Consonantal codes in 582 Austronesian languages

A catalogue of form–meaning correspondences built without genealogy or reconstruction

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

Abstract

We present a catalogue of 10,994 consonantal codes grouping 84,884 lexical forms from 582 Austronesian languages. A code is a consonant skeleton shared by words for the same concept across distinct languages. Its construction uses no etymology, no proto-forms and no cognacy judgements: only the transcribed form, the concept it expresses, and the requirement that the coincidence occur in at least three languages.

The procedure groups 39.1% of the available forms, with a precision 41.6 times above chance. 85.4% of the resulting codes have three consonants or fewer. Calibrated against expert Indo-European cognacy under equivalent conditions, the method shows precision 0.96 and recall 0.32: it finds a reliable minority of the relations that exist.

The codes of widest reach reproduce the pan-Austronesian vocabulary that comparative linguistics had already established — `l·m` for *five* in 252 languages, `m·t` for *eye* in 246, `t·l·n` for *ear* in 54 — without having reconstructed anything. We argue that this catalogue has three uses: generating testable genealogical hypotheses in families where reconstruction is sparse, producing an inventory of substitution pairs without aligning anything, and grounding a method of lexical learning based on predictability across related languages, which in Austronesian reaches a median of 50.6% of the vocabulary per language — a figure which, as declared in §6.4, depends on the corpus's sampling density and cannot be read as a measure of divergence.

1. What is being sought, and why

This series asks whether something useful can be done by giving up reconstruction, and the method paper describes the resulting instrument: a catalogue of consonantal codes, more modest and cheaper than the comparative method, which does part of its work without making its claims.

1.1 What this work does not attempt

This is worth stating first, because the object invites over-reading.

We do not reconstruct proto-forms. No ancestral form is proposed, not even implicitly.

We do not claim relatedness. That two words share a code does not say they descend from a common one.

We do not distinguish inheritance from borrowing. Without proto-forms this is impossible, and we will not pretend otherwise. A group of words sharing a code may be common inheritance, ancient borrowing, recent borrowing, convergence or coincidence. The catalogue does not choose.

We do not assign meaning to a code. `l·m` does not mean *five*. It is what the 252 languages that say *five* share, under those conditions of information. We write `l·m|five` and never `l·m = five`.

1.2 Where this fits, and what it is not

The procedure is not new. Reducing a word to its consonants and ignoring vowels is the *consonant class matching* introduced by Dolgopolsky (1964) and formalised by Turchin, Peiros and Gell-Mann (2010). It should not be confused with ASJP, which is a different thing: ASJP transcribes into a 41-symbol alphabet that preserves the vowels and compares with normalised, chance-corrected Levenshtein distance (Brown, Holman, Wichmann and Velupillai 2008; Holman et al. 2008, 2011). The benchmarking literature treats ”Turchin” and ”ASJP” as distinct methods, and compares both against SCA and LexStat (List 2012, 2014; List, Greenhill and Gray 2017, which uses Austronesian as a test set and B-cubed as its metric; Jäger 2018; Rama et al. 2018). What is done here is a variant of the first, with three declared differences:

	Prior work	Here
Length considered	first two consonants of the root	full skeleton, unbounded arity
Output	distance matrix between languages	catalogue of codes by concept
Use	phylogenetic classification	consultable inventory, classifying nothing

The first difference is the substantive one, and it is what organises §2.3 and §2.4: Dolgopolsky and Turchin et al. truncate at two positions, whereas here arity is the object of study rather than a fixed parameter. The second is the object of the work: no tree is sought, a list is.

A difference we do not claim. It would be convenient to say that no sound classes are used here and that this is why the grouping is more precise. It is not true: the canonical normalisation of §2.1 is a layer of classes — it collapses `ɣ` with `x`, and affricates and postalveolars with `s` — which is why it is published in full before any result. The difference from SCA or from Dolgopolsky's classes is not having them or not, but that ours are fewer and declared.

And this is not multilateral comparison. The surface profile of this work — resemblance of form, same gloss, three languages, no reconstruction — matches that of the multilateral comparison the discipline rejected, and it is worth stating where it parts company. Here the criterion is countable and falsifiable: every grouping is compared against a permutation breaking the form–meaning relation (§3.4), every code carries its enrichment over the base frequency (§3.3), and no new genealogical grouping is proposed — neither between languages nor between families. Multilateral comparison failed precisely for want of those controls.

2. What a code is

2.1 The consonant skeleton

From each word we extract the sequence of its consonants, preserving order and discarding vowels. Malagasy *mainty* ’black’ has the segments `m`, `ai̯`, `ⁿt`, `i`; its skeleton is `m·n·t`.

The canonical normalisation groups each consonant under a representative, and it is worth having to hand because it decides which codes coincide. It derives from the mapping in `ingest/recompute_skeleton.py`:

Written	as	Written	as
h ħ ɦ ʔ ʕ ʜ ʡ ʢ	h	n ŋ ɲ ɳ ɴ	n
l ʎ ɫ ǁ ɬ ɫ ɭ ɮ ʟ	l	r ɹ ɺ ɻ ɽ ɾ ʀ ʁ	r
x ç ɣ ɧ ʝ χ	x	ɡ ɟ g	g
c ɟ	c	d ɗ	d · m ɱ → m
ʃ ʒ ʂ ʐ ɕ ʑ and every affricate (ʦ ʣ ʧ ʤ t͡s t͡ʃ...)	s		

Four consequences the reader will meet in the data, explained here. The glottal stop **ʔ** is written **h** (so *ʔepat* gives h · p · t) and the velar nasal **ŋ** is written **n** (so *taliŋa* gives t · l · n). Prenasalised consonants occupy two positions (^{n}t → n · t), which is why *mainty* comes out ternary. And — the widest-reaching — affricates are read by their destination and written **s**, together with the displaced sibilants: this is why m·s|star gathers *masoa* with *mat*^{s}*oi* and *maʧo*, and m·s|BE DEAD OR DIE gathers *mase* with *-meʧ*.

This table has two costs worth seeing before the results. The first is that the claim ”no sound classes are used here” would be false, which is why §1.2 does not make it. The second is that the pairs this table merges are invisible to §6.3: x ↔ ɣ cannot appear in the substitution inventory because it is already resolved here, and that section's s ↔ t row is measured over an alphabet in which t͡ʃ already counts as s. The substitution matrix measures, in part, this table.

Vowels are not lost: they are stored in a column of their own, alongside the syllabic template. The reason is practical and appears in §7.3: the analysis works on the skeleton, but the morphological segmentation that would reach the root needs the vowels, because Austronesian affixes are syllables (ma–, ka–, –an) and in the skeleton ma– is indistinguishable from a root m.

2.2 A code is a skeleton that recurs across languages

A code is a consonant skeleton appearing for the same concept in at least three distinct languages. Both requirements matter:

- The same concept, because without semantic control the grouping collapses. We verified this: in earlier work on this same corpus, grouping by shared form without control of meaning produced networks that chained homographs until they gathered unrelated words — 96 networks averaging 25.9 languages with code coverage collapsed to 13.1%. The concept is what prevents that chaining.
- At least three languages, because two forms from the same language are not a comparative fact, and two languages yield too many coincidences.

2.3 Unary, binary and ternary codes

The arity of a code is its number of consonantal positions.

A unary code has one consonant: m, t, n. A binary has two: l·m, m·t, t·l. A ternary, three: b·l·n, t·l·n, k·l·t.

Arity determines how much a code can discriminate, and the reason is arithmetic. With an inventory of some thirty consonants there are thirty possible unary codes, nine hundred binary and twenty-seven thousand ternary. Across 216,864 forms this means a unary code falls on thousands of words on average and a ternary on a handful.

The data confirm this bluntly. In Austronesian the unary code n is associated with 148 concepts having three or more languages. A sample, bearing no relation to one another:

I, drink, he/she, this, sand, eat, mother, thou, fish, six, at, that, what, his/her, if, lie down, and, coconut, water, name, where, sleep, foot, chew, stick, tooth, mosquito, house, here, cry, blow, you, big, say...

That is not a code: it is a frequent consonant surfacing by frequency. We measured the precision of each arity — given two forms from distinct languages with the same code, how often do they share a concept? — and the result is unambiguous:

Arity	Distinct values	Precision	Times above chance
1 (unary)	28	2.66%	3.7×
2 (binary)	555	6.81%	14.7×
3 (ternary)	5,045	13.51%	30.4×

The method's power lives in binaries and ternaries. Unary codes are included in the catalogue for completeness, but on their own they support no claim.

One caveat about this table. 18.0% of binary and ternary codes (1,418 of 7,881) repeat a consonant: `s·s|breast`, `m·m|father`, `n·n|mother`, `l·l|tongue`. These are CVCV forms with a repeated consonant, whether by reduplication — a productive grammatical process in Austronesian —, by nursery word, or simply by the shape of the root: `s·s|breast` is PAn *susu, monomorphemic, and `l·l|tongue` gathers Malagasy *lela* with Polynesian *alelo*, neither of which is productive reduplication. The cause varies; what matters here is the consequence, common to all of them: their space of possibilities is far smaller than that of a binary with two distinct consonants, so their probability by chance is higher. They are inside the figures above and have not been excluded; this is declared so that anyone wishing to redo the calculation without them knows how many they are.

2.4 The arity criterion as a control

Before generating the catalogue we declared a criterion: codes should come out unary, binary or ternary, and the appearance of quaternaries was to be taken as grounds for doubting the analysis itself, not as a finding. A four- or five-consonant code in a family of short words suggests the word has not been segmented properly.

The result, with nothing filtered:

Arity	Codes	%
1	1,509	13.7%
2	4,797	43.6%
3	3,084	28.1%
4	1,200	10.9%
5 or more	400	3.6%

Figure 1: Distribution of codes by arity. 85.4% have three consonants or fewer.

85.4% of codes meet the criterion. The remaining 14.6% are in the catalogue, counted and accessible, which is what the criterion asked for: quaternaries are not hidden, they are flagged for review. Their likely source is the absence of morphological segmentation (§7.3).

