The catalogue of consonantal codes

A comparative method without genealogy, cognacy or reconstruction

Alejandro Toledo Martínez · ORCID 0009-0000-1277-9697 · 2026

Abstract

The comparative method has been applied with enormous unevenness: in the corpus this series draws on, kinship judgements exist for one macrosystem out of 225. Any study leaning on the historical layer of the lexicon will measure, without meaning to, the density of the comparative literature.

We describe a method that gives up reconstruction — neither improving nor replacing it: doing without it — and builds a catalogue of consonantal codes from three things and only three: a form's consonant skeleton, the concept it expresses, and a threshold of three distinct languages. The procedure does not know that families have branches; genealogy enters at the end and only to read what came out.

We give the full protocol, the nulls with what each breaks and preserves, and the threshold rules, including the three corrections that five applications forced: the null is computed exactly rather than estimated — estimating it published an interval measuring our own draw for three papers; beating a shuffling null authorises claiming a correspondence and nothing more, because reading it as migration requires a proto-form the method does not have; and the threshold is fixed bounded by two limits, the achievable ceiling and the false-positive rate, because the $2\times$ used for five papers turned out to be exceeded by chance in 64% of cases for one of the statistics employed.

The criterion is tested in the one place where the answer is known. Against Indo-European gold cognacy, concept plus identical skeleton gives 93.3% precision against a 26.0% base, with 13.6% coverage: high precision and low coverage, which is what a catalogue should be.

We also compare the instrument's cost across the five macrosystems applied, and find that collision — the measure of what the skeleton does not distinguish — is not explained by word length: Pama-Nyungan and Indo-European have the same mean length and differ by eighteen points.

1 · The problem, and the wager

Historical linguistics has a powerful and expensive instrument: the comparative method. It reconstructs proto-forms, establishes correspondences and produces trees. It works, and has done for two centuries.

But it has been applied with enormous unevenness. Indo-European has dense reconstruction and expert-reviewed cognacy; most other families have neither. In the corpus this work draws on, kinship judgements exist for one macrosystem out of 225.

From which something follows that is worth saying up front: any comparative study leaning on the historical layer of the lexicon will measure, without meaning to, the density of the comparative literature rather than the past of the languages.

1.1 · The wager

This work asks whether something useful can be done by giving up reconstruction. Not improving it, not replacing it: doing without it and seeing what remains.

What remains is a catalogue of consonantal codes: a more modest and far cheaper object that does part of genealogy's work without making any of its claims. It is built from three things and only three — a form's consonant skeleton, the concept it expresses, and a threshold of distinct languages — and it does not know that families have branches.

1.2 · The order matters: catalogue first, genealogy last

This is what distinguishes the method from an ordinary comparison, and it is easy to lose sight of.

The catalogue is built without genealogy. No branch classification enters, no proto-form, no cognate list, no hypothesis about what descends from what.

Genealogy enters afterwards, and only to read what came out. When codes are crossed with branch labels, it is done on an already-closed catalogue, and the question is not "what codes does the classification predict?" but the inverse: "does the catalogue, on its own, recover something the classification recognises?" The proto-forms appearing in the tables are there so the reader can check what was recovered — the method did not see them, did not use them and does not produce them.

That order is not a stylistic preference: it is what makes the instrument work where there is no reconstruction, which is what it was designed for. A method that needed the classification to build its catalogue could not be applied precisely where it is needed.

1.3 · What is not attempted

The object is easy to over-read, so it is worth bounding precisely.

No proto-forms are reconstructed. No ancestral form, not even implicitly.

No kinship is claimed. That two words share a code does not say they descend from a common one.

Inheritance is not distinguished from borrowing. Without proto-forms this is impossible, and there is no point pretending otherwise. A group of words sharing a code may be common inheritance, old borrowing, recent borrowing, convergence or coincidence. The catalogue does not choose.

No meaning is attributed to a code. One writes `n·m|name`, never `n·m = name`. The code does not mean "name": it is the shape "name" repeatedly takes.

