invisible to Δ by construction — not because a language lacks it.

We restore what symmetry throws away, in the one setting where this needs no reconstruction: an *attested* ancestor. Across three documented ancestors — Latin, Vedic, and Old Church Slavonic — we test whether directing the comparison from ancestor to daughter, with a null that respects the clustering of the data, recovers conditioned laws the symmetric view misses.

It does, with a clean and principled boundary. Latin $/k/$ palatalises before front vowels ($p = 0.0002$, cluster-level permutation) — a law that post-dates Latin and is recovered on the Latin$\to$Romance branch. Old Church Slavonic$\to$Slavic recovers the pervasive softening of consonants before front vowels across ten consonants (all surviving Benjamini–Hochberg correction). Vedic$\to$Indo-Aryan recovers *nothing* of the kind, exactly because the Indo-Aryan palatalisation of velars pre-dates attested Vedic — the change is not on that branch to find.

The lesson: direction recovers precisely the conditioned laws that occurred on the attested-ancestor$\to$daughter branch, and no others. Symmetry hides those laws; an attested ancestor — orientation without reconstruction — brings them back.

We are careful about scope. "History" here means *conditioned, directional sound laws*, not genealogical relatedness, which symmetric methods do capture (the companion finds symmetric genealogical signal in Austronesian). The claim is about the sound-law layer, and it is bounded accordingly.

1. Introduction: the arrow of time as a hidden variable

The companion paper proposed a change of order. Instead of reconstructing an ancestor and deriving the documented forms from it, it first builds a mathematical object from the attested correspondences alone, and only afterward sets it beside the historical apparatus as an independent check.

That object is the repertoire O. Each correspondence between two aligned sounds a, b becomes an *operator* $\Delta(a, b)$ — the set of phonological features on which the two sounds differ. A family is then a region of a feature space, and one can measure its shape.

The companion paper found that shape behaves representationally: combine two operators and you land back in the repertoire more often than chance, and that closure does not obviously track genealogy. It left a puzzle. If the shape of the correspondence space is hard to pin on genealogy, where do the regular sound laws — a catalogue of hundreds — actually show up?

The answer is almost embarrassingly simple, and it hides in the definition of the operator.

$\Delta(a, b) = \Delta(b, a)$.

The operator is symmetric. It records *that* two sounds differ, and in which features, but not *which way* the change ran, nor *what neighbourhood* licensed it. A conditioned sound law is exactly a statement about orientation and environment: Latin $/k/$ *becomes* an affricate *before a front vowel*. Strip the direction and the trigger and what remains — "k and an affricate differ in continuancy and place" — is a fact about two sounds, not about a change. The law is quotiented out before measurement begins.

This paper makes that precise and tests it. We keep the symmetric object; then we break its symmetry in the one place where doing so smuggles in no reconstruction: a family whose ancestor is *documented*. Latin is written down. Its Romance daughters are written down. The direction from Latin to Romanian is not inferred — it is attested. The same holds for Vedic and for Old Church Slavonic. That lets us orient the operator, restore the environment, and ask whether the conditioned laws return.

They return — and, revealingly, only where the change happened on the branch we are looking at.

A scope note, up front. By "history" we mean the *conditioned, directional sound laws* that the symmetric view cannot represent. We do not mean genealogical relatedness or subgrouping. Symmetric, distance-based methods capture genealogy well — that is most of computational phylogenetics — and the companion paper itself finds genealogical signal in the *symmetric* Austronesian repertoire. The blindness we demonstrate is specific to the sound-law layer, and we keep the claim there.

1.1 Relation to the companion paper and to prior work

This is a second study on the same object, not a revision of the first. The operator, the feature set, and the alignment are carried over unchanged from the companion paper [23], to which we refer for their motivation and for the null-model apparatus behind the symmetric results.

Two commitments are worth stating for the historical linguist. First, we do not claim to have discovered that Latin $/k/$ palatalises; that is textbook Romance philology [24, 25]. Recovering it is our *positive control* — evidence that the directed method has power on a case whose answer is known — not our result. The result is the *pattern* across three ancestors: direction recovers conditioned laws exactly on the branches where they occurred.

Second, conditioned correspondences are not new terrain for computation. List [29] infers context-sensitive correspondence patterns across many languages; probabilistic reconstruction models descent as a directed stochastic channel [28]; and work on the regularity and directionality of sound change [27, 30] long predates us. Our narrow contribution is to show, on documented data and against a *correctly clustered* null, what is lost when direction is integrated out, that an attested ancestor is enough to get it back, and that recoverability has a clean temporal boundary.