3. Method

3.1 Data

Integrative Corpus, Austronesian subset: 582 languages, 216,864 forms with transcription and an assigned concept. A note on that 582, since three figures coexist in the corpus and it must be said which one is used: the language table has 586 rows and 584 distinct names — two pairs of rows are the same language recorded twice — and of those 584 names, 582 contribute at least one form with a skeleton. Every figure in this paper counts by language name over those 582, except where stated otherwise. Forms come from standardised wordlists (Lexibank, IDS) under CC-BY licences, and the concept is a Concepticon node rather than a free gloss. Phonetic transcription coverage is 100% and skeleton coverage 93.8%.

3.2 Construction

For each concept, forms are grouped by identical skeleton. A group counts if it gathers three or more languages. No other information intervenes: not family, not branch, not geography, not etymology.

3.3 Enrichment: separating a code from a coincidence

A skeleton may recur because the words are related or because it is made of frequent consonants. To tell these apart, observed is compared against expected:

> enrichment = *observed* / *expected*, where observed = languages of concept c whose form has skeleton e ÷ languages having c; expected = forms in the whole family with skeleton e ÷ forms in the whole family.

An enrichment of 1× means the skeleton occurs in that concept as often as it occurs anywhere. The table below uses a single consonant rather than a full skeleton, because the contrast shows better that way; the calculation is the same with e of one position.

The contrast is sharp:

Set	Element	Observed	Base in family	Enrichment
eight	w	41.2%	2.3%	17.8×
blood	d	25.1%	2.2%	11.4×
eye	m	77.5%	15.5%	5.0×
hundred	m	15.5%	15.5%	1.0×
buy	m	14.1%	15.5%	0.9×

Figure 2: Reach against enrichment. Each point is a binary or ternary code; the vertical axis is logarithmic. A code may gather many languages and say nothing if it is made of frequent consonants.

The same consonant `m` is informative in *eye* and empty in *buy*. In the first it appears five times more than expected; in the second, less than by chance. Enrichment is what turns a constancy into a fact.

But it does not measure antiquity, and this is worth knowing before reading the tables. High enrichment says a rare sequence recurs; nothing more. §6.2 shows that the highest-enrichment codes are systematically the most local — clusters of neighbouring languages — and that reconstruction candidates have moderate enrichment. Ranking by enrichment is not ranking by historical depth.

3.4 Null models

Every result is compared against a permutation that breaks the form–meaning relation while preserving the marginal distributions. Specifically: within each language the concepts are shuffled, so that each language keeps its inventory of skeletons and only the cross-language correspondence is destroyed.

This choice is not cosmetic. A more obvious null — shuffling skeletons within each concept — produces an artefact: 40.7% of (language, concept) pairs have more than one form, and in 32,495 cases those forms repeat a skeleton within the same language. On shuffling, these intra-language duplicates are distributed across distinct languages and manufacture groups that did not exist. Under that null, chance "grouped" more than the real data and the method appeared useless. We record this as a methodological warning: a badly built null does not produce noise, it produces an inverted result.

3.5 Calibration

The method was calibrated against IE-CoR (Heggarty et al. 2023), expert-established Indo-European cognacy: 2,618 sets, 22,554 forms, 170 concepts, 152 languages. Calibration was run under conditions equivalent to this work — using only form, concept and skeleton, with no etymology or proto-forms — and with the B-cubed metric, standard in automatic cognate detection.

Criterion	Precision	Recall	F1
Identical skeleton	0.963	0.317	0.477
null	0.620	0.150	0.241

Why calibration was not done within the family. Austronesian has published expert cognacy — the ABVD of Greenhill, Blust and Gray (2008), with judgements on some 210 concepts across more than a thousand languages — and calibrating against it would have been preferable. It was not done because the ABVD is not in the Integrative Corpus: the Austronesian subset holds not a single cognate set. Ingesting it is pending work, and until then extrapolation from Indo-European is what there is, with the discount that deserves.

When the method groups two words, it is right 96.3% of the time; it finds 31.7% of those that exist. That profile defines its use: it does not reconstruct a lexicon's full partition, it proposes reliable groups.

And part of that 0.963 is not the skeleton's doing. Calibration runs within sets of already-fixed concept, so the semantic restriction is doing work the figure does not separate out. The converse matters more: by requiring an identical Concepticon node, the method cannot find any relation involving semantic shift — the *louse* $\rightarrow$ *flea* type, or the hand/five and eye/face chains — and much etymological work lives there. The recall of 0.32 includes, by construction, zero cognates with meaning change. With one caveat: IE-CoR is Indo-European basic vocabulary, the favourable case. Elsewhere it will be worse.

A second calibration result explains the low recall and is not a defect of the method. Starting from the true cognate sets, only 12.8% have a homogeneous skeleton, and the most frequent skeleton covers on average 52.9% of the set. Real cognates diverge in form: sound change does its work. What a code can recover is the part that still resembles itself.

4. Results

4.1 The catalogue

Codes (≥ 3 languages)	10,994
Concepts with at least one code	543 of 1,823
Forms catalogued	84,884 of 216,726 (39.2%)
Languages	582
Binary and ternary	7,881 (71.7%)
Median enrichment (bin. and tern.)	14×
Languages per code (median)	4

Figure 3: Distribution of reach: how many codes are shared by how many languages. The median is four languages per code.

The grouping covers 39.1% of forms with a precision 41.6 times above chance (15.92% against 0.383% for the null). A median of four languages per code indicates that most are small and local; those of wide reach are a tiny minority.

4.2 The codes of widest reach

Concept	Code	Arity	Languages	Enrichment
five	`l·m`	2	252	120×
eye	`m·t`	2	246	53×
three	`t·l`	2	241	86×
be dead or die	`m·t`	2	204	47×
breast	`s·s`	2	184	95×
bird	`m·n`	2	172	36×
louse	`k·t`	2	135	65×
father	`t·m`	2	105	76×
seven	`p·t`	2	101	92×
stone	`b·t`	2	98	113×
hand	`l·m`	2	93	41×
six	`n·m`	2	77	61×
eight	`w·l`	2	75	152×
moon	**`b·l·n`**	3	71	149×
sleep	**`m·t·r`**	3	54	90×
ear	**`t·l·n`**	3	54	78×
lightning	**`k·l·t`**	3	51	137×
nine	**`s·j·m`**	3	49	531×

Figure 4: The twenty binary and ternary codes of widest reach, by number of languages.

These codes coincide with the pan-Austronesian vocabulary established by comparative linguistics. We note this as an observation, not a validation: nothing was scored against any reconstruction, because the corpus contains no Proto-Austronesian.

5. Worked examples

5.1 **`l·m|five`** — a binary code of maximum reach

252 languages. Coverage of 48.3% of those having the word; enrichment 120×.

Language	Form	Language	Form
Puyuma (Formosa)	*ḷima*	Aklanon (Philippines)	*lima*
Kalinga	*limá*	Tolomako (Vanuatu)	*lima*
Ambrym North	*lim*	Ifugao	*lema*
Naman	*iləm*	Rano	*i-lim*
Atta	*li:ma:*	Nulnesa	*lima*

The code `l·m` survives vowel variation (*lima*, *lema*, *lim*, *iləm*), loss of the final vowel and prefixation (*i-lim*). That is exactly what a consonant skeleton does: ignore what varies fast and keep what varies slowly.

Note that *hand* has the same code `l·m` in 93 languages. These are two distinct sets — each defined by its concept — that happen to share a code. The catalogue does not merge them, and rightly so: merging would be a hypothesis, not a datum. But it leaves them visible side by side, and that adjacency is information. In Austronesian that particular coincidence is well known — 'five' and 'hand' are related, unremarkable in base-five numeral systems — and the catalogue exposes it alongside every other coincidence between concepts, without asserting it. It is a generator of questions, not answers.

5.2 **`t·l·n|ear`** — a ternary across ten thousand kilometres

54 languages, enrichment 78×.

Language	Form	Region
Arta	*taliŋa*	Philippines
Itawis	*taliŋa*	Philippines
Bare'e	*taliŋa*	Sulawesi
Chamorro	*talaŋa*	Micronesia
Bali-Vitu	*taliŋa*	New Britain
Kove	*taliŋa*	New Guinea
Gitua	*taliŋa*	New Guinea
Gedaged	*tiliŋa*	New Guinea
Lou	*teliŋa*	Admiralties
Lote	*talŋa-*	New Britain
Itbayaten	*tilwan*	Batanes
Label	*talŋa*	New Ireland
Nias	*taliŋa*	Sumatra
Pukapuka	*talinga*	Cook Islands

A ternary code with three positions to get right, sustained across the whole Austronesian dispersal. Vowel variation is wide (*a-i*, *a-a*, *i-i*, *e-i*) and the skeleton does not move. In *Lote* and *Label* the medial vowel disappears (*talŋa*) and the code stays intact, because the code does not count vowels.

The last two rows give the measure: between Nias, off Sumatra, and Pukapuka, in the Cook Islands, lie 10,755 km of ocean, and the word is the same. That is the distance the section title refers to; the first twelve rows, all from the Philippines–Melanesia arc, stay within some 4,200 km, and without the last two the title would not hold.

The same reach — 54 languages — belongs to `m·t·r|sleep`, which reaches 12,682 km between Sakalava in Madagascar and Mwerlap in Vanuatu. They are not, therefore, a pair that separates reach from distance: both are long-reach and long-distance, and saying otherwise would manufacture a contrast the data do not give.

And **`m·t·r|sleep`** brings its own warning, which belongs here rather than buried in the limits. The etymon is PMP *ma- + *tuduR: stative prefix plus root. The `m` of that code is not root material, it is morphology — exactly the defect §7.3 and the ”Qualities” section of Appendix A identify as the method’s worst. The code works because the prefix was inherited along with the root, not because the instrument knew how to separate them.

5.3 **s·j·m|nine** — the highest enrichment

49 languages, 531×. The sequence s·j·m is rare in Austronesian, so its recurrence in forty-nine languages for the same concept is extraordinarily improbable by chance. Compare a code of equal reach made of frequent consonants, which would sit at 1–2×.

