A Point Cloud of Languages

A reconstruction-free, feature-decomposable dissimilarity between language systems: an Indo-European study, with a portability check on two further families

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

Abstract

We define a dissimilarity between two documented language systems that uses no reconstruction, no family tree, and no branch labels. From concept-aligned wordlists we detect coderivative sets statistically (LexStat), align each language pair, and average — over aligned consonant slots — the number of primary phonological features that differ. All analysis matrices are complete observed matrices with zero imputed values. On 50 Indo-European doculects, a system's nearest neighbour shares its Glottolog-derived branch 97.7% of the time on the 44 systems whose branch has more than one member (creoles excluded; 97.9% over all doculects; silhouette +0.34); a label-permutation test on the fixed matrix puts both far above chance ($p_{\mathrm{MC}} \approx 10^{-4}$, Monte-Carlo floor at $B = 10{,}000$; chance purity 0.16), and the result survives macro-averaging, removal of 14 near-duplicate systems (collapse criterion: the dissimilarity itself, a disclosed circularity), and the minimum shared-slot threshold up to values that shrink the sample, and shows no linear correlation with the number of retained consonant slots. We are deliberately careful about what this shows. Branch recovery is not a novelty: a plain edit distance matches our purity and an unfiltered all-concept baseline matches or beats it — so the branch signal is generic lexical-phonological similarity, not an artefact of the coderivation filter. The claimed contribution is feature-level attribution: the dissimilarity decomposes additively into per-feature components (demonstrated on real pairs; the attribution is conditional on the chosen alignment), computed on the same correspondence units, by construction, as the operator repertoire of the pilot study. As a portability check (Appendix B), the same pipeline runs unchanged on Nakh-Daghestanian, which shows strong branch association, and on a small complete Austronesian subset (14 of 45 doculects survive the completeness requirement, and they fall almost entirely within one Oceanic subgroup) which shows positive but weaker association — too small and branch-imbalanced for a robust portability claim. Each family is analysed in a separate field (never a shared distance space); Appendix B compares only scale-normalised descriptive summaries and draws no historical, causal, or relational cross-family inference. We make no areal claim.

1. Introduction

The pilot study argued that a family's inventory of feature-difference operators, read at the level of *types*, is representational rather than genealogical: the operators are largely those general phonology makes available, and the genealogical signal lives one level up, in how those operators are *distributed* across languages. This paper looks at the simplest projection of that distribution — a pairwise dissimilarity between whole systems — and asks a deliberately modest question: is branch structure legible in it, built without any historical apparatus?

The stance is the programme's: discover first, contrast later. We construct the map from correspondences in concept-aligned lexical data and consult branch labels, contact history, and baselines only

afterwards. We say coderivative, not *cognate*: operationally, an *algorithmically linked form set* (recurrent correspondence), with no claim of single-ancestor descent. One tension must be conceded openly: the detector we use, LexStat, was designed and validated as an automatic *cognate* detector — a single-ancestor construct. What we take from it is its output — clusters bound by recurrent segmental correspondence — treating that output as a correspondence structure rather than a descent claim; our "reconstruction-free" applies to the *inputs* (no protoforms, no tree, no labels), not to the intellectual lineage of the detector. Readers who prefer to read "coderivative" as "statistically detected cognate" lose nothing in the mathematics; the terminological choice marks a difference of interpretive commitment, not of computation.

Two honesty commitments frame the paper. First, the input is not "phonology alone": it is phonological dissimilarity computed *within concept-matched, statistically inferred lexical correspondences*. "No reconstruction", "no tree", and "no branch labels" are the defensible claims. Second, recovering known branches is a sanity check, not the contribution — §5 shows simpler baselines do as well or better on that task. What the representation claims instead is feature-level attribution (§5.1): unlike an edit distance, d factors exactly into which phonological features carry the difference.