2. Two views of a correspondence

Fix two aligned segments a, b with panphon feature map ϕ.

The symmetric operator is the one the companion paper studies:

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

It is a contrast, not a change. By construction $\Delta(a,b) = \Delta(b,a)$, and it carries no environment.

The directed operator records a source, a target, and the neighbourhood in which the pairing occurred:

$$T = \big(a \to b \mid \text{env}\big), \qquad \text{env} = (\text{left neighbour, right neighbour, position}).$$

The two are related by a forgetful map. Symmetrization is the quotient that identifies T with its reverse $T^{-1} = (b \to a \mid \cdot)$ and integrates out the environment: $\Delta = \sigma(T)$.

This is the whole mechanism in one line, and we are candid that it is definitional: any information that lives in the fibre of σ — orientation, or environment — cannot survive into Δ. What is *not* definitional, and is the empirical content of the paper, is the consequence: because named sound laws are overwhelmingly oriented or conditioned, the symmetric repertoire should look representational while the directed view should surface the laws — *and* the directed view should surface a law only when the change occurred on the branch being directed. The Indo-European data test that consequence, including its boundary.

3. Data and method

Carried over. The operator, the 12-feature panphon subset (`cont, voi, nas, ant, cor, lab, back, round, strid, hi, lo, son`), and the Needleman–Wunsch alignment with feature-distance cost are those of the companion paper. No protoforms, laws, or sound classes enter the computation.

Cognacy. Expert cognacy is primary here: IE-CoR [6] throughout the symmetric results (§4.1–4.2), with the statistical (LexStat) pipeline used only in the controlled comparison that isolates the effect of cognacy method.

The directed corpora. For §4.3 we take three attested ancestors and their *modern* daughters in IE-CoR (modern only, to avoid counting a lineage twice at two time depths):

- Latin $\to$ Italian, Spanish, Portuguese, Catalan, French, Romanian, Neapolitan.
- Vedic (Vedic: Early) $\to$ Assamese, Bengali, Bhojpuri, Hindi, Kashmiri, Marathi, Nepali, Punjabi, Sinhalese, Urdu.
- Old Church Slavonic $\to$ Belarusian, Bulgarian, Czech, Kashubian, Macedonian, Polish, Russian, Serbo-Croat, Slovak, Slovene, Lower and Upper Sorbian, Ukrainian.

Within each expert cognate set we align the members by multiple-sequence alignment and read, at every consonant column, the correspondence between the ancestor segment and each daughter segment, together with the ancestor's neighbourhood. Because the ancestor is documented, its neighbourhood *is* the environment of the change.

A caveat on the attested ancestor. "Attested" is not "identical to the true spoken ancestor." IE-CoR Latin is *Classical* Latin, and Romance descends from spoken/Vulgar Latin; the palatalisation is a Proto-Romance change post-dating the Classical standard. Classical Latin is thus a *proxy* — a documented near-ancestor that supplies orientation without a reconstructed *form*. The palatalisation result survives because its trigger, the front vowels e, i, is inherited faithfully; but we do not overstate "attested" into "reconstruction-free in every respect." Choosing the reference stage is itself a modelling choice.

The null, done at the right unit. This is the methodological core, and it corrects the naive test. The trigger is read from the *ancestor* row, so it is identical across all daughters of one cognate-set column, and the daughter reflexes of one etymon are correlated. Correspondence tokens are therefore not exchangeable. We permute at the (cognate-set, ancestor-column) cluster level — one trigger label per cluster, shuffled across clusters, then expanded to tokens — with $B = 10{,}000$. We report the number of independent clusters beside the token count, so the effective sample size is visible. Mutual information is reported both as the plug-in estimate and with the Miller–Madow finite-sample bias correction; inference rests on the permutation p, not on the point estimate.

The trigger, kept clean. The palatalisation contrast is *front vowel* versus *non-front vowel* — restricted to cases where a vowel follows — so that intervocalic lenition (following consonant or word edge) is not conflated with frontness. Word-initial position is included (the iconic *ce-/ci-* environment is not discarded). Across a source consonant's exploratory test we apply Benjamini–Hochberg correction; $/k/$ palatalisation is the one *pre-specified* positive control, tested alone.

Scope. Indo-European throughout. The companion's Austronesian arm is not re-run here.

4. Results

4.1 The symmetric repertoire under gold cognacy

The object survives the move to expert cognacy. Under IE-CoR gold cognate sets the Indo-European repertoire has $|O| = 65$ operators (support ≥ 8), with a composition-realization index $C(O) = 0.169$ and triple density $\tau = 0.166$. These are lower than the companion's Lexibank headline ($C(O) \approx 0.22$), and the gap has an honest explanation: it is corpus density, not cognacy method.