Forms: *siyam* (Bikol, Binukid, Ibanag, Dupaningan), *siyám* (Alangan, Cuyonon, Hanunóo), *syam* (Aklanon, Bolinao), *si’yʌm* (Balangaw).

This is the case that best illustrates why enrichment is indispensable. A code with 49 languages may be trivial or extraordinary depending on what it is made of, and only comparison with its base frequency tells the difference.

And it also illustrates its limit, which §6.2 measures — to the point of inverting the reading of the case. The languages of this code are not ”all Philippine” by accident of sampling: every one of them is, and the reason is known. Proto-Austronesian for ”nine” is *Siwa; *siam/*siyam is a Philippine-level innovation, not an etymon inherited from the deep node. That is: the highest enrichment in the entire catalogue belongs to an innovation of a single subgroup. Far from weakening the result, it is the best possible illustration of §6.2’s criterion — high enrichment does not mean antiquity, it means a rare sequence recurring among neighbours — which is why the example stays where it is instead of being swapped for a more comfortable one.

A transcription dependency, which this code teaches better than any other. s·j·m exists because the source writes *siyam* with a glide; written *siam*, the same etymon would give s·m, a banal binary of low enrichment. Nine of the sixteen forms in the Appendix A block carry the *y*, three do not. The code is no more fragile than others: it is simply that here the fracture point is in plain sight.

5.4 **m·t** — a code that over-groups, and what is learned from it

The binary code m·t is associated with several concepts: *eye* (246 languages), *be dead or die* (204), *snake* (95), *right* (71), *black* (39). It deserves examination, because not all of these are the same phenomenon.

’Eye’ and ’die’ are two distinct etyma converging on the same skeleton. This is code homophony, and the catalogue reflects it correctly: two separate sets, each with its concept.

’Black’ is another matter. Its forms are *maeto*, *mainty*, *mainte*, *maitem*, *maitum*, *metan*, *maʔitɨm*. Set against those for ’eye’ — *mata*, *matá*, *mata-* — they are plainly not parallel:

Concept	Form	Skeleton	Analysis
eye	*mata*	m·t	full root
black	*maitem*	m·t·m	prefix + root + final nasal
black	*metan*	m·t·n	prefix + root + final nasal
black	*maʔitɨm*	m·h·t·m	with the glottal preserved
black	*mainty*	m·n·t	with prenasalised nasal

In ’black’ the initial m is not part of the root: it is a prefix. And the skeleton cannot tell, because ma– and root m are both m. Only the languages preserving more material — *maitem*, *maʔitɨm* — reveal it.

This is a real limit of the method and should not be glossed over. A code can fuse a prefix with a root. The remedy would be morphological segmentation, which is unresolved (§7.3). In the meantime the catalogue signals it: the sets for 'black' appear distributed across `m·t`, `m·t·m`, `m·t·n`, `m·h·t·m` and `m·n·t`, and that dispersal is visible and is information.

5.5 A case of regular variation: **`t·l`** and **`t·r`** for 'three'

'Three' appears with two codes: `t·l` in 241 languages and `t·r` in 43. These are not two independent facts: they are the `l/r` alternation, one of the most regular correspondences in Austronesian.

Our substitution analysis (§6.3) detects it as the most frequent pair in the whole family: **`l`** ↔ **`r`** with 4,040 cases in the validation half, 3.3 times above expectation. Treating `l` and `r` as equivalent merges the two codes into one covering 284 languages (the two sets are disjoint: no language appears in both).

6. What this can be used for

6.1 Generating testable genealogical hypotheses

The catalogue does not claim relatedness. But it produces candidates ordered by strength, and that is precisely what a reconstruction programme needs in order to know where to begin.

The clearest case is that of codes crossing the family's primary division. That division is worth stating correctly, because whether the procedure means anything depends on it. The established classification (Blust 2009, 2013) does not split Austronesian in two, but into around ten primary subgroups: nine Formosan and Malayo-Polynesian. "Formosan" is a geographical label, not a clade — and that is exactly why counting Formosan branches works: if two languages of Taiwan belong to different primary branches, their common ancestor is already Proto-Austronesian. An etymon present in several Formosan branches is therefore the classic candidate for reconstruction to that node.

Our catalogue identifies them without knowing that rule. The denominators first, since they govern everything else: of those nine Formosan subgroups the corpus holds five — Atayalic, East Formosan, Northwest Formosan, Tsouic and Western Plains Austronesian — so "five branches" means *all there are here*, not all there are. 46 binary or ternary codes cross three branches or more. The table shows the five that cross four or more — the complete list from the top — and, among those crossing three, the ones reaching 40 languages, so the large rows stay comparable:

Concept	Code	Formosan branches	Formosan languages	Malayo-Polynesian	Other	Total
three	`t·r`	5	9	32	2	43
four	**`s·p·t`**	5	12	0	3	15
seven	`p·t`	4	11	86	4	101
five	`r·m`	4	6	28	3	37
I	`j·k`	4	5	4	0	9
five	`l·m`	3	6	243	3	252
three	`t·l`	3	6	232	3	241
father	`t·m`	3	4	99	2	105
stone	`b·t`	3	5	92	1	98
six	`n·m`	3	6	67	4	77
four	`p·t`	3	5	64	3	72

The second row is the best result in the table, and repays reading slowly. `s·p·t|four` crosses all five Formosan branches in the corpus and has zero Malayo-Polynesian languages. The catalogue does not know why; the discipline does: the etymon is PAn *Sepat, that initial `s` is the reflex of *S, and Proto-Malayo-Polynesian loses *S — hence PMP *epat, which in our catalogue are the other two entries for "four", `h·p·t` (36 languages, with the initial glottal) and `p·t` (72). That is: the procedure separated the PAn level from the PMP level in this numeral by itself, using one of the diagnostic phonemes by which that division is defined, and knowing neither the phoneme nor the division. That numerals are also the domain from which Sagart (2004) argues the Formosan subgrouping makes the coincidence more useful still: there is something to check it against.

The table is restricted to binary and ternary codes, and that restriction is a rule, not a tidy-up. §2.3 established that unary codes support no claim on their own; admitting them here would have produced, among others, `n|mother` as a reconstruction candidate — a pair §2.3 uses as its canonical example of noise, since that same unary code occurs with 148 distinct concepts. A criterion declared in one section has to govern the others, and the "Other" column — languages with no branch assigned in the corpus — is there so the reader can check that the figures add up.

These are hypotheses, not conclusions: the catalogue says there is a form–meaning correspondence traversing the deepest division in the family, and that it therefore deserves examination by the comparative method. It does not say it is inheritance. It could be ancient borrowing, and numerals are notoriously borrowable.

But the list is short, ordered by strength and reproducible. For a family without reconstruction, that is the difference between knowing where to start and not knowing.

Note that the list is ordered by number of Formosan branches, not by enrichment. The reason is in §6.2: enrichment rewards locality, and ordering by it would put precisely the worst candidates on top.

And a complementary use, perhaps more valuable: the list of what does not fit. Of the 10,994 codes, 96.8% do not leave a single macroarea. The real macroareas are three — Papunesia (561 languages), Africa (22) and Eurasia (1) — and two languages in the corpus have none assigned; they are excluded from the count, and the figure does not change. The 349 that do cross (3.2%) are the short list of anomalies. Their breakdown says more than the percentage:

Macroareas crossed	Codes
Africa ↔ Papunesia	298
Eurasia ↔ Papunesia	48
all three	2
Africa ↔ Eurasia	1

That is: the anomaly list is 85% Malagasy, the Austronesian language of Madagascar whose nearest relative is in Borneo. The instrument does not know that and finds it anyway. With only three macroareas the bar is low, so the figure is worth the short list it produces rather than the percentage itself; but that short list contains the most famous long-distance dispersal in the family, and without the catalogue there would be no way to find it among 216,864 forms.

6.2 Do the codes match the kinship of the languages they group?

The question is measurable, and the answer forces a correction in how enrichment is read.

Geographic distance between languages is used here, and it must be said plainly that this is a proxy and why it was accepted. The reason we gave — that Glottolog's label does not discriminate, since 567 of the 582 languages fall under Malayo-Polynesian — holds only at the primary level. Glottolog does not give one label per language: it gives a nested classification path fifteen or more levels deep in the Oceanic part, and from it comes the standard measure, the depth of the most recent common node. For Austronesian there is also a published phylogeny with branch lengths (Gray, Drummond and Greenhill 2009). Neither has been used here, and that is a shortcoming of this work, not a property of the family. §6.2 measures geographic proximity; the version with node depth is pending.

And the proxy is especially poor *in this family*. Austronesian is the textbook case of long-distance dispersal: Malagasy sits some 6,500 km from its nearest relative in Borneo (Dahl 1951; Adelaar 1995), the Polynesian outliers are embedded in Melanesia, Chamorro and Palauan are primary Malayo-Polynesian branches in Micronesia. §5.2 itself celebrates a code joining Madagascar to the Pacific. That is: we have chosen as our proxy for kinship the very variable this family falsifies in its most cited cases. What follows still stands — the signal is large and runs in the predicted direction — but read it as proximity, not kinship, and expect the tree-based version to move the figures.

Languages sharing more codes are closer together, and the relation is monotonic:

Codes shared	Pairs	Mean distance	÷ chance
1	64,297	3,373 km	0.94
2	22,447	2,950 km	0.82
3–5	16,505	2,341 km	0.65
6–10	4,853	891 km	0.25
11–20	2,744	553 km	0.15
two random languages	40,000	3,582 km	1.00

On average, the groups each code forms are more compact than expected for their size: between 0.40 and 0.65 times the dispersion of a random group of the same number of languages. The code captures real structure. What it cannot say is of what kind: inheritance and contact predict exactly the same thing here.

That average, however, hides what matters, and the two tables below bring it out: the most enriched codes drop to 0.24 and the reconstruction candidates rise above 1.00, that is, they scatter more than chance. Mean compactness describes the bulk, not the part that counts.