Relation to existing distance traditions. The method inherits from, and should be read against, several lines: ASJP-style normalized Levenshtein distances between wordlists (LDN/LDND), ALINE and SCA/PMI-based alignment scoring, and phonetic-distance dialectometry. Credit where due: ALINE is itself feature-based (its alignment scores decompose into feature saliences), and phonetic dialectometry has long used feature-based segment distances; only the Levenshtein/ASJP line is genuinely feature-opaque. What our formulation adds relative to the feature-aware lines is not the use of features but the *bookkeeping*: the final system-level dissimilarity is an exact additive sum of per-feature, per-pair components ($d = \sum_f d_f$, §5.1) over correspondence slots within inferred coderivative sets — so any entry of the matrix can be opened into a feature ledger tied to specific correspondences, the unit on which the pilot's operator analysis is defined. We do not claim superior classification (§5 shows parity); the contribution is this attribution layer, and its scope relative to ALINE-type scoring is deliberately narrow.

2. A dissimilarity between systems

Let $\phi_f(x) \in \{+, -, 0\}$ be primary feature f of segment x (panphon). For aligned segments a, b,

$$\text{fd}(a, b) = \big|\{f \in \text{PRIM} : \phi_f(a) \neq \phi_f(b)\}\big|,$$

over the twelve consonantal features `cont, voi, nas, ant, cor, lab, back, round, strid, hi, lo, son`.

Family membership is used only to *delimit each field* — i.e. to select which doculects enter a given run; no internal branch classification, tree, or protoform enters the dissimilarity computation. Coderivative sets are detected with LexStat (no etymologies, protoforms, or family labels supplied). Within each set, each language pair is aligned by a feature-cost Needleman–Wunsch, and we average fd over slots where both aligned segments are consonants:

$$d(\ell, \ell') = \text{mean}\{\text{fd}(a, b) : (a, b) \text{ a shared consonant slot of } \ell, \ell'\},$$

defined only when at least `MINSLOT` slots are observed. Observed range here $\approx [0.5, 3.5]$ (theoretical $[0, 12]$).

Missing pairs are never imputed. If a pair has fewer than `MINSLOT` shared slots, d is undefined for it. Form the *observed-distance graph* — one vertex per doculect, an edge wherever d is defined — and take its maximum clique: the largest set of doculects that are all pairwise observed. The analysis matrix is the (complete) submatrix on that clique, computed exactly by Bron–Kerbosch (feasible at these sizes;

$\leq$ 50 vertices). When more than one maximum clique exists we report the count and take the first in input order; for the fields here the choice is inconsequential (Indo-European and Nakh-Daghestanian graphs are already complete, so the clique is the whole sample). This replaces an earlier internal version that filled undefined entries with the global mean — which manufactures artificial equidistance and was removed (see Appendix B, where its effect on Austronesian is documented rather than hidden). Every matrix reported here has zero imputed entries, the surviving doculects are listed explicitly (Appendix A / B), and each field reports how many the clique requirement removed.

We call d a dissimilarity, not a metric. Because each pair is computed on its own set of shared slots, the triangle inequality and identity-of-indiscernibles are not guaranteed. It is symmetric and non-negative; that is all we assume. Known blind spots, stated plainly: slots where a consonant corresponds to a gap or to a vowel do not enter the average, so consonant loss, epenthesis, and consonant–vowel change are not penalized — two systems can look close on the consonants that survive alignment even if much structure was lost. Vowels also do not contribute directly to d, but they remain in the full-form alignment and its feature cost, and can therefore influence *which* consonant slots end up paired — so "consonants only" describes the scored quantity, not the alignment. A companion indel/gap statistic is future work; per-pair gap counts are summarised in §4 and exported with the reproducibility materials (`data/results/pair_stats_ie.csv`).

3. From dissimilarity to picture

We embed $D = [d(\ell, \ell')]$ with classical (Torgerson) MDS and build a $k = 3$ nearest-neighbour network for display; purity is a $k = 1$ statistic computed on the full matrix. The plane is only a projection: on the Indo-European analysis matrix the first two MDS axes carry 28.2% of the positive variance, with 6.2% negative inertia — D is only approximately Euclidean, and the 2-D figure is a lossy shadow. All statistics use the full matrix. (LexStat's scorer is seeded, so the figure and the analysis matrix are built from the *same* deterministic run and their summaries agree; re-running with a different `SEED` is one of the robustness checks.)