1.4 · What can be claimed, then

One thing only, and that is why the method is cheap: that certain form–meaning correspondences recur across languages more than they would by chance, with chance measured explicitly (§4).

Everything else — what that recurrence means, whether it is inheritance or contact, what produced it — falls outside. It is less than the comparative method gives, and it can be had where the comparative method has not been applied.

1.5 · Where this sits

Reducing a word to its consonants is not new: it is the *consonant class matching* of Dolgopolsky (1964) and of Turchin, Peiros and Gell-Mann (2010), evaluated alongside SCA and LexStat in the automatic

comparison literature (List 2012, 2014; List, Greenhill and Gray 2017; Jäger 2018). It should not be confused with ASJP (Brown et al. 2008; Holman et al. 2008), which keeps the vowels and compares by chance-corrected Levenshtein distance.

The variant used here differs in two ways:

It does not truncate. Classical consonant class matching compares the first two consonants. Here the whole sequence is kept, because arity is an object of study and not a fixed parameter (§2.4): that a macrosystem produces many four-consonant codes is a diagnostic about its morphology, and truncation would hide it.

The output is not a distance matrix but a catalogue searchable by concept. The difference is not one of format: a distance matrix answers "how similar are these two languages?" and a catalogue answers "what shape does this concept repeatedly take in this macrosystem?" The second question does not average over the lexicon, and so each code can be inspected, disputed and refuted one at a time.

A difference that is NOT claimed. It would be convenient to say that sound classes are not used here. That is false: the skeleton's canonical normalisation is a class layer, and that is why every application publishes its whole table before any result (§2.1).

2 · The instrument

2.1 · The consonant skeleton

The sequence of a word's consonants, with the vowels removed — not lost: they stay in a separate column — and each consonant written with a canonical representative by qualic value.

The skeleton is not produced by this method: it is a corpus layer, computed with the same rules for all 225 macrosystems, and its definition, its normalisation decisions and its cost are described in the corpus paper. Here it is taken as input.

What does belong to this method is the normalisation table each macrosystem publishes, derived by aligning every form with its skeleton and not written by hand. Publishing it is compulsory, because whatever the normalisation merges stops being visible, and the reader is entitled to know what.

2.2 · What a code is

A skeleton that appears, for the same concept, in three distinct languages or more.

The concept requirement is not decorative. Without it, the skeleton groups by formal resemblance and gathers unrelated words; with it, the grouping has to get two things right at once, and it is that double condition that produces whatever signal the method can have.

The three-language threshold is a declared decision, not a discovery. With two, any coincidence between neighbours produces a code.

2.3 · The instrument's cost: collision

A skeleton does not distinguish what vowels distinguish. The clean measure of that, with no genealogy involved, is: what proportion of forms shares a skeleton, within its own language, with a form of another concept.

It is high, and varies a great deal by macrosystem (§5). In Turkic it is 59.4%, and the example that shows it best is `k·l`, which appears with ten concepts: *kül* "ash", *kel-* "come", *köl* "lake", *kül-* "laugh". Three are told apart by the vowel, which is exactly what the skeleton does not record.

But the fourth is more interesting, and forces a qualification the method must carry in writing: **kül** "ash" and **kül-** "laugh" are already homophones in Old Turkic. No procedure keeping the vowels would separate them either. Part of the collision is not a cost of the instrument but a fact about the language, and confusing the two overstates the method's price.

2.4 · Arity

A code may have one consonant, two, three or more.

A unary sustains no claim on its own: there are few consonants and any of them coincides by chance. A code of four or more is grounds for doubting the analysis itself, because it usually means the morphology has not been separated — and across the five applications that suspicion proved founded more than once.

The figures a paper in this series defends are restricted to binaries and ternaries. It is a declared restriction and it costs: in Pama-Nyungan it leaves out 447 codes of 2,092.