Enrichment measures locality, not antiquity

Crossing the two magnitudes yields a negative relation (r = −0.25). The population changes here and that must be said: the two tables below are computed over the 1,563 binary and ternary codes present in eight languages or more, not the 7,881 of the rest of the paper, because below eight a group's geographic dispersion is too noisy to compare. The table's cuts are the quartiles of *that* population:

Enrichment	Dispersion ÷ chance
2–8×	0.60
8–21×	0.54
21–71×	0.41
71–2208×	0.24

And at the extremes it is unambiguous:

	Code	Languages	Formosan branches	Enrichment	Dispersion
Maximum of this population 0.17	‘g·l·s	ice‘	9	0	2208×
0.06	‘x·m·n	beard‘	9	0	1468×
0.07	‘p·p·s	turmeric‘	10	0	700×
Candidates from §6.1 1.25	‘t·r	three‘	43	5	20×

	Code	Languages	Formosan branches	Enrichment	Dispersion
	‘r·m	five‘	37	4	39×
1.21					
	‘b·t	stone‘	98	3	113×
0.26					

(The maximum across the whole catalogue, without the eight-language filter, is 13,554× and belongs to codes of three or four languages — precisely the range that filter excludes as noisy.)

The highest-enrichment codes have not a single Formosan branch and form local clusters. The extreme case says it better than the table: **`g·l·s|ice`**, at 2208×, is twelve Malagasy varieties saying *gilasy*, *glasy*, *glatsy* — French *glace*, a colonial loan a century old. The second, `x·m·n|beard`, is nine languages of New Georgia sharing the same *ɣumi-na*, possessive included. Note too that ”ice” is neither basic nor inheritable vocabulary in tropical languages: the concept itself was already a warning.

In other words: the most enriched code in the whole Austronesian family is a Gallicism. The reconstruction candidates of §6.1, by contrast, have moderate enrichment and are more dispersed than chance.

This corrects a reading the rest of the paper encourages. Enrichment is presented in §3.3 as what separates a code from a coincidence, and it does; but high enrichment does not indicate historical depth. It indicates that a rare sequence recurs, and a rare sequence recurring among neighbours is exactly what a areal borrowing or a recent innovation produces. Ranking reconstruction candidates by enrichment would rank them nearly backwards.

Two axes, and both are needed

A reading criterion follows that the catalogue did not have:

	Compact	Dispersed
High enrichment	local innovation or borrowing	*rare; worth examining*
Moderate enrichment	uninformative	candidate for old retention

A caveat on the geographic axis. It measures kilometres, not time depth, and the two axes can contradict each other: `s·p·t|four` crosses all five Formosan branches in the corpus — tied with `t·r|three` at the top of §6.1 — and yet comes out compact (0.04×), because twelve of its fifteen languages are in Taiwan, a small island. A code can be old and geographically concentrated if its family is. It is the cleanest counterexample to this section’s reading, and it comes from the same catalogue.

Two further caveats, bearing on every figure in §6.2. The first is Galton’s problem, which §6.4 states and which operates identically here: a group’s compactness depends on where the corpus is dense, and this corpus holds fifteen Malagasy varieties and dozens of Vanuatu lects. The second is the source list: the data come from dozens of separate surveys, and two languages from the same survey share concept inventory, transcription convention and geographic neighbourhood at once. Part of ”sharing codes → being close” is ”coming from the same source”. Neither is controlled here, and the right correction is known: the null should sample from the languages attesting that concept, stratified by source, rather than from all 582; and the genealogical question is answered with a partial Mantel test — code against geography, controlling for tree distance, and the reverse. Both are pending.

6.3 Substitution pairs, discovered without aligning anything

Within each set one can take pairs of languages whose skeletons have the same length and differ in exactly one position. That difference is an already-aligned substitution.

```
ROPE   Taiof  afon  → f·n      Tolai  vinau → v·n     ⇒  f ↔ v
HAIR   Emae   furu  → f·r      Nalik  vur   → v·r     ⇒  f ↔ v
```

Accumulated across the family (206,856 clean substitutions) and validated on a half of the languages that did not take part in the discovery, the most frequent are (counts are from the validation half):

These are not 'correspondences' in the technical sense. A correspondence is regular, directional and environment-conditioned, between two specific languages. This is an unordered substitution matrix aggregated over 582 languages, with neither environment nor direction. Frequency is not regularity. What it produces are candidates: the starting point of a correspondence analysis, not its result.

Pair	Cases	Times above chance	Reading *(our gloss, not a result)*
f ↔ v	1,073	13.9×	labial voicing
v ↔ β	438	13.4×	labial
p ↔ ɸ	133	11.6×	labial
s ↔ z	466	8.0×	sibilant voicing
f ↔ p	897	7.1×	labial
k ↔ x	1,478	5.1×	dorsal spirantisation
f ↔ h	1,029	4.8×	debuccalisation `*p > f > h`
l ↔ r	4,040	3.3×	liquids
s ↔ t	3,321	2.7×	palatalisation `*t > s`

From pairs to classes. Taking the pairs that hold up in the validation half and closing them transitively, the structure that appears depends on the threshold, and the whole sweep is more informative than any one of its rows:

Threshold	Classes obtained
≥2×	`{b c d f g h j k l n p r s t v w x z ɸ β}` — all chained
≥3×	`{b f g h j k p s t v w x z ɸ β}` · **`{l r}`**
≥4×	**`{b f h p v w ɸ β}`** · **`{k s x z}`**
≥6×	`{b f p v ɸ β}` · `{s z}`
≥10×	`{f p v ɸ β}`

The three classes Appendix A writes as `1`, `2` and `3` are not one threshold's output, and it is best not to pretend they are:

	Segments		Appears at
1	b f h p v w ɸ β	labials, plus `h`	≥4×
2	k s x z	dorsals and sibilants	≥4×
3	l r	liquids	≥3×

The liquids detach first (at ≥3×), but at that threshold the other two still form a single undifferentiated block; at ≥4× that block splits cleanly in two and `l ↔ r` — which stands at 3.3×, as the table above shows — no longer appears. The published trio is therefore a reading of the whole sweep, not a single run. Saying so costs a paragraph; declaring "a threshold of 4×" would have produced a table that this very section's `l ↔ r` row contradicts.

None of these pairs was sought. They emerged from counting differences in aligned pairs. And they are among those any description of Austronesian mentions, including the debuccalisation `*p > f > h` that yields Hawaiian.

But it must be said what this procedure cannot see, because that is exactly where the family's classic problem lives. The method takes pairs of skeletons of the same length differing in one position; by construction, it cannot see any zero reflex. And Austronesian's most-cited correspondence sets are precisely the ones including ∅: PAn *R, whose reflexes are *r, l, g, y, h*, ∅ depending on the branch; PAn *q (*ʔ, h, k*, ∅); and the very PAn *S (*s, h*, ∅) that produces the `s·p·t|four` of §6.1. None of the three emergent classes captures *R or *q. Put differently: these pairs are the ones any description mentions in the phonetics chapter; the ones it mentions in the subgrouping chapter are different, and this instrument is blind to them. To which add the warning of §2.1: pairs already merged by the canonical normalisation — `x ↔ ɣ`, `s ↔ ʃ`, `s ↔ t͡ʃ` — cannot appear here either.

With those two blind spots declared, for a family without historical description this procedure would produce a first inventory of candidate pairs in a matter of minutes.

6.4 Language learning

This is the use we least expected and the one with the most favourable figures.

If a learner knows the code for a concept, they know the approximate shape of that word in every language of the group. Learning `t·l·n|ear` gives simultaneous access to *taliŋa*, *talaŋa*, *tiliŋa*, *teliŋa*, *talŋa* and *tilwan* across twelve languages in eight regions. It is not the exact word — the vowels are missing — but it is the frame, and the frame is what is preserved.

We measured what proportion of each language's vocabulary has a code shared with others:

Language	Entries	With shared code	%
Merina (Madagascar)	619	592	95.6%
Masikoro	204	194	95.1%
Penrhyn (Polynesia)	384	364	94.8%
Vezo	613	570	93.0%
Betsileo	817	754	92.3%
…			*median 50.6%*
Tomoip	749	105	14.0%
Atayal (Formosa)	1,144	132	11.5%

The median is 50.6%: half the basic vocabulary of an average Austronesian language is predictable from the code if other languages of the family are known. In the Malagasy group, which is highly cohesive, it exceeds 92%. The median carries the same caveat as the extremes, and the next paragraph states it: it is a median over an unbalanced sample, not over the family.

The low extremes look equally informative — Atayal, Formosan, sits at 11.5% — but that reading does not hold without a control this work lacks. The corpus contains some fifteen Malagasy varieties (Merina, Betsileo, Bara, Antanosy, Antankarana, Masikoro, Vezo, Sakalava…) and no close sister of Atayal. Languages are counted as independent observations and they are not: this is Galton's problem. So 95.6% against 11.5% measures, in the first instance, how many close relatives each language has in this corpus, and only then something about divergence. Calling it a 'measure of distance' would require controlling for sampling density, and that has not been done. The warning applies to all three figures, the median included: if 95.6% cannot be read as cohesion nor 11.5% as divergence, neither can 50.6% be read as a property of Austronesian. What it does support is the pedagogical claim, which is more modest and needs no such control: for a speaker of a particular language with relatives in the corpus, that share of their vocabulary is predictable from the code, and that holds however skewed the sample.

For a teaching method this suggests an order: teach first the codes of widest reach — those covering two hundred languages — and afterwards the realisation rules of each particular language. The learner learns once and applies many times.

7. Limits

7.1 What the catalogue cannot say

Inheritance or borrowing. There is no way to tell without proto-forms, and in Austronesian the borrowing strata are massive and well known: Malay as donor across insular Southeast Asia, Sanskrit and Arabic on top of it, and trade Malay in eastern Philippines and Melanesia. Several catalogue codes fall in that contested zone — `b·l·j|house` (*balay/balai*), `d·g·m|needle`, `s·l|path` — and the catalogue contains them unmarked. The numerals of §6.1 are the typical case: they cross the whole family and are also notoriously borrowable.