4. Results (Indo-European, 50 doculects, 4 creoles)

Branch labels (scoring only) are derived from Glottolog by a sample-dependent tree cut (nodes covering at most half the sample); this makes "branch" a function of the sample, which is one reason the cross-family §7 is kept exploratory. Creole policy, stated once: Glottolog classifies creoles under their lexifier's branch (Jamaican Creole $\rightarrow$ Germanic), and we keep those labels wherever creoles appear; but because their genealogical status is contested, the headline purity excludes them (they remain present in the matrix as potential neighbours, and in the all-doculect row). All numbers in this section come from the analysis matrix (`make analysis`).

Significance. A label-permutation test on the fixed matrix: the matrix comprises the 46 non-creole systems; the statistic is nearest-neighbour purity over the 44 of them whose branch has more than one member (a singleton can never match, so the two Albanian-label singletons are excluded from the numerator and denominator but remain as potential neighbours); the null shuffles the branch labels of all 46 systems, 10,000 times. Result: purity 0.977 against a null of 0.155 ± 0.067 and silhouette $+0.344$ against -0.194 ± 0.036, with no permutation reaching the observed values ($p_{\mathrm{MC}} = 1/10001 \approx 1.0 \times 10^{-4}$, the Monte-Carlo estimate with $(k+1)/(B+1)$ correction at $B = 10{,}000$). Caveat: label exchangeability is an approximation — doculects within a branch often share dataset and transcription conventions; blocked-by-source permutation is future work. (Three related "chance" figures appear in this paper and are different quantities: 0.155 is this permutation-null mean; 0.158 is the analytic expected same-branch probability of a random neighbour; §7's 0.16 is a separate 2,000$\times$ permutation

null on the network matrix. They agree to two decimals here, but they are not the same estimator.)

Purity, reported at several denominators.

Denominator	purity
All doculects (multi-member branches)	0.979 (47/48)
Multi-branch, creoles excluded	0.977 (43/44), descriptive Wilson 95% [0.88, 1.00] (ignores neighbour dependence)
Macro-average by branch (equal weight per branch)	0.952
After collapsing near-duplicate pairs ($d < 1.3$; 14 dropped)	0.967 (29/30)

The near-duplicate ablation uses d itself as the collapse criterion, which is admittedly circular in a limited sense; a metadata-based collapse (same Glottocode / documented dialect relation) is the better design and is left as future work — the manifest (Appendix A) contains the fields needed to do it. One labelling artefact is noted and neutralized: Glottolog splits "Albanian (Tosk)" and "Standard Albanian" into different top labels although they are one language; as two singletons they are excluded from multi-branch purity, so the reported purity is *conservative* — merging them would add a correct pair.

Coverage. Median 661 consonant slots per pair (range 142–5681), median 400 gaps, median 226 shared concepts; within this sample, d shows no linear correlation with the number of retained slots (`corr(d, #slots) = -0.07`). This does not rule out subtler coverage effects (variance, gap proportion, neighbour identity); per-pair coverage is published with the results so they can be probed.

5. Baselines — what the representation does and does not add

On the multi-branch non-creole set (n = 44), same corpus and concepts:

Method	purity	silhouette
Per-language marginal profile	0.477	−0.126
Edit distance on consonant skeletons	0.977	+0.205
All-concept dissimilarity, no LexStat filter	1.000	+0.324
Ours (pairwise feature dissimilarity)	0.977	+0.344