2.5 · Enrichment, and a warning about it

A code appearing in many languages may be notable or trivial depending on what it is made of: one built from the family's most frequent consonants will turn up everywhere while saying nothing. Enrichment corrects for that:

reach	number of distinct languages the code appears in
coverage	reach ÷ languages attesting that concept
base frequency	times that skeleton occurs in the macrosystem ÷ total skeletons
enrichment	coverage ÷ base frequency

It is a ratio of two proportions and has no units. The × accompanying it is not the same × as a ratio over a null, which compares two other things — and confusing the two is easy to do when reading a table.

The warning: enrichments are not comparable across macrosystems. Across the five applications the median value runs from 15.0 in Austronesian to 190.5 in Turkic, an order of magnitude, and that difference does not say Turkic has thirteen times better codes: it says both quantities entering the ratio depend on the size and shape of the macrosystem. Within a macrosystem enrichment ranks well; across macrosystems it should not be used without normalising first, and this work does not do so.

3 · The protocol

Five applications taught that the order of steps is not indifferent. This is the protocol as it stands, with the corrections each family forced.

3.1 · Phase 0: hygiene, before looking at anything

Before building a catalogue the raw population is audited: doculect duplicates, misclassified languages, citation forms, impossible skeletons.

It goes first because afterwards it cannot be seen. A duplicated doculect inflates the coverage of every code it appears in, and once the catalogue is built the inflation is indistinguishable from signal. In Uralic, phase 0 found a Kam-Sui language classified as Finnic, flagged by three independent checks: it is 6,446 km from the family's centre, it marks tone in digits when no Uralic language does, and its forms are θaːm^{453} "three" and ðɐm^{214} "water". That is 1,008 forms that would have entered the catalogue as Uralic material.

Phase 0 runs on the RAW population and publishes its figure. What it removes, it declares.

3.2 · One macrosystem per measurement

A rule adopted after measuring it: macrosystems are not mixed within a single measurement.

This is not purism. In this material the between-language variance dominates any effect one looks for, and pooling produces artefacts shaped like findings. In a single day's work three appeared: a gradient running from 63% to 86% dissolved entirely when looked at within each language — Simpson's paradox — and the same happened with arity and with nuclei.

From the same rule it follows that the value of a class is relative to the macrosystem that produced it. Transplanting a class discovered in Austronesian to read Turkic produces figures shaped like data and empty of content.

3.3 · Stripping the citation form

Sources do not cite alike. In Turkic, `ids` gives verbs almost always in the infinitive — 85% carry `–mAK` — while other sources give the bare stem. The same verb then comes out with two different skeletons, `j·ʃ·m·q` against `j·ʃ`, and forms no code: two consonantal positions belonging not to the word but to the dictionary.

Removing them is legitimate and it is dangerous, so it comes with three conditions:

1. Declared per family. No automatic stripping. Cutting endings that "look like suffixes" is the quick route to manufacturing exactly the codes one wants to see.
2. Double condition. The orthography and the skeleton's tail must both match. If only one matches, nothing is touched — so a root that happens to end in `–mak` survives. It is exclusion by construction, not by judgement.
3. Measured and published. How many forms were touched and how many codes were gained or lost. A stripping rule that does not pay is withdrawn.

3.4 · Declare before measuring

Every study in this series writes down, before running anything: what is measured, on what population, with what null, with what threshold, and what would count as falsification. Then it is run once and published whatever comes out.

This is not a formality. In this series' tone studies, three measurements fell below their bar and all three were published; and on two occasions the protocol prevented a result that would have come from moving the threshold after seeing the data.

The order is: dispersion → null alone → ceiling → threshold → observed, once. The dispersion step — is there variance to measure? — was added after skipping it twice and discovering afterwards that one of the two variables did not vary, so the measurement had no possible answer.

3.5 · The single door

Every script reads a family's data through one function, which always applies the declared decisions and counts what it applied.

It exists because it did not: the catalogue applied all the decisions and the appendix only some, so the catalogue excluded a language and the appendix displayed it, on two consecutive pages of the same paper.