We built the repertoire two ways on the *same* IE-CoR forms — expert cognacy and LexStat clustering of those very forms — and compared them at matched size against the same null. Gold matches or exceeds LexStat at every size from sixteen operators up (triple-density permutation-Z of $+3.2, +3.3, +3.4$ for gold at $K = 16, 20, 24$ versus $+2.3, +1.6, +2.2$ for LexStat). Expert cognacy does not degrade the additive structure. What Lexibank had was more *data* — correspondence supports in the tens of thousands where IE-CoR reaches the hundreds — which fills the linear-algebraic structure in more completely. The gold headline is smaller and correct.

We report the empirical permutation p (0.005–0.014 at these sizes) rather than the Gaussian Z, because the null distribution over small operator subsets is discrete and right-skewed and the Z overstates significance.

4.2 Is the additive structure representational? The control is size-confounded

The companion paper read the additive structure as *representational* on the strength of a concept-permuted control D_R: shuffle the concept labels to destroy all cognacy, and the additive closure did not fall. We tried to harden that control under gold cognacy and found, honestly, that it does not yield a robust verdict — the comparison is confounded by repertoire size.

Concept-permutation produces a *larger* repertoire than expert cognacy at the same support threshold (about 104 operators versus 65), because shuffling manufactures extra ”recurrent” correspondences. Since $C(O)$ depends on both size and the support distribution, size-matching the two repertoires gives opposite answers depending on *how* one matches:

matching of D_R to D_G's size (65)	D_R $C(O)$	vs $D_G = 0.169$
none (native, 0.207	O higher, but larger	≈ 104)

Latin /k/: the same correspondences, with and without the arrow of time

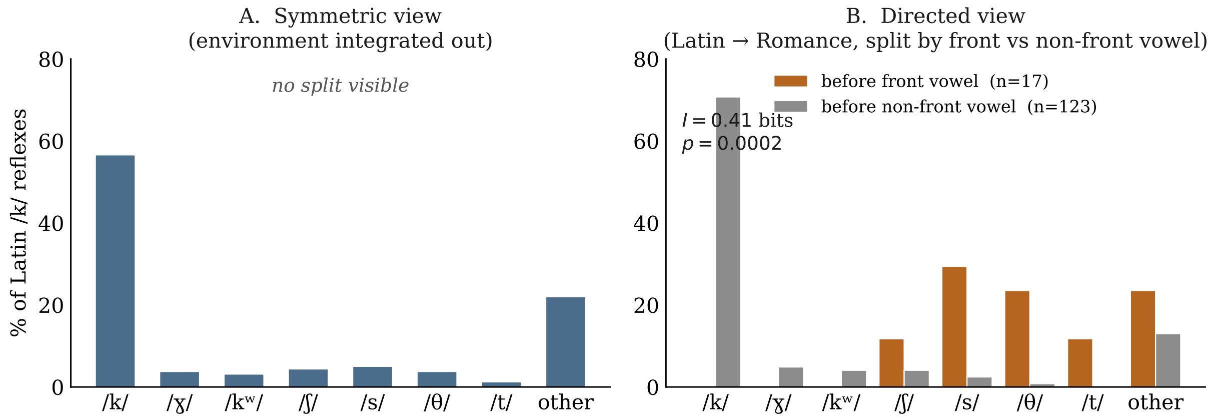

Figure 1: The same Latin $/k/$ correspondences, with and without the arrow of time. Panel A (symmetric). The marginal distribution of daughter reflexes, orientation and environment integrated out — all the symmetric operator Δ can see. The velar $/k/$ dominates a blended distribution; no conditioning is visible. Panel B (directed). The identical events, split by whether the following Latin vowel is front. Before a front vowel the reflexes are coronal and sibilant ($s, \theta, \int$); before a non-front vowel they stay velar (k). Mutual information 0.41 bits (Miller–Madow 0.33), cluster-level permutation $p = 0.0002$. Nothing about the data differs between the panels except whether the arrow of time is kept.

matching of D_R to D_G's size (65)	D_R $C(O)$	vs $D_G = 0.169$
top-65 by support (raise the threshold)	0.194	higher — but this is D_R's most-closed core
random 65-operator subsets (300 permutations)	0.131 $[0.098, 0.166]$	lower — D_G at the 99th percentile

Raising the threshold keeps D_R's high-support, low-weight operators (the additively central ones), inflating its $C(O)$; random subsampling dilutes with the tail, deflating it. The two fair-looking matchings disagree, so the $C(O)$ comparison cannot settle whether the additive structure is representational or carries genealogical signal. We therefore withdraw the strong "representational, not genealogical" reading at the level of $C(O)$.