Three readings. (i) Branch recovery is generic: edit distance ties our purity; the unfiltered all-concept variant matches or beats it. (ii) The structure is not generated by the LexStat coderivation filter: removing that similarity-based filter *strengthens* the structure rather than destroying it. (This rules out the filter as the source; it does not remove every alignment↔scoring dependency, since both still use phonological information and the baselines are not perfectly matched.) (iii) The +0.020 silhouette edge over the all-concept baseline is small and untested for stability (a paired concept-bootstrap is future work), and the baselines are not perfectly matched (the all-concept variant uses one form per concept); we do not rest any claim on that margin. (iv) The marginal baseline (0.477) confirms that the pairwise comparison keeps information a per-language profile discards. Why keep the LexStat-filtered variant as the main object at all? An honest answer first: not because of decomposability — the feature decomposition of §5.1 is a property of the feature-difference metric and would hold equally for the unfiltered all-concept variant computed the same way. The filtered variant is retained for *continuity of units*: its slots are correspondences within coderivative sets, which is — by construction, not by demonstrated result — the same unit the pilot's operator analysis is defined on, so the two studies' quantities compose.

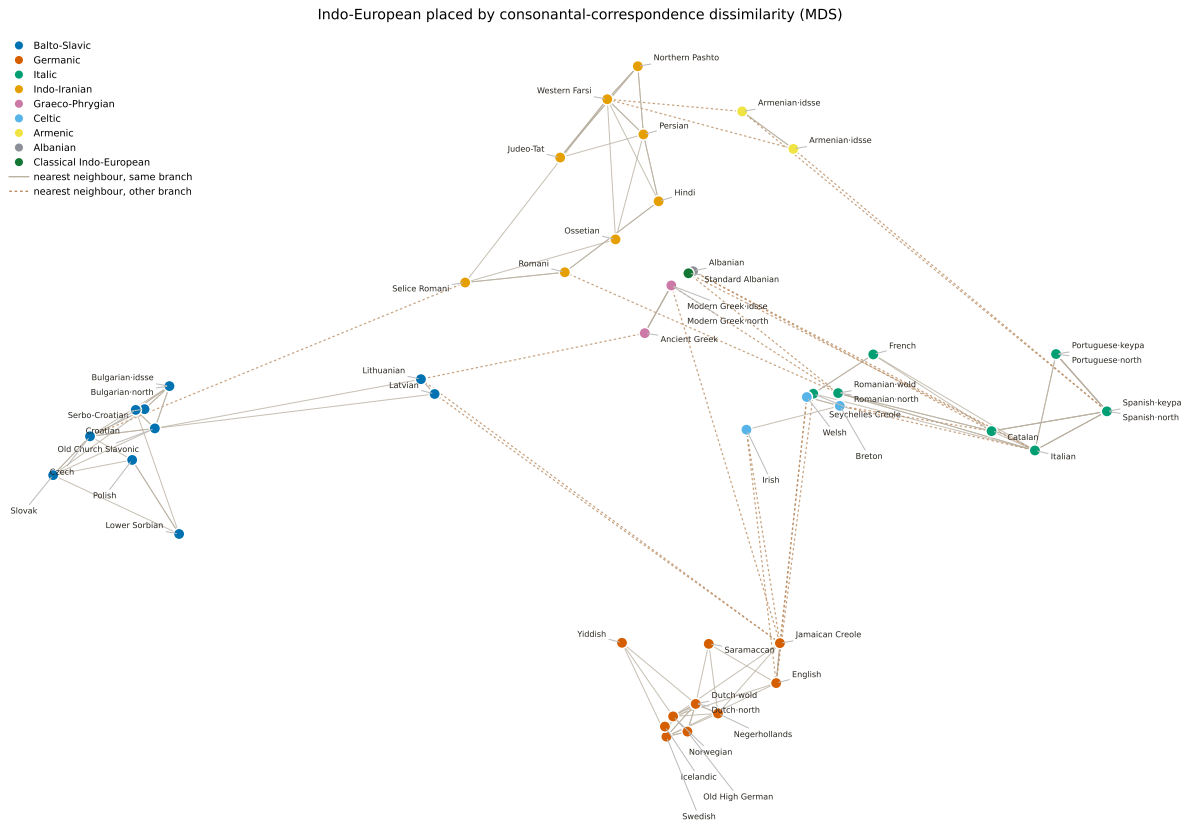

Figure 1: Indo-European doculects placed by the consonantal-correspondence dissimilarity (classical MDS of the network matrix; the plane shows under a third of the variance, so central overlaps are partly projection). Nodes coloured by branch — colours added after layout. Solid links join same-branch nearest neighbours; dashed links join the few cross-branch nearest neighbours. Every node is labelled, with leader lines separating labels in dense clusters and a dataset suffix distinguishing same-name twins (e.g. Spanish·keypa / Spanish·north); an interactive version — hover for each node's three nearest neighbours — is the published artifact.