And that query carries an **`ORDER BY`** ending in a unique column. Without it PostgreSQL promises no order and gave none: five of six scripts were non-deterministic, and the branch-pair substitution table published `b p` at 0.5× or `b n` at 1.9× depending on the run.

4 · The nulls

A catalogue without a null says nothing: any procedure that groups produces groups. The question is always how many chance would have produced.

4.1 · What a null must break, and what it must preserve

A null must break exactly the link being claimed, and preserve intact everything else that is real.

This series' null shuffles concepts within each language. It breaks the form–meaning relation — which is what is claimed — and preserves each language's inventory, its number of forms and its word lengths, which are facts.

The nulls that were discarded teach more than the one used:

- Placing an element at a random position. Discarded: it destroys a real constraint unrelated to the hypothesis and produces a null that is trivial to beat — a large ratio for a boring reason.
- Comparing, within one language, marked forms against unmarked. Discarded after checking: in Vietnamese the 42% "without tone" was one source at 0% against another at 58%. It would compare sources, not language states.

4.2 · The null is computed exactly, not estimated

For three papers the null was estimated by shuffling eighty thousand times, and its interval was published as if it measured the uncertainty of the data. It measured that of the draw. This was discovered by removing six forms and seeing the null move by 14%: a sensitivity impossible if the interval described the population.

The closed form is

```
(Σ_c n_c² − Σ_cl n_cl²) / (N² − Σ_l n_l²)
```

and on substituting it the published figures changed: 225× → 218×, 77× → 78×, 39× → 37×.

Rule: if the null has a closed form, the closed form is used; and it is validated against brute-force shuffling, because an unvalidated closed form is exactly what published a wrong null for three papers.

4.3 · What beating a null authorises

This is the easiest thing to overstate, and it is worth stating narrowly.

> Beating a shuffling null demonstrates a CORRESPONDENCE. Nothing more.

Reading that correspondence as *migration*, as *inheritance* or as *change* requires claiming there existed an earlier form from which both descend. That is a proto-form, and this method has none. The null does not manufacture one.

The nulls do not fail by construction; what fails is what one allows oneself to say on beating them. It is the characteristic scope error of this work: claiming more scale than has been earned.

4.4 · The threshold is fixed beforehand, and against two limits

A threshold chosen after seeing the result is not a threshold. But fixing it beforehand is not enough either: it must be fixed against the two limits that bound it, and neither is obvious.

From above, the achievable ceiling. Every measure has a maximum — for a mutual information, `min(H(a), H(b))` — and a threshold above it declares the measurement impossible in advance. It must be computed and published.

From below, the false positives. A null has a mean and also a dispersion, and the dispersion is not given by the closed form. Checked: with a threshold of $2\times$ the null's mean, 64% of pairs exceeded it by chance in more than one shuffle in twenty. The $2\times$ this series used for five papers does not hold for a mutual information.

The rule, formulated after correcting it once: the minimum ceiling is a cap, not a target; within it, one picks the lowest value whose false-positive rate falls below 0.5%. The first formulation — "the highest still achievable" — worked by coincidence in a macrosystem where the two limits crossed at a single value, and in the next one it gave $15\times$, which no real effect reaches.

4.5 · Pseudoreplication

A set of language pairs is not a set of independent observations if the same languages appear in many pairs. In one study of this series, 243 pairs came from eleven recombined languages.

It is the error that invalidated an earlier version of a paper in this same series. The summary is done by language, not by pair, and the pair count is reported as what it is, not as a sample size.