What survives is weaker and honest: the additive structure is *real* (§4.1, it beats the opportunity-set null), but the representational-versus-genealogical question is not decidable from the symmetric closure statistic. The next section shows where that question *is* decidable — under direction.

4.3 Direction restores the laws — on the branch where they occurred

Now we break the symmetry, across three attested ancestors, with the cluster-level null.

Latin $\to$ Romance (the pre-specified control). Latin $/k/$ before a front vowel yields coronal and sibilant reflexes — $s, \theta, \int$ — while before a non-front vowel it stays velar k (70%). Mutual information 0.41 bits (Miller–Madow 0.33), cluster-level $p = 0.0002$ over 29 independent columns. The textbook Romance palatalisation is recovered by a pipeline told nothing about it, and — symmetrized — unable to see it. No other Romance consonant survives Benjamini–Hochberg correction: the directed Latin$\to$Romance signal is *one clean law*, exactly the one the handbooks name.

Old Church Slavonic $\to$ Slavic (the branch is rich). Here the conditioned change happened *after* the attested ancestor, in the daughters — the Slavic softening of consonants before front vowels — and the directed view recovers it at scale. Ten source consonants show significant front-vowel conditioning, all surviving Benjamini–Hochberg:

OCS source	clusters	I	I (MM)	p_{clust}	q_{BH}
$/l/$	34	0.337	0.322	0.0001	< 0.001
$/n/$	29	0.273	0.263	0.0001	< 0.001
$/r/$	41	0.219	0.193	0.0001	< 0.001
$/\upsilon/$	23	0.226	0.194	0.0001	< 0.001
$/t/$	68	0.125	0.106	0.0001	< 0.001
$/p/$	19	0.125	0.118	0.0004	0.001
$/s/$	22	0.114	0.090	0.0052	0.009
$/m/$	18	0.095	0.081	0.0074	0.011
$/d/$	25	0.125	0.108	0.0227	0.030
$/b/$	8	0.279	0.263	0.0260	0.031

The reflex splits are the Slavic hard/soft system read straight off the corpus. $/n/$ goes to the palatal ɲ$/n^j$ before a front vowel but stays plain n (97%) elsewhere; $/l/$ goes to soft l^j before front and velarised, hard l^γ elsewhere; $/r/$ gives the palatalised ɾ^j/ʐ (compare Czech *ř*, Polish *rz*) before a front vowel. Direction reads the softening off the corpus; the symmetric operator averages it away.

Vedic $\to$ Indo-Aryan (the branch is empty — and that is the point). Vedic $/k/$ before a front vowel stays velar (k 73%, k^h 17%); nothing survives correction. This is not a failure of the method — it is its boundary made visible. The Indo-Aryan palatalisation of velars is an Indo-Iranian change that *pre-dates* attested Vedic; Vedic already shows its palatal reflexes, so the change is not on the Vedic$\to$daughter branch to recover. The corpus is also smaller (96 cognate sets). Where Latin *precedes* the palatalisation it feeds, we recover it; where Vedic *follows* its palatalisation, we cannot — because there is nothing left on that branch to find.

The pattern is the finding. Across three attested ancestors, direction recovers conditioned laws precisely on the branches where the change occurred: the palatalisation Latin precedes (recovered), the softening the Slavic daughters underwent after OCS (recovered, tenfold), and the palatalisation Vedic already completed (absent). The symmetric view sees none of these; the directed view sees exactly those that lie on its branch.

4.4 The environment axis and the recoverability boundary

This closes a loop with the companion series, which argued that a protoform-agnostic pipeline — by dropping vowels, prosody, and direction — goes blind to exactly the law-classes those ingredients condition, so its null results map which missing ingredient each law needs. Section 4.3 is the constructive half: restore direction and the front-vowel environment, and the conditioned laws on the branch light up.

The Vedic null sharpens the map. Direction is necessary but not sufficient; the change must also *post-date the reference stage*. An attested ancestor recovers the laws downstream of it and is silent about the ones upstream. That is a usable boundary, not a defect: it tells you which reference stage would recover which law.

5. Discussion

The contribution is a duality with a boundary.

A phonological correspondence carries two separable kinds of information. Its symmetric part — which features differ — behaves like a fact about representation and inventory; it is additively organised and stable under expert cognacy, and (§4.2) the symmetric closure statistic cannot by itself separate representation from genealogy. Its directed part — which way the change ran, in what environment — is where conditioned sound laws live, and they become legible the moment an attested ancestor orients the pair.