5.1 Feature-level decomposition — demonstrated, not promised

By construction $d(\ell, \ell') = \sum_f d_f(\ell, \ell')$, where d_f is the fraction of aligned consonant slots on which feature f differs. Real values from the analysis matrix (`src/decompose.py`):

Feature f (d_f)	Spanish–Portuguese	Persian–Hindi	Spanish–Hindi
cont	0.20	0.30	0.39
voi	0.15	0.28	0.35
nas	0.06	0.15	0.23
ant	0.17	0.24	0.32
cor	0.16	0.27	0.29
lab	0.10	0.18	0.19
back	0.09	0.20	0.21
round	0.00	0.01	0.01
strid	0.17	0.16	0.19
hi	0.18	0.27	0.42
lo	0.07	0.12	0.21
son	0.14	0.26	0.37
total d	1.50	2.43	3.18
consonant slots	4117	512	552
doculects	`keypano`	`idsseg/neur`	`keypano/neur`

(Doculects are named to avoid ambiguity — the corpus holds two Spanish and two Portuguese; here both are from the `keypano` dataset.) The sum reproduces the total exactly. We read it cautiously: in each pair *continuancy* and *high* are the largest individual components, but several other features (anterior, coronal, strident, voicing) contribute at comparable levels, so these are aggregate feature-difference frequencies, not process labels — identifying a change like lenition would require inspecting the underlying correspondences (the per-concept, per-slot ledger is produced by `src/provenance.py` — `decompose.py` gives only the aggregate feature table above; a worked example is in Appendix C). One caveat is declared: the alignment is optimized with a feature cost, so per-feature attribution is conditional on the chosen alignment; a fixed-alignment sensitivity check is future work.

6. Cross-branch proximities: rare and insufficient to identify contact

Of the 50 systems, only three have a nearest neighbour in another branch — two of which are the Albanian labelling artefact of §4 (Glottolog splits one language into two top labels):

System	branch	nearest neighbour	branch	d	note
Irish	Celtic	Jamaican Creole	Germanic	2.76	no documented contact; sampling / lexical overlap / accidental proximity
Albanian (Tosk) ↔ Standard Albanian	(label split)	each other	—	0.80	same language; artefact, discounted

None of the cross-branch nearest neighbours corresponds to a documented contact situation. (An earlier, unseeded run had shown a Romani → Romanian link — a real Balkan-contact case — but it does *not* survive the seeded canonical run, where Romani's nearest neighbour is same-branch; we mention this precisely because a result that flips between runs is not one to build on.) The Irish case is computed on the full, creole-inclusive matrix (not the projection), so it is a genuine proximity under this dissimilarity, most plausibly reflecting sampling, lexical overlap, or corpus effects. Note

an asymmetry we flag rather than hide: because the headline purity of §4 excludes creoles, this one anomalous proximity cannot lower that headline — it is visible only here. We make no areal claim: a proper test requires a contact matrix defined *before* inspecting the map — deferred.

A portability check — running the identical pipeline on two further families (Nakh-Daghestanian and a small Austronesian subset), held in separate fields — is reported in Appendix B. It is not part of the paper's claims.