5 · What the instrument costs, measured across five macrosystems

The method has been applied to five. The table comes from the published catalogues and from the single door, with the same decisions each paper applied:

macrosystem	languages	forms	C/word	collision	codes	bin+ter	≥ 4	median enrich.
Austronesian	582	216,728	2.78	46.2%	10,990	7,881	1,600	15.0
Turkic	32	54,875	2.85	59.4%	1,515	1,150	180	190.5
Uralic	27	80,032	3.28	59.1%	1,474	1,175	133	154.9
Indo-European	207	319,155	3.35	52.5%	10,097	7,445	1,947	138.1
Pama-Nyungan	154	45,829	3.35	34.0%	2,092	1,563	447	40.3

5.1 · Word length does not explain collision

It would be the expected thing — shorter words, more collision — and it does not hold. Pama-Nyungan and Indo-European have exactly the same mean length, 3.35 consonants, and their collision differs by eighteen points: 34.0% against 52.5%.

So the instrument's price depends not only on how many consonants a word has, but on how the available consonants are distributed: a large inventory used evenly produces fewer clashes than a small one with a few very frequent consonants.

Pama-Nyungan is the macrosystem where the instrument costs least of the five measured, and this comparison independently reproduces the verdict of its own paper, which was written without it.

5.2 · Enrichment is not comparable across macrosystems

Already flagged in §2.5, and here is the figure: the median runs from 15.0 to 190.5. Thirteen times.

It does not say Turkic has thirteen times better codes. It says both proportions in the ratio depend on the size and shape of the macrosystem, and that enrichment ranks well within one and means nothing between two. It is the kind of figure that invites a comparative table which should not be made.

5.3 · High arity signals unseparated morphology

Codes of four consonants or more run from 12.0% in Turkic to 21.4% in Pama-Nyungan, and in every case investigated they turned out to be unseparated morphology. In Turkic, stripping the citation form (§3.3) brought them down from 29% to 12%.

A high proportion of arity $\geq$4 is a diagnostic, not a result: it says go back to phase 0.

5.4 · What applying it five times taught

Each macrosystem taught something different about the instrument, and none of the five lessons could have been anticipated from the design:

- Austronesian — the first — showed that the catalogue exists: that giving up reconstruction leaves something.
- Turkic showed what its figures mean, by having a known history to check against. And it showed the cost of normalisation: merging every sibilant into `s` makes lambdacism invisible, which is one of the correspondences defining the family's primary split.
- Uralic was where the instrument could finally do what it promises, with the highest null ratio.
- Indo-European was done blind, without looking at the reconstruction, so that a comparison could be made afterwards.
- Pama-Nyungan is where it costs least, and suggests the price depends on the inventory and not only on length.

6 · The criterion, tested where the answer is known

Everything above rests on a wager: that shared concept plus identical skeleton identifies related words often enough. It is a wager testable in one place and one only — Indo-European, the sole macrosystem

with expert cognacy judgements.

`iecor-gold` is used (152 languages, 170 basic-vocabulary concepts, controlled judgements) and the question is: of the pairs the criterion proposes, how many are genuinely cognate?

6.1 · The result

base — same-concept pairs that are cognate	389,457 / 1,495,154	26.0%
code — of those, the ones sharing the whole skeleton	52,780 / 56,559	93.3%
coverage — cognate pairs the criterion proposes	56,559 / 389,457	13.6%

The criterion has high precision and low coverage. Nine of every ten codes are a real cognate group; and it finds one of every seven that exist.

That is exactly what a catalogue should be. An instrument that gives up reconstruction cannot aspire to find everything; what it cannot afford is for what it does find to be noise. At 93.3%, it is not.

6.2 · A tension with §2.4, and it is not resolved in our favour

By number of consonants in the shared skeleton:

arity	precision
1	92.8%
2	93.8%
3	92.6%
4+	93.7%

It is flat. A unary code is as accurate as a ternary, and that contradicts §2.4, which says a unary sustains no claim on its own.

The contradiction will not be papered over, but neither is it resolved in favour of unaries, for a reason that lies in the population measured: all these forms have cognacy judgements, that is, they already belong to some set. The measure says how many of the proposed pairs are cognate among forms that are in the cognate inventory, not among forms in general. And the starting base — 26% — is already very high, because the basic vocabulary of a well-attested family is dense in cognates.