Age. The catalogue dates nothing.

Relatedness between families. The code space is small and coincidences across families are expected. Comparing codes between families requires first measuring equivalence, and that has not been done.

Meaning. A code does not mean. It groups.

7.2 Source bias

The forms come from standardised wordlists. They are not basic-vocabulary lists, and saying so would be inaccurate: the corpus covers 1,823 concepts, among them ”ice”, ”turmeric” and ”needle for sewing”. They are dozens of separate field surveys, each with its own concept selection, its own transcriber and its own regional scope; the core of all of them is indeed basic lexicon, but the tail is not. This has two consequences. The bias towards the conservative lexicon is real and says nothing about the rest of the language: the 39.1% coverage figure refers to that subset, not to the Austronesian lexicon. And the provenance of each list is an uncontrolled confound: two languages from the same survey share concepts, transcription convention and geographic neighbourhood at once, which bears directly on §6.2, where it is declared.

7.3 Morphological segmentation, unresolved

This is the most serious limitation. The analysis works on whole words: root and affixes have not been separated. In a family with Austronesian affixation and reduplication this has two consequences:

It underestimates the grouping. Two forms of the same root with different affixes are not recognised. The 39.1% is a floor.

And it fuses prefixes with roots, as seen for ’black’ (§5.4).

We attempted a distributional affix detector and it failed instructively. Applied to the skeleton, in Hawaiian it marked seven of the language’s eight consonants as prefix *and* as suffix: it had not discovered morphology, it had discovered the inventory. The cause is that the separability criterion — ”it can be removed and the remainder exists on its own” — saturates in languages with a small inventory and CV syllable structure.

Repeated on the form with vowels, the same detector found real morphemes: `te–` (Māori definite article), `–ŋa` (nominaliser, 154 cases), `ka–/ke–/nā–` (Hawaiian articles). Segmentation needs the vowels the code discards, and that is the path for the next stage of this work.

And there is a second reason not to use it, visible only against the null. The detector raises raw grouping — from 62.2% to 77.2% — and that figure, alone, looks like a large gain. But the same trimming that brings related words together also brings unrelated ones together, so the null rises with it, from 39.0% to 70.4%:

Step	In a group	Null	Advantage	$\varkappa$
Whole word	62.2%	39.0%	23.2	0.380
+ reduplication	64.8%	—	—	—
+ discovered affixes	77.2%	70.4%	6.8	0.230

The advantage over chance falls from 23.2 points to 6.8, and grouping becomes barely 10% better than shuffling the concepts ($1.60\times \to 1.10\times$). The $\varkappa$ column is the Hubert-Arabie correction, `(obs - null) /(1 - null)`, which normalises by the attainable ceiling and so allows comparison across families: $\Delta\varkappa = -0.151$.

(The first version of this table called "whole word" the step that had already stripped reduplication — 64.8% — and compared from there. The other seven papers in the series start from the step before, so this family's figure was not comparable with theirs. Corrected on 2026-09-09; the reduplication row is published separately, as in the Indo-European paper.)

It is the largest drop among the eight families measured, and it runs the same way in all of them — Turkic −0.009, Afro-Asiatic −0.004, Sino-Tibetan −0.020, Uralic −0.035, Indo-European −0.094, Pama-Nyungan −0.102, Trans-New-Guinea −0.118. The catalogue in this paper does not use automatic segmentation, and this measurement is why.

What the complete series does not confirm is the explanation this paragraph used to give — that "the larger the raw gain a paper reports, the more negative the real one". Sino-Tibetan, with the shortest word of the eight, drops −0.020, less than Indo-European. See §5.1 of the Sino-Tibetan paper.

7.4 Onomatopoeia

Some codes of very high enrichment are onomatopoeic: *chicken* = `k·k·r·k`, of the *kokorako* type. The method groups them rightly — they share form and meaning — but they are not inheritance. The catalogue does not distinguish them.

7.5 One family

All of the above is Austronesian. That the procedure works here does not guarantee it will work in families of other typologies. We measured the form profile of thirteen families and the differences are large: Indo-European has 4.65 consonants per word and Sino-Tibetan 2.22. In families of short words and high tone — Sino-Tibetan, Tai-Kadai — a two-consonant skeleton discriminates very little, and moreover between 33% and 45% of their segments are tonal and fall outside the skeleton. For those families this instrument is insufficient, and saying so is part of the result.

8. Conclusion

Giving up reconstruction, a useful object can still be built. The catalogue of consonantal codes groups 39.1% of Austronesian forms with a precision 41.6 times above chance, reproduces the known pan-Austronesian vocabulary without having sought it, and yields three things that for this family did not exist in consultable form: an ordered list of reconstruction candidates, an inventory of substitution pairs discovered without aligning anything — candidates for correspondence, not correspondences, in the sense §6.3 delimits — and a measure of lexical predictability across languages that can ground a teaching method.

It does so while claiming very little. A code is not a proto-form, implies no relatedness, does not distinguish inheritance from borrowing and means nothing. It is a form–meaning correspondence that recurs across languages, with its reach and its enrichment measured.

We argue that this modesty is the condition of its usefulness. An instrument that claims little can be applied where there is no literature to support it — within the scope §7.5 sets: in families of long words and no tone, which is where the form profile allows it. Not in all of them.

References

Adelaar, K. A. (1995). Borneo as a cross-roads for comparative Austronesian linguistics. In P. Bellwood, J. Fox and D. Tryon (eds.), *The Austronesians*. ANU. — [Barito origin of Malagasy]

Aikhenvald, A. Y. and Dixon, R. M. W. (eds.) (2001). *Areal Diffusion and Genetic Inheritance*. Oxford University Press.

Bagga, A. and Baldwin, B. (1998). *Algorithms for scoring coreference chains*. LREC Workshop on Linguistic Coreference. — [B-cubed metric]

Blust, R. (2013). *The Austronesian Languages*. Revised edition. Asia-Pacific Linguistics, ANU. — [reference classification of the family]

Blust, R. and Trussel, S. *Austronesian Comparative Dictionary*. Online.

Bowern, C. (2013). Relatedness as a factor in language contact. *Journal of Language Contact* 6. — [borrowing in basic vocabulary]

Brown, C. H., Holman, E. W., Wichmann, S. and Velupillai, V. (2008). Automated classification of the world's languages. *Sprachtypologie und Universalienforschung* 61. — [the ASJP method: LDN/LDND]

Dahl, O. C. (1951). *Malgache et Maanjan*. Egede-Instituttet.

Dolgopolsky, A. B. (1964). Gipoteza drevnejšego rodstva jazykovyx semej Severnoj Evrazii s verojatnostnoj točki zrenija. *Voprosy Jazykoznanija* 2. — [consonant classes]

Gray, R. D., Drummond, A. J. and Greenhill, S. J. (2009). Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323. — [Austronesian phylogeny with branch lengths]

Greenhill, S. J., Blust, R. and Gray, R. D. (2008). The Austronesian Basic Vocabulary Database: from bioinformatics to lexomics. *Evolutionary Bioinformatics* 4.

Hammarström, H., Forkel, R., Haspelmath, M. and Bank, S. *Glottolog*. Max Planck Institute for Evolutionary Anthropology.

Heggarty, P. *et al.* (2023). Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. *Science* 381. — [IE-CoR]

Jäger, G. (2018). Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data* 5.

Kessler, B. (2001). *The Significance of Word Lists*. CSLI Publications. — [chance resemblance in word lists]

Key, M. R. and Comrie, B. (eds.). *Intercontinental Dictionary Series*.

List, J.-M. (2012). SCA: phonetic alignment based on sound classes. In *New Directions in Logic, Language and Computation*. Springer.

List, J.-M. (2014). *Sequence Comparison in Historical Linguistics*. Düsseldorf University Press. — [LexStat]

List, J.-M., Greenhill, S. J. and Gray, R. D. (2017). The potential of automatic word comparison for historical linguistics. *PLOS ONE* 12(1). — [benchmarking of Turchin, SCA and LexStat with B-cubed; Austronesian among the test sets]

List, J.-M., Forkel, R., Greenhill, S. J. *et al.* (2022). Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data* 9.

List, J.-M., Rzymski, C., Greenhill, S. J. *et al.* *Concepticon*. Max Planck Institute for Evolutionary Anthropology.

Nichols, J. (1992). *Linguistic Diversity in Space and Time*. University of Chicago Press. — [the geography/genealogy confound]

Rama, T., List, J.-M., Wahle, J. and Jäger, G. (2018). Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? *NAACL-HLT*.

Ranacher, P., Neureiter, N., van Gijn, R. *et al.* (2021). Contact-tracing in cultural evolution: a Bayesian mixture model to detect geographic areas of language contact. *Journal of the Royal Society Interface* 18.

Ross, M. (1988). *Proto Oceanic and the Austronesian Languages of Western Melanesia*. Pacific Linguistics.

Ross, M., Pawley, A. and Osmond, M. *The Lexicon of Proto Oceanic*. Pacific Linguistics.

Sagart, L. (2004). The higher phylogeny of Austronesian and the position of Tai-Kadai. *Oceanic Linguistics*

1. — [Formosan subgrouping argued from the numerals]

Ringe, D. (1992). On calculating the factor of chance in language comparison. *Transactions of the American Philosophical Society* 82.

Toledo Martínez, A. (2026). *Integrative Corpus: a multi-layer lexical database of 225 macrosystems, with row-level provenance and its verification apparatus.* — [the data] · https://doi.org/10.13140/RG.2.2.31244.68486

Turchin, P., Peiros, I. and Gell-Mann, M. (2010). Analyzing genetic connections between languages by matching consonant classes. *Journal of Language Relationship* 3.

Wichmann, S., Holman, E. W. and Brown, C. H. (eds.). *The ASJP Database*. — [database; the method is in Brown et al. 2008 and Holman et al. 2008]

Note: references are given with author, year and venue. Before any submission they should be completed with pages and DOIs, and the cited editions verified.