7. Controls and robustness

The label-permutation test (§4) is the primary significance check. The following controls were run on the n=50 configuration with the zero-imputation policy (`make controls`; outputs in `data/results/controls_ie_n50.txt`):

- Reproduction: an independent control pass gives purity 0.977, silhouette +0.324 — matching the seeded canonical values of §4 (0.977 / +0.344) to run-to-run tolerance.
- Concept-permuted ablation: shuffling which forms share a concept and rerunning the full pipeline collapses purity to 0.28 (range 0.20–0.38 over three reruns) and silhouette to ≈0 — toward, though not fully to, the chance level 0.16. The residual above chance is expected: concept-shuffling destroys coderivative structure but preserves each language's segment inventory and phonotactics, which carry some branch signal on their own. Three reruns are too few for a null tail, so this is reported qualitatively; significance is carried by the 10,000× label-permutation test of §4, not by this ablation.
- Minimum shared-slot threshold: since the minimum observed per-pair coverage is 144 slots, thresholds below that are vacuous by construction; the sweep therefore extends above it. Purity is 0.977 at `MINSLOT` $\in \{20, 40, 100\}$ (n=50), 0.977 at 200 (n=48), and 1.000 at 400 (only 27 systems survive the completeness requirement). The result is not a knife-edge of the threshold. (The LexStat detection threshold `THR` is *not* varied here — a sweep over it, gap penalties, and alignment costs is future work.)
- Feature subsets: dropping stridency leaves the result unchanged (0.977/+0.323); dropping all five place features gives 0.955/+0.319; manner-only gives 0.932/+0.329 — mild variation, no knife-edge.
- Subsampling: drawing 22 of 50 systems 500× keeps purity at mean 0.934, 95% CI [0.79, 1.00].
- One doculect per Glottocode (metadata-based, addressing near-twin/duplication effects directly): collapsing the 7 duplicated Glottocodes (Bulgarian, Romanian, Spanish, Modern Greek, Portuguese, Western Farsi, Dutch) to one representative each leaves 39 systems with purity 0.973 (36/37) — the branch structure is not carried by duplicate doculects. (Representatives are chosen by a metadata rule — most shared concepts — not by the dissimilarity; the value is identical, 0.973, across all $2^7 = 128$ representative choices.)
- Leave-one-dataset-out (source-confound control): dropping each of the four contributing Lexibank datasets in turn keeps purity in [0.935, 0.976] — no single source drives the result.
- Branch-cut granularity: varying the Glottolog tree-cut fraction `TF_BRANCH_MAXFRAC` $\in \{0.25, 0.33, 0.5, 0.67\}$ (which changes what counts as a "branch") gives purity 0.932 / 0.977 / 0.977 / 0.977 — the result does not depend on the particular branch definition.

8. Limitations and scope

The input is concept-aligned lexical data, not phonology alone. The dissimilarity is corpus- and aligner-dependent, is not a metric, ignores gaps and consonant–vowel correspondences, and weights concepts by their consonant count (a concept-balanced variant is future work). The 2-D figures show under a third of the variance. Branch labels are a sample-dependent Glottolog cut, so purity is not directly comparable across differently-sampled families — which is why Appendix B is an exploratory portability check only and draws no cross-family inference. The twelve-feature instrument was not validated for cross-family measurement invariance (glottalization, uvularity and pharyngeal oppositions are not directly represented — relevant for Nakh-Daghestanian). Branch recovery is not distinctive versus edit distance; the near-duplicate collapse uses d itself; the silhouette margin over the strongest baseline is untested for stability. Direction and time are out of scope (the directed-layer paper).

9. Reproducibility

Pinned dependencies (`requirements.txt`, exact versions; Python 3.12.13). Each experiment's `MAXLANG` is fixed in the `Makefile`: `LEX_PATH=… make compute` reproduces the n=50 Indo-European field; `make compute-nd` / `make compute-an` the other fields (Appendix B); `make analysis` §4–§6; `make compare` the portability figure (all three slugs by default); `make controls` §7; `make manifest-all` the three manifests (Appendix A); `make figure` redraws from bundled results without the corpus. Analysis/control caches are keyed by `MAXLANG` so a different configuration cannot silently reuse them. Permutation seeds are fixed in the scripts; the maximum-clique selection is exact (Bron–Kerbosch). The one irreducible source of run-to-run variation is LexStat's stochastic scorer (`get_scorer`), which we do not seed — hence the small second-decimal differences between the analysis matrix and the independently-regenerated figure matrix, disclosed in §3. Complete observed matrices, coordinates, manifests, survivor lists and analysis logs are bundled under `data/results/`. Code MIT; text, figures, data CC BY 4.0.