The prudent reading: in Indo-European basic vocabulary, a single shared consonant plus the concept predicts cognacy at 93%. It does not authorise using unaries in a macrosystem where cognate density is unknown, which is the case for the other 224. The §2.4 restriction stands, now as a declared decision rather than a measured fact, which is a more honest position than the previous one.

6.3 · And what this test cannot say

`iecor-gold` is 170 basic vocabulary concepts in a macrosystem with two centuries of work behind it. The criterion could behave worse in non-basic vocabulary, and there is no way to know: no gold cognacy exists outside it.

That is the programme's asymmetry as the corpus is integrated today, and it is worth putting it that way: outside Indo-European there are published cognate sets — the *Austronesian Comparative Dictionary*, Uralic databases, Bantu and Sino-Tibetan comparative work — that are not integrated here. Integrating them would widen the test bench, and that is pending work rather than a limit of the field.

What is asymmetric at root is that the instrument was designed for the macrosystems without a comparative tradition, and those are precisely the ones that cannot serve as a test bench. The answer is not to pretend the calibration transfers, but to declare the scope of what was checked and apply the instrument knowing what is unknown.

7 · Falsifiers and limits

7.1 · What would falsify this method

A method that cannot be falsified is not a method. These are the results that would sink it, and none has occurred:

That the catalogue did not beat its null. If the real grouping covered roughly what shuffling covers, there would be no form–meaning correspondence to measure. Across the five applications the real grouping exceeds the null by a factor of three to seven.

That the criterion did not discriminate where it can be checked. If concept + skeleton did not raise the cognate rate above its base in Indo-European, everything built on it would be noise. It gives 93.3% against 26.0% (§6).

That codes appeared equally within and across branches. This was checked in one study of the series: the ratio over the null is 1.29× within branch and 0.99× across — exactly the null. That the branch distinction appears as an output of a design that does not use it as an input is what authorises applying the instrument where there are no branches.

That high arity were not explained by morphology. It was: stripping the citation form brought codes of four or more down from 29% to 12% in Turkic.

7.2 · The limits that will not be resolved

It does not distinguish inheritance from borrowing, and will not. This is inherent in giving up protoforms.

It is not comparable across macrosystems beyond collision. Enrichment is not (§2.5, §5.2), and classes discovered in one do not hold in another.

Coverage is low by design. It finds one in seven of the cognate groups that exist (§6.1). A catalogue is not an inventory.

It depends on a layer it does not control. The skeleton is produced by the corpus; its normalisation decisions — what is merged and what is not — determine which correspondences can be seen. In Turkic, merging the sibilants makes lambdacism invisible, and lambdacism is the correspondence defining the family's primary split. That cost is declared in each application and cannot be eliminated, only measured.

And it cannot be calibrated where it is needed. §6.3.

7.3 · What this method is not

It is not an alternative to the comparative method nor a cheap version of it. It produces less, claims less and costs less, and its only advantage is that it can be applied where the other has not been. Where there is reconstruction, the comparative method is better at everything.

8 · How to apply it to a new macrosystem

The full protocol, in the order it must be done. Each step is where it is because skipping it cost something in one of the five applications.

1 · Phase 0, on the raw population. Doculect duplicates by glottocode, misclassified languages, citation forms, impossible skeletons. What is removed is published, with the reason. *It goes first because afterwards the inflation is indistinguishable from signal.*

2 · Declare the family's decisions. Which doculects are excluded and by what mechanical rule; whether there is citation-form stripping, with its double condition. *Nothing automatic: a rule nobody wrote is not applied.*

3 · Publish the normalisation table, derived and not written by hand, with the cost it imposes: which correspondences this family can no longer see because of what the normalisation merges.

4 · Build the catalogue without genealogy. Skeleton, concept, three-language threshold. Nothing else.

5 · Check the arity. If codes of four or more exceed 15%, go back to step 2: that is unseparated morphology, not a finding.