Data and reproducibility

Catalogue	`output/catalogo-austronesio/catalogo-austronesian.tsv` — 10,994 rows
Per-form detail	`output/catalogo-austronesio/catalogo-austronesian-lenguas.tsv` — 84,884 rows
Construction	`analysis/fases/catalogo_familia.py`
Grouping and nulls (§4.1)	`analysis/fases/agrupamiento_familia.py` → `docs/AUSTRONESIO-PRIMERA-APLICACION.md`
Calibration (§3.5)	`analysis/estudios/fase_c_calibracion.py` → `docs/FASEC-CALIBRACION.md`
Morphological splitting (§7.3)	`analysis/fases/fase2_despiece.py` → `docs/FASE2-DESPIECE.md`
Pairs and classes (§6.3)	`analysis/fases/fase6_clases_emergentes.py` → `docs/FASE6-CLASES-EMERGENTES.md`
Per-set step	`analysis/estudios/peldano_por_conjunto.py` — supporting material, not cited in the body
Codes and kinship (§6.2)	`analysis/fases/codigos_vs_parentesco.py`
Appendix A	`analysis/fases/anexo_codigos.py`

Three warnings about this table, because a reproducibility table that is not faithful is worse than none. (i) A good share of the body's figures live not in a script's output but in the run reports under `docs/`, which is why those are listed alongside. (ii) The counts from `agrupamiento_familia.py` and those of the catalogue do not use the same criterion: the former groups from two languages up and counts by `lect_id`, the catalogue requires three and counts by name; hence the 39.1% coverage and the 15.92% precision of §4.1 come from different populations even though they are cited in one sentence. (iii) The Formosan-branch figures of §6.1 and the macroarea figures are recomputed directly from the two published TSVs, not from `codigos_vs_parentesco.py`, which does not calculate them.

Data sources: Lexibank (List, Forkel, Greenhill et al. 2022, CC-BY-4.0), IDS (Key & Comrie, CC-BY-4.0), Concepticon and Glottolog (List et al.; Hammarström et al., CC-BY-4.0), IE-CoR (Heggarty et al. 2023, CC-BY-4.0). All redistributable with attribution.

Appendix A · One hundred commented codes

The catalogue holds 10,994 codes; the full data are in `output/catalogo-austronesio/`. What follows is a readable sample: the hundred binary and ternary codes of widest reach, each with sixteen languages and their words.

Three things about the presentation. The sixteen languages are sampled across the alphabetical list, not taken from its head: sixteen languages beginning with A would hide the geographic spread, which is what the example has to show. Language names are shortened where long; the forms are never altered. And the thematic grouping is declared by hand as a reading order — it does not come from the data and is not a semantic classification.

Emergent codes are written `1`, `2` and `3` for the three classes of §6.3 — `1` = b f h p v w ɸ β, `2` = k s x z, `3` = l r. The commentary preceding each section points at what can be seen in the block. It is observation, not analysis: the reader is expected to look at the words.

Reverse index: codes recurring across concepts

The thematic order that follows groups by meaning, and in doing so disperses precisely the opposite: one skeleton falling on unrelated concepts, which is the phenomenon of §2.3 and the reason short codes support no claims. This table brings it back into view before the blocks begin.

Read it with the enrichment alongside: when a code appears in five concepts and in all of them with low figures, it is not a code, it is a frequent consonant.

Code	Concepts it appears with, and their enrichment
`m·t`	eye (246, 53×) · be dead or die (204, 48×) · snake (95, 23×) · right (71, 16×) · die (43, 56×) · black (39, 8×)
`t·n`	cry (109, 27×) · earth (92, 22×) · year (59, 16×) · mother (56, 13×) · intestines (47, 13×) · burn (40, 10×)
`m·n`	bird (172, 36×) · drink (81, 17×) · laugh (58, 13×) · tongue (38, 8×) · male person (37, 8×)
`r·n`	hear (108, 39×) · day (68, 26×) · leaf (40, 14×) · rain (precipitation) (36, 14×)
`n·m`	mosquito (104, 44×) · six (77, 61×) · drink (69, 28×) · house (62, 26×)
`l·m`	five (252, 120×) · hand (93, 41×) · come (35, 16×)
`t·l`	three (241, 86×) · rope (70, 26×) · egg (47, 18×)
`n·n`	mother (76, 19×) · eat (36, 9×) · yellow (36, 10×)
`m·m`	father (58, 24×) · chew (51, 35×) · piss (48, 159×)

Numerals

Numerals are the method's best case, and a good place to start.

The same etymon appears split across several codes, and the differences between them are regular correspondences. *Four* comes out as `p·t` (72 languages), `h·p·t` (36) and `v·t` (35): the same word with the initial glottal preserved — which the skeleton writes `h`, §2.1 — and with *p* > *v*. There is a fourth code for *four* that does not enter this appendix because it has only 15 languages, and it is nonetheless the most interesting of the four: **`s·p·t`**, which crosses all five Formosan branches in the corpus and no Malayo-Polynesian one (§6.1). The three above are the PMP level; the missing one is the PAn level. That a reach criterion should leave out the best candidate is, in itself, this block's lesson. *Five* as `l·m` (252) and `r·m` (37); *three* as `t·l` (241) and `t·r` (43) — the `l/r` alternation of §6.3, here visible word by word. *Six* as `n·m` (77) and `h·n·m` (40).

The catalogue does not merge them, deliberately: merging would be proposing a correspondence, which is the analyst's work and not the inventory's. What it does is place them side by side, with their figures, so the correspondence can be seen.

Note the enrichment too: `s·j·m|NINE` reaches 531× with only 49 languages, while `p·t|four` sits at 34× across 72 of them. More languages is not more signal.

`l·m|five` · binary · 252 languages · enrichment 120× · emergent `3·m`

Agta limʌ	As lim	Bikol li'ma	Dupaningan Agta lima
Ibanag lima	Kairiru lim	Koronadal Blaan limɨ	Lonwolwol lim
Manobo 'lima	Merei lima	Nasarian i-ləme	Piamatsina lima
S: Raga: Avavanva lima	Simbo lima	Tagbanwa lima	Uripiv: Durpif e-lim

`t·l|three` · binary · 241 languages · enrichment 86× · emergent `t·3`

'Oyaoya tolu	Antambohoaka telo	Banggi taolu	Bolaang Mongondow tolu
Futuna tolu	Itawis tallu	Kavalan F69 tulu	Lelepa tolu
Mamanwa tolo	Merina_Malagasy telo	Navut tolu	Nume i-tol
Saaroa F69 to:lo	South Efate itol	Tiale tolu	Vaeakau-Taumako tolu

`p·t|SEVEN` · binary · 101 languages · enrichment 92× · emergent `1·t`

Agta pitu	Arta pitu	Bantik pitu	Binukid 'pitu
Cebuano pito	Hiligaynon pito	Ilokano pito	Itawis pitu
Kalinga pitú	Kapampangan pitú	Manobo pitu	Masbatenyo pito
Rano: Chinambong ɛ-pʉt	Sangil 'pitu	Surigaonon pito	Tonsea pitu

`n·m|SIX` · binary · 77 languages · enrichment 61×

Agutyanen enem	Ayta Abellen anɨm	Basay TsYM91 anəm	Bolaang Mongondow onom
Cebuano onom	Hiligaynon anem	Ibanag annam	Iraya anúm
Kalinga onom	Kinaray-a anəm	Manide anum	Moken nam
Romblon unóm	Sarangani Blaan nam	Tagalog anem	Ughele ɔnɔmo

`w·l|EIGHT` · binary · 75 languages · enrichment 151× · emergent `1·3`

Agta wʌlu	Arta walu	Bantik walu	Bolaang Mongondow wolu
Gaddang walu	Hiligaynon walo	Inagta Alabat wa'lu	Isneg walu
Kalinga walú	Koronadal Blaan walu	Manide walo	Maranao walo
Ratagnon waló	Sambal wa'lo	Subanun walu	Tontemboan walu

`p·t|four` · binary · 72 languages · enrichment 34× · emergent `1·t`

Agutyanen epat	Ayta Abellen apat	Bare'e pata	Bontok opat
Cebuano upát	Guiangan appat	Inati apat	Itawis appat
Kalinga opat	Lelepa pati	Misima epat	Ratagnon apat
Sinauna a'pat	Subanun pat	Thao B96 pat	Tungag / Tungak / apuat

`s·j·m|NINE` · ternary · 49 languages · enrichment 531× · emergent `2·j·m`

Aklanon - Bisayan syam	Ayta Mag-indi siyam	Bikol siyam	Bolinao syam
Cuyonon siyám	Hanunóo sijam	Inibaloi siyam	Itawis siyam
Kagayanen syam	Kalinga siyam	Kinaray-a siyam	Manobo siyam
Ratagnon siyám	Sambal siyam	Subanon siyam	Tagalog siyam

t·r|three · binary · 43 languages · enrichment 20× · emergent t·3

Angavae i–ɛtɹ	Babuza Ts82 tarua	Emae toru	Ifira–Mele e–toru
Kalinga tul̠u	Kavalan F69 turu	Mangareva toru	Mansaka turu
Nukumanu toru	Puyuma tero	Rarotongan toru	Seediq F69 teru
Tarpia tor	Tikopia toru	Uliveo i–tor	Yami atoro

h·n·m|SIX · ternary · 40 languages · enrichment 308× · emergent 1·n·m

Agta ʔʌnnʌm	Balangaw ʔɨ'nɨm	Gaddang ʔanɨm	Hanunóo ʔunum
Ifugao ʔenem	Inibaloi ʔɨnɨm	Itneg ʔɨ'nɨm	Kalagan ʔɨnɨm
Kamayo ʔunúm	Kapampangan ánam	Manobo ʔɨnɨm	Manobo ʔʌnʌm
Palawan Batak ʔɨnɨm	Sakizaya ʔənəm	Tagalog ánim	Tagbanwa ʔɨnɨm

r·m|five · binary · 37 languages · enrichment 39× · emergent 3·m

Arosi rima	Bauro erima	Emae rima	Futuna–Aniwa rima
Kanakanabu F69 rima	Kavalan F69 rima	Maori rima	Nukumanu rima
Paiwan rima	Port Sandwich ɛ–rim	Rarotongan rima	Saaroa F69 rima
Tahitian rima	Tikopia rima	Tuamotuan rima	Uruava ruma

h·p·t|four · ternary · 36 languages · enrichment 226× · emergent 1·1·t

Agta ʔʌppat	Bikol ʔə'pat	Dumagat ʔɨ'pat	Hanunóo úpat
Ifugao ʔepat	Inibaloi ʔɨpat	Itbayaten ʔapat	Kagayanen ʔáppat
Kalinga ʔo'pat	Kamayo ʔupát	Mamanwa ʔopat	Manobo ʔɨpat
Palawan Batak ʔə́pat	Sambal aʔpat	Tagalog ápat	Tagbanwa ʔɨpat

v·t|four · binary · 35 languages · enrichment 44× · emergent 1·t

Aore a_v̈ati	Bilur vauat	Burmbar e–vɛt	Kandas vat
Lunga Lunga aivat	Magori vati	Mbirao vati	Molima veyata
Mpotovoro v̈at	Nese –vat	Piva evati	Sisingga vati
Siviti i–vat	Sowa ivêt	Tasmate vati	Tolomako vati

Body parts

A pattern appears here that recurs throughout the appendix: the same etymon with and without a possessive suffix yields two distinct codes.