We are deliberately narrow about the word "history." Symmetric, distance-based methods recover genealogical relatedness and subgrouping — that is their core business, and the companion paper finds such signal in the *symmetric* Austronesian repertoire. What symmetry hides is not genealogy but the *conditioned, oriented sound law*. The correct general statement is therefore scoped: symmetric correspondence invariants are blind to the conditioning and direction of individual laws, and recover them only once orientation is supplied.

Two things follow for the protoform-agnostic programme. First, you do not need reconstructed protoforms to see conditioned history — you need *orientation*, and an attested ancestor supplies it directly, as Latin, Vedic, and OCS do. Second, recoverability is bounded in time: the reference stage you can orient from determines the laws you can recover. Where no ancestor is attested, a dated stratum or a chosen reference stage — a gauge, in the sense the research programme develops — would play the same role, and the Vedic result predicts exactly what such a choice can and cannot reach.

6. Limitations and scope

Three attested-ancestor families is a pilot, not a survey; the Vedic arm is small (96 cognate sets) and yields a null we read structurally, not as an estimate. The directed test pools reflexes across a family's daughters, so it detects *that* an ancestral consonant is conditioned, not how each daughter resolves it; a per-daughter grammar and a minimum-description-length rule treatment belong to the systematic follow-up (Paper 2 of this series). Mutual information detects conditioning; it is not a rule with scope and ordering, and we report Miller–Madow-corrected values with permutation inference precisely because the plug-in estimate is biased upward. The attested ancestors are documented proxies for their spoken predecessors, not the predecessors themselves (§3). And the symmetric $C(O)$ comparison of §4.2 is inconclusive by design once size is controlled — a limitation we state as a result.

None of these touch the paper's core: that the symmetric operator is blind to conditioned laws by construction, and that direction from an attested ancestor recovers exactly the laws on its branch — recovered on Latin$\to$Romance and OCS$\to$Slavic, correctly absent on Vedic$\to$Indo-Aryan.

7. Reproducibility

All results run from `transformations/` against IE-CoR. The symmetric gold repertoire and the corpus regimes are produced by `src/regimes.py`; the size-confounded D_R comparison by `paper2/dr_control_v2.py` (random-subset and top-K matchings, 300 permutations); the gold-versus-LexStat survival comparison by `paper2/paper1_gold_recheck.py` and `paper2/paper1_gold_matched.py`; the directed multi-ancestor test — cluster-level permutation, Miller–Madow MI, front-vs-non-front-vowel trigger, BH correction — by `paper2/directed_multi.py`. Results are cached in `data/db/paper2.db` (tables `paper1_gold_recheck`, `paper1_dr_v2`, `directed_multi`). Feature set: panphon 12-feature primary subset. Cognacy: IE-CoR expert sets; LexStat threshold 0.55, 100-run scorer. Permutation seed 20260804; $B = 10{,}000$.

References

The companion paper [23] gives the full apparatus behind the symmetric object and its null models.

[6] Heggarty, P., Anderson, C., Scarborough, M., et al. (2023). Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. *Science*, 381, eabg0818. (Dataset: IE-CoR.)

[23] Toledo Martínez, A. (2026). *Additive Structure of Phonological Correspondences: A protoform-agnostic method for discovering mathematical patterns in documented linguistic systems.* Preprint (companion, this series).

[24] Alkire, T., & Rosen, C. (2010). *Romance Languages: A Historical Introduction.* Cambridge University Press. (Palatalisation of Latin velars before front vowels.)

[25] Recasens, D. (2020). *Palatalization in the Romance Languages.* Oxford Research Encyclopedia of Linguistics. Oxford University Press.

[27] Hoenigswald, H. M. (1960). *Language Change and Linguistic Reconstruction.* University of Chicago Press. (The directionality of merger and split.)

[28] Bouchard-Côté, A., Hall, D., Griffiths, T. L., & Klein, D. (2013). Automated reconstruction of ancient languages using probabilistic models of sound change. *PNAS*, 110(11), 4224–4229. (Descent as a directed stochastic channel.)

[29] List, J.-M. (2019). Automatic inference of sound correspondence patterns across multiple languages. *Computational Linguistics*, 45(1), 137–161. (Context-sensitive correspondence patterns across many lects.)

[30] Hruschka, D. J., Branford, S., Smith, E. D., Wilkins, J., Meade, A., Pagel, M., & Bhattacharya, T. (2015). Detecting regular sound changes in linguistics as events of concerted evolution. *Current Biology*, 25(1), 1–9.

Note (pilot preprint): bibliographic details verified against publisher records where possible; a few page numbers should be confirmed at typesetting.