6 · The null, exact and validated. Shuffle concepts within each language. If it has a closed form, use the closed form — checked against brute-force shuffling.

7 · The threshold, before looking at the observed value, bounded above by the achievable ceiling and below by the false-positive rate measured on the null itself.

8 · Run once. And publish the result whatever it is.

9 · Only now, genealogy, and only to read what came out: does the catalogue recover something the classification recognises?

10 · Write down what this family taught about the instrument, which tends to be the most useful part of the exercise and cannot be anticipated. The five taught five different things (§5.4).

Where the material is, and how to request it

Everything — the pipeline, the catalogues, the five papers and each study's pre-declarations — is in the programme repository. The data comes from the Integrative Corpus, described separately.

Both repositories are private today and access is granted on request, by writing to alex@pixelspace.com with who you are and what for. Whoever receives it has full read access and can fork them to their own account.

The reason is in the corpus paper: until the source versions are pinned, a public clone would promise a reproducibility that document declares does not yet exist. It is a state, not a policy, and it lifts when those versions are pinned.

The licence does not depend on this: CC-BY-SA-4.0 on the content, and ShareAlike is inherited.

9 · References

This section was missing. The text cited by author and year — Dolgopolsky, List, Jäger, ASJP — without any of those citations having an entry, which is the very defect an external review flagged as a blocker in the corpus paper. The entries identify the work, not the version used; for the corpus the data come from, the exact fingerprint of the files is in its source manifest.

9.1 · The other two works in this series

- Toledo Martínez, A. (2026). *Integrative Corpus: a multi-layer lexical database of 225 macrosystems, with row-level provenance and its verification apparatus.* — The resource everything here runs on. This work's catalogue builds no data: it reads it from there, and the division between the two is in §1.5.
- Toledo Martínez, A. (2026). *Meulemans: a read-only query layer over a multi-layer lexical corpus of 225 macrosystems.* — The other consumer of the same corpus, and the useful contrast: it displays the lineage this method forbids itself to look at.

9.2 · The consonant skeleton and automatic comparison

- Dolgopolsky, A. B. (1964). Gipoteza drevnejšego rodstva jazykovyx semej Severnoj Evrazii s verojatnostnoj točki zrenija. *Voprosy Jazykoznanija* 2, 53–63. — The *consonant class matching* the skeleton of §2.1 descends from.
- Turchin, P., Peiros, I. & Gell-Mann, M. (2010). Analyzing genetic connections between languages by matching consonant classes. *Journal of Language Relationship* 3, 117–126. — The version that truncates to the first two consonants, which §1.5 deliberately departs from.
- List, J.-M. (2012). LexStat: Automatic detection of cognates in multilingual wordlists. *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, 117–125.
- List, J.-M. (2014). *Sequence Comparison in Historical Linguistics*. Düsseldorf: Düsseldorf University Press.
- List, J.-M., Greenhill, S. J. & Gray, R. D. (2017). The potential of automatic word comparison for historical linguistics. *PLoS ONE* 12(1), e0170046.
- Jäger, G. (2018). Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data* 5, 180189.

9.3 · ASJP, with which it should not be confused

- Brown, C. H., Holman, E. W., Wichmann, S. & Velupillai, V. (2008). Automated classification of the world's languages: A description of the method and preliminary results. *STUF – Language Typology and Universals* 61(4), 285–308.
- Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A. & Bakker, D. (2008). Explorations in automated language classification. *Folia Linguistica* 42(3–4), 331–354.
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10(8), 707–710. — The distance ASJP corrects for chance, and which §1.5 distinguishes from what is done here.

9.4 · The nulls

- Vinh, N. X., Epps, J. & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research* 11, 2837–2854. — The finite-sample bias of mutual information and its closed-form correction, which is what `cod/metricas.py` implements and what §4.2 generalises into a rule.

9.5 · The reference pattern of §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. — IE-CoR, where `iecor-gold` comes from: the cognacy judgements the instrument is calibrated against.