10. References

1. Toledo Martínez, A. (2026). *Additive Structure of Phonological Correspondences* (pilot study). https://github.com/toledoal/phonological-correspondences
2. Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika* 17(4), 401–419.
3. Kruskal, J. B., & Wish, M. (1978). *Multidimensional Scaling*. Sage.
4. Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20, 53–65.
5. Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *JASA* 22, 209–212.
6. Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48(3), 443–453.
7. Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., & Levin, L. (2016). PanPhon: a resource for mapping IPA segments to articulatory feature vectors. *Proc. COLING 2016*, 3475–3484.

8. List, J.-M. (2012). LexStat: automatic detection of cognates in multilingual wordlists. *Proc. EACL 2012 Workshop LaTeCH*, 117–125.
9. List, J.-M., Forkel, R., Greenhill, S. J., Rzymski, C., Englisch, J., & Gray, R. D. (2022). Lexibank, a public repository of standardized wordlists. *Scientific Data* 9, 316.
10. Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2023). *Glottolog 4.8*. MPI-EVA. https://glottolog.org
11. Wichmann, S., Holman, E. W., Bakker, D., & Brown, C. H. (2010). Evaluating linguistic distance measures (LDN/LDND). *Physica A* 389(17), 3632–3639.

11a. Wichmann, S., et al. *The ASJP Database* (version consulted). https://asjp.clld.org

1. Kondrak, G. (2000). A new algorithm for the alignment of phonetic sequences (ALINE). *Proc. NAACL 2000*.
2. Jäger, G. (2018). Global-scale phylogenetic linguistic inference from lexical resources (PMI distances). *Scientific Data* 5, 180189.
3. Nerbonne, J., & Heeringa, W. (2010). Measuring dialect differences. In *Language and Space* (pp. 550–567). De Gruyter.
4. Thomason, S. G., & Kaufman, T. (1988). *Language Contact, Creolization, and Genetic Linguistics*. UC Press.
5. Schuchardt, H. (1922). *Hugo Schuchardt-Brevier* (L. Spitzer, Ed.). Niemeyer.

Appendix A — Doculect manifests

The doculects of all three fields, with Lexibank dataset, doculect ID, Glottocode, branch used for scoring, and concept count, are in `data/results/doculect_manifest_{ie,an,nd}.csv` (regenerate with `make manifest-all`). Each manifest lists the candidate pool (the top-`MAXLANG` doculects by form count); the analysed field can be smaller after the completeness requirement — in particular the Austronesian manifest lists 45 candidates while the complete-matrix field is n=14. Provenance is exposed so source/transcription confounds can be audited — e.g. several near-twin systems share a source dataset, and most Nakh-Daghestanian doculects come from a single segmented source, which the reader should weigh when comparing fields.

Appendix B — Portability check on two further families (not part of the paper's claims)

The identical pipeline runs on two other families, each on its own field (a separate distance space — never shared, which would assert a direct relation between families we do not claim). Each field's analysis matrix is the maximum observed clique (§2). We compare descriptive, scale-normalised outputs, but draw no historical or causal inference from their differences, because the samples are not commensurable (Indo-European mixes ancient, modern and creole doculects; Nakh-Daghestanian is dialect-dense and largely single-source; the Austronesian clique falls almost entirely within one Oceanic subgroup). Significance is each field's own 10^{-3}-resolution label-permutation test (2,000 permutations); *adj. purity* $= (P_{obs} - P_{null})/(1 - P_{null})$.

Field (own max-clique matrix)	n	eff. dim. (PR)	silhouette (p)	purity	chance	adj. purity (p)
Indo-European	50	13.3	+0.33 (p≈.0005)	0.96	0.16	0.95 (p≈.0005)
Nakh-Daghestanian	45	9.4	+0.31 (p≈.0005)	0.98	0.29	0.97 (p≈.0005)
Austronesian	14	3.8	+0.28 (p≈.003)	0.93	0.51	0.86 (p≈.009)

Reading. Nakh-Daghestanian shows strong branch association, confirming the pipeline is portable to a very different consonant system. The 14-doculect Austronesian subset shows positive, significant association, but it is too small and too branch-imbalanced (one Oceanic subgroup dominates, so chance purity is already 0.51) to support a robust portability claim; we report it for transparency, not as evidence. The surviving doculects of each field are listed in `data/results/cloud_{ie,nd,an}.txt`.