Eye gives m·t (246 languages) and m·t·n (65). *Breast* gives s·s (184) and s·s·n (50). That final –n is the third-person possessive, and in many languages it comes hyphenated in the data itself: *mata-n*, *susu-na*. The code treats it as one more consonant, because it does not know it is morphology.

This is not a flaw in the catalogue: it is morphological information emerging from it. That a concept appears systematically with and without a final consonant in the same position is exactly the clue that would lead to identifying a suffix, and it is the route of the pending segmentation (§7.3).

Tongue is the most dispersed case: d·l (52), l·l (42), m·n (38) and d·l·h (37). d·l and d·l·h are the same etymon with and without the final glottal — *dila* against *dilaʔ* — and l·l is compatible with them. m·n is another word (*mea*, *mea-na*, *mena-*), and its enrichment of 8× gives it away: among the lowest in the appendix, which the paper teaches to read as noise. Incidentally, *mea* / *mea-na* shows again the possessive suffix this section opens with.

m·t|eye · binary · 246 languages · enrichment 53×

Agta mʌtʌ	Arta mata	Bantoanon mata	Bughotu mata
Emae mata	Hiw məta-	Isneg mata	Kavalan F69 mata
Malalamai mata	Mansaka mata	Mota matai	Palawan Batak mata
Ririo mat	Sinaugoro mata-	Tangoa m̈ata-	Vaghua mata-

s·s|BREAST · binary · 184 languages · enrichment 95× · emergent 2·2

'Are'are susu	Anuki susu	Bantik susu	Buli sus
Hanunóo susu	Itbayaten soso	Kankanaey soso	Lau susu-
Malo susu	Marino sus	Na:ti sʉsʉ-	Pangasinan suso
Roria susu-	Sewa susu	Tagalog su:su	Torau susu

l·m|hand · binary · 93 languages · enrichment 41× · emergent 3·m

Akolet lama	Babuza Ts82 lima	Bontok lima	Fortsenal lima-
Ibanag lima	Kandas a_lama-	Lakalai lima	Lunga Lunga a_lima
Marino lima-i	Navut lima-	Pangasinan lima	Puyuma ɭima
S: Raga: Avavanva lima	Sinauna alíma	Tiale lima	Western Fijian lima

m·t·n|eye · ternary · 65 languages · enrichment 77×

Alor m'ɑtaŋ	Bilur omatana	Burmbar mata-n	Gedaged matan
Iduna matana	Kusaghe mata-na	Likum mita-n	Malango mata-na
Misima mata|na	Nahavaq mɛtɛ-n	Nggela mata-na	Roviana mata-na
Siviti mɛta-n	Tesmbol mɛtɛ-n	Ubir mata|n	Uripiv: Durpif mɛtɛ-n

t·l·n|ear · ternary · 54 languages · enrichment 78× · emergent t·3·n

Alor t'iluŋ	Bare'e taliŋa	Biliau talaŋa-u	Gitua taliŋa
Kagayanen talíŋa	Kove taliŋa	Malalamai taliŋa	Mangap-Mbula talŋa-
Mutu taliŋa	Nias taliŋa	Pukapuka talinga	Simbo taliŋa
Tigak talaŋa	Uruava talina	Vangunu taliŋi	Western Bukidnon teliŋa

d·l|tongue · binary · 52 languages · enrichment 92× · emergent d·3

Alangan dila	Ayta Ambala dila	Balangaw 'dilʌ	Batangan dila
Capisano dila	Gaddang dila	Hanunóo dila	Ifugao di:la
Isinay dila	Itneg 'dila	Kalinga dela	Kallahan dila
Kinaray-a dila	Pangasinan dila	Sangil 'dila	Tadyawan dilâ

s·s·n|BREAST · ternary · 50 languages · enrichment 133× · emergent 2·2·n

'Are'are susu-na	Bauro susu-na	Dori'o susu-na	Koro süsu-ŋ
Lau susu-na	Lunga Lunga a_susuna	Marino susuŋa	Mbirao čuču-na
Nduke susu-na	Pangkumu susi-n	Roviana susu-na	Shark Bay sisia-ŋ
Talise čuču-na	Tawaroga susu-na	Ubir susu|n	Uruava susú-na

t·n|INTESTINES · binary · 47 languages · enrichment 13×

Ambrym teña	Aore tine-	Buol tinae	Cebuano tiná-i
Gedaged taen	Kapingamarangi tinae	Lihir atien	Malo tine
Masbatenyo tinâi	N: Havai: Lombaha tinae	Orkon tine-	Sangil ti'nai
Sowa têne-	Surigaonon tina-i	Tangoa tine-	Uvea tinae

`l·l|tongue` · binary · 42 languages · enrichment 18× · emergent `3·3`

Antaimoro lela	Antambohoaka lela	Antankarana lela	Bara lela
Buhutu lelo	Itbayaten lila	Luangiua alelo	Masikoro lela
Mikea lela	Niue alelo	Pukapuka alelo	Sakalava lela
Sihanaka lela	Tikopia alelo	Tsimihety lela	Wuna lela

`m·n|tongue` · binary · 38 languages · enrichment 8×

'Are'are mea-na	Ambrym mæ-ŋ	Burmbar mɛ-ŋ	Dori'o mea-na
Ghanongga mea-na	Iamalele meana	Kusaghe mea-na	Kwara'ae mea-n
Luqa mea-na	Marovo mea-na	Namakir mena-	Port Sandwich mɛ-n
Sa'a mea-na	Tawala mena-	To'ambaita mea-na	Uliveo mea-n

`d·l·h|tongue` · ternary · 37 languages · enrichment 196× · emergent `d·3·1`

Aklanon - Bisayan dí:laʔ	Binukid 'dilaʔ	Bolinao dilaʔ	Dumagat di'laʔ
Iban dilah	Inati dilaʔ	Kalagan dilaʔ	Mamanwa dilaʔ
Manobo 'dilaʔ	Manobo dilaʔ	Mansaka dilaʔ	Romblon díla'
Sambal 'dilaʔ	Subanon dilah	Surigaonon dilaʔ	Tiruray dilaʔ

Kinship and person

Kinship separates two things worth not confusing.

Father appears with four codes, and their enrichments order them by themselves: `t·m·n` (153×), `t·m` (76×), `m·m` (24×) and `t·t` (22×). The first two are one etymon with and without a suffix; the last two are nursery words — *mama*, *tata* — of the kind that turns up in wholly unrelated families and that the comparative method sets aside for that reason.

Enrichment tells them apart without knowing any of this: a nursery word is made of the most frequent and most repeated consonants, so its enrichment falls well below that of an inherited etymon: 24× and 22× against the 76× and 153× of the first two. It is a case where the figure does the work the comparativist's criterion would do.

`t·m|FATHER` · binary · 105 languages · enrichment 76×

'Auhelawa tama	Anejom etma-	Bariai tama	Buhutu tama
Gitua tama-	Kiribati táma	Luqa tama	Manam tama
Morouas tama-	Navut tama-	Papitalai tama-	Saliba tama
South Efate tem	Tangoa tam̈a-	Tolai tama-	Varisi tama-

`k·m|YOU` · binary · 103 languages · enrichment 70× · emergent `2·m`

Akei ikomiu	Atchin: Orap kami	Bare'e komi	Butuanon kamo
Hanunóo kamu	Koro kimi	Löyöp kimi	Masbatenyo kamo
Mwerlap kami	Ninde kamo	Ratagnon kamó	Satawalese kámi
Tadyawan kam	Tasmate kamiu	Tunjung kam	Vurës ekimi