A correction made and disclosed. An earlier version of this section filled unobserved pairs with the global mean. For Austronesian this was fatal: 52.6% of its 990 candidate pairs were the fill constant, which *manufactures* a flat, nearly-equidistant, high-dimensional geometry — exactly the "signature" we had then reported and interpreted. That analysis was wrong and is withdrawn; the zero-imputation / max-clique policy above replaces it.

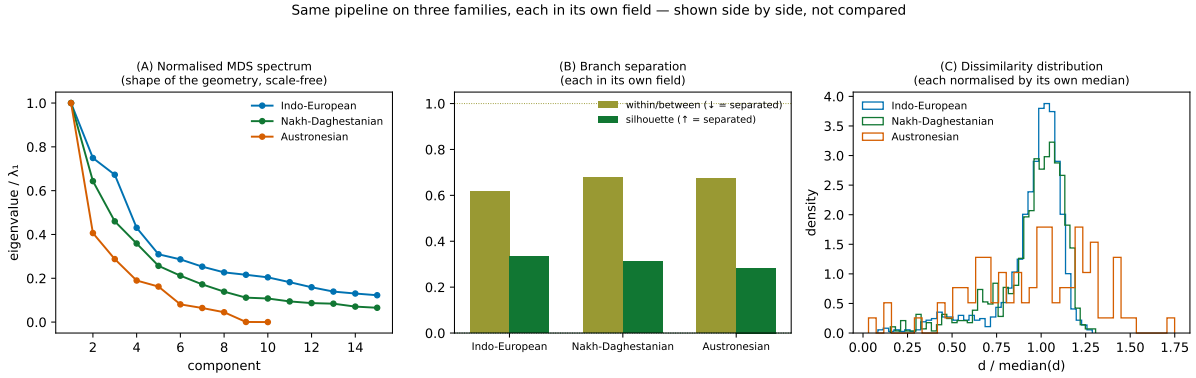

Figure 2: The same pipeline on three families, each on its own complete observed matrix, shown side by side (descriptive only; never a shared space, and no cross-family inference). (A) normalised MDS spectra; (B) branch separation; (C) each field's dissimilarity distribution normalised by its own median. The Austronesian field is n=14 and branch-imbalanced.

Appendix C — Provenance of the feature ledger (worked example)

Each unit of the dissimilarity opens not only into features (§5.1) but into the specific correspondences that produced them. `src/provenance.py` re-runs the *same* seeded LexStat pipeline, so it reproduces the exact coderivative sets (cogids), form pairs and consonant slots that build d — not a re-alignment. For Spanish–Portuguese it lists 4117 consonant slots across the shared cogids, and the per-feature sums reproduce the §5.1 row exactly (total $d = 1.499$; e.g. voicing $629/4117 = 0.153$, continuancy 0.198, strident 0.171). Concrete slots (concept, cogid, the two forms, the two segments, differing features):

Concept	cogid	Spanish	Portuguese	slot	differing features
accuse	31	akusaɾ	ɐkuzaɾ	s/z	voi
admit	152	aðmitiɾ	ɐdmitiɾ	ð/d	cont
admit	152	aðmitiɾ	kõfisaɾ	t/s	cont, strid
accuse	31	denunθjaɾ	ɐkuzaɾ	θ/z	voi, strid

(The full per-slot table is `data/results/provenance_Spanish_Portuguese.csv`, 2026 non-identity slots.) This is the difference between an opaque aggregate distance and an auditable one: the 0.153 voicing component of the Spanish–Portuguese ledger (§5.1) is, concretely, the sum of correspondences like *s/z* in *accuse* — and the counts add back up to the reported d.