`n·n|MOTHER` · binary · 76 languages · enrichment 19×

Ali nʏanʏ	Antanalana nene	As nen	Biliau nin
Futuna-Aniwa nana	Hula ina-na	Koluwawa inana	Mahafaly nene
Mikea nene	Nauru inen	Orkon nina	Sakalava neny
Talise na-na	Tawaroga ena-na	Tsimihety niny	Western Fijian Nene

t·m·n|FATHER · ternary · 70 languages · enrichment 153×

Ambrym temæ-ŋ	Bellona tamana	Bunama tamana	Gedaged tamán
Haku tamana	Koro tama-ŋ	Lonwolwol timæ-n	Marovo tama-na
Muyuw taman	Nauna tama-n	Nombotkote atɪmæn	Roviana tama-na
Sikaiana tamana	Tape etmen	Ughele tama-na	Vaghua tama-na

m·m|FATHER · binary · 58 languages · enrichment 24×

'Are'are mama	Amis mama	Avava mom	Bughotu mama
Fifti: Khatbol mama	Gitua mama	Inati mama	Kokota mama
Leleghia mama	Löyöp mam	Mindiri mamə	Mota mama
Neverver mama	Papitalai momo	Rukai mamaa	Tomoip mimma

t·n|MOTHER · binary · 56 languages · enrichment 13×

Akei tina-	Baluan tina	Buhutu tina	Dehu thin
Kiribati tína	Luqa tina	Maleu tna	Marovo tina-
Motu tina-	Nukumanu tinna	Piamatsina tina-	Samoan tinā
Sinaugoro tina-	Sudest tïnae	Tangoa tina-	Tutuba tina-

n·t|CHILD · binary · 55 languages · enrichment 31×

Akei natu-	Aore natu-	Fortsenal natu-	Iamalele natu
Kandas nat	Mafea natu-	Manam natu	Minaveha natu
Morouas natu-	Namakir nati-	Nese nat-	Patpatar natu
Sivisa Titan nat	Sye nete-	Tiale natu-	Tungag / Tungak / nat

t·t|FATHER · binary · 51 languages · enrichment 22×

Alavas: Alavas tɛte	Avok tata	Burmbar tata	Ctl: Nalemba: Nar tata
Kapampangan tata	Maii tata	Mpotovoro tɛte	Nahavaq tatai
Ninde tatai	Nombotkote tata	Patpatar tata	Rano: Chinambong tɛte
Siviti tate	Tasmate tata	Tirax tata	V'ënen Taut tate

k·m|WE · binary · 47 languages · enrichment 103× · emergent 2·m

Alangan kamí	Ayta Ambala kami	Bare'e kami	Bikol kami
Bunun F69 kam	Cebuano kami	Gaddang ikkami	Inati ikám
Kalinga kamí	Kapampangan ikami	Maloh ikam	Mansaka kami
Palawan Batak kámi	S.W. Palawano kami	Surigaonon kamo	Tausug kami

n·k|CHILD · binary · 41 languages · enrichment 34× · emergent n·2

Ambel nikwɛ	Ayta Ambala anak	Babuyan anak	Bantoanon anak
Bolinao anak	Cebuano anak	Hiligaynon anak	Iban anak
Itbayaten anak	Kankanaey anak	Lala naku	Masbatenyo anak
Pangasinan anak	Sambal anak	Surigaonon anak	Waray-Waray anák

n·r|THEY · binary · 39 languages · enrichment 27× · emergent n·3

Agutyanen nira	Ambae niɛr	Antaisaka anareo	Araki nira
Banoni nari	Halia nori	Lelepa naara	Morouas nira
N: Havai: Lombaha ŋere	Nakanamanga: Sesa nara	Piamatsina iŋire	Rano: Chinambong nir
Tiang ner	Uripiv: Durpif niɾ	Vunapu inire	Wailapa inira

`s·n|WHO` · binary · 38 languages · enrichment 16× · emergent `2·n`

Amara seiŋa	Babuyan sino	Bikol sin-o	Hiligaynon sin-o
Ilokano sinno	Kalinga 'sinu	Kankanay si'no	Kwamera sin
Mansaka sini	Ratagnon sino	Sa sien	Siviti sena
Sori seŋ	Surigaonon sin-o	Tambotalo sin	Unua asiŋo

`m·n|MALE PERSON` · binary · 37 languages · enrichment 8×

'Are'are mane	Ambel man	As -man	Bughotu mane
Gimán maón	Kokota mane	Longgu mwane	Malango mane
Minyaifuin man	Mortlockese mwæ:n	Numfor man	Oroha mane
Puluwatese m^{w}áán	Satawalese m^{w}áán	Tawaroga mwane	Waropen mano

`n·k|THOU` · binary · 36 languages · enrichment 35× · emergent `n·2`

Akei iniko	Araki niko	Baetora niko	Koro nik
Lemerig næk	Marino iniko	Mwerlap iniok	**N:** Leonda nieko
Nokuku nik	Orkon nak	Rano: Chinambong nik	Tangoa e^{ŋ}ko
Tigak nuk	Tonsea ni ko	Vera'a enike	Wailapa iniko

`l·l·k|MALE PERSON` · ternary · 35 languages · enrichment 169× · emergent `3·3·2`

Agta lʌlʌki	Ayta Mag-antsi lalaki	Bikol lalake	Bolaang Mongondow lolaki
Capisano lalake	Cuyonon lalaki	Hanunóo lalaki	Ibanag lalaki
Inati lalaki	Itawis lalaki	Kalinga laláki	Kapampangan lalaki
Masbatenyo lalaki	S.W. Palawano lelaki	Sambal la'laki	Tagbanwa lalaki

Nature and sky

Stone with three codes — `b·t` (98), `v·t` (49), `f·t` (38) — is the labial series of §6.3 laid out within a single concept: *b*, *v*, *f* in the same position, across three groups of languages.

Sea illustrates the other extreme: `t·s` (57 languages, 68×) and `d·g·t` (49, 248×) are not variants of one etymon, they are two different words. The second's enrichment gives it away: a rare sequence repeated in forty-nine languages.

`k·l·t|LIGHTNING` (137×) and `b·l·n|moon` (149×) are textbook ternaries: three positions to get right, wide reach and high enrichment.

`b·t|stone` · binary · 98 languages · enrichment 113× · emergent `1·t`

Agta bʌtu	Ayta Ambala bato	Balangaw bʌ'to	Bikol ba'to
Bujan bat	Dupaningan Agta bitu	Hoanya Ts82 ba:tú:	Ilokano bato
Itbayaten bato	Kalinga ba'tu	Kapampangan batu	Mamanwa bato
Manobo bato	Pazih F69 batu	Sambal ba'to	Tadyawan bató

`t·n|earth` · binary · 92 languages · enrichment 22×

Ambae tan	Antaimoro tany	Antanosy tany	Bantik tana
Betsimisaraka tane	Ctl: Nalemba: Nar tano	Gitua tano	Lonwolwol tan
Mangap-Mbula toono	Merina_Malagasy tàny	N: Havai: Lombaha tano	Nulnesa tono
Sa tan	Sio tono	Tsimihety tany	Wab ton

b·l·n|moon · ternary · 71 languages · enrichment 149× · emergent 1·3·n

Agutyanen bolan	Bahasa Indonesia bulan	Bikol bo'lan	Bolinao bulan
Cuyonon bulán	Hoanya Ts82 bu-lan	Ifugao bu:lan	Iraya bulán
Kalinga bu'lan	Kavalan F69 bu:lán	Manide bilan	Manobo bolan
Pangasinan bulan	Sambal bolan	Tadyawan bulán	Tboli bulon

l·n|wind · binary · 69 languages · enrichment 30× · emergent 3·n

Akei lani	Aore laŋ	Butmas-Tur laŋ	Eton laŋ
Lamenu laŋ	Lonwolwol leŋ	Malo laŋi	Mwerlap laŋ
Narango lan	Nulnesa laŋ	Port Vato leŋ	S: Raga: Avavanva laŋi
Siviti laŋ	Tasmate lan	Uripiv: Durpif liŋ	Vurës leŋ

r·n|day · binary · 68 languages · enrichment 26× · emergent 3·n

Akei rani	Apma ren	Ctl-E: Tawet: Nav ran	Ghanongga rane
Larevat irɛn	Malo rani	Mokilese reen	N: Leonda rani
Nduke rane	Piamatsina ran	Roro rani	Sakao ɛrɛn
Sivisa Titan raŋ	Tambotalo raŋia	Tolomako ran	Vaghua ranə̥

t·n|year · binary · 59 languages · enrichment 16×

Akei tauni	Antandroy taogne	Berawan_Long_Tera taun	Bikol táon
Buol taun	Inibaloi tao-en	Malo tauni	Merei tauon
Morouas taone	Numfor taun	Roria taon	Sihanaka taona
Tambotalo taunua	Tiale taoni	Tsimihety tagno	Wailapa taun

t·s|sea · binary · 57 languages · enrichment 68× · emergent t·2

Akei tasi	Angavae ε-t^{h}a·s	Bare'e tasi	Ctl: Nalemba: Nar tas
Ghari tasi	Kandas tasi	Loniu tas	Mafea tasi
Mbirao tasi	Narango tas	Nokuku tɒs	S: Raga: Avavanva tasi
Sissano tias	Tambotalo tasia	Tape tes	Tiang tas

k·l·t|LIGHTNING · ternary · 51 languages · enrichment 137× · emergent 2·3·t

Agta kilat	Ayta Abellen kilat	Bikol kil-at	Buol kilat
Dumagat ki'lat	Guiangan kilat	Ifugao ki'lat	Iranun kilat
Kalagan kilat	Kinaray-a kilat	Maloh kilat	Manobo kilat
Maranao kilat	Ratagnon kilát	Subanon kilat	Tausug kilat

d·g·t|sea · ternary · 49 languages · enrichment 248×

Aklanon - Bisayan dá:gat	Ayta Mag-antsi dagat	Bikol dagat
Bolaang Mongondow dagat	Capisano dagat	Dumagat di'get
Hiligaynon dagat	Kagayanen dágat	Kamayo dá:gat
Mamanwa dagat	Manobo 'dagat	Manobo dagat
S.W. Palawano dagat	Subanon dagat	Tadyawan dagát
Tausug dagat		

`v·t|stone` · binary · 49 languages · enrichment 61× · emergent `1·t`

```
Angavae ɛ–vɐt       Antambohoaka vato  Antankarana vato  Avava a–vat
Betsimisaraka vato  Iaai veto          Kavalan F69 vato  Mahafaly vato
Meramera vatu       Mikea vato         Namakir vat       Raga vatu
Sihanaka vato       Talise vatu        Tesmbol i–vɛt     Vezo vato
```

`b·n|night` · binary · 41 languages · enrichment 41× · emergent `1·n`

```
Andra biŋ    Babuza Ts82 bi–ini  Bariai boŋ         Bughotu boŋi
Lau boŋi     Malalamai boŋ       Marshallese bʷéŋʷ  Mokaren biŋ
Neve'ei buŋ  Pak bu:ŋ            Ponam biŋ          Saipan Carolinian bʷooŋ
Simbo boŋi   Tawaroga bwɔŋi      Tiang bɨŋ          Tulu bI:n
```

`m·s|star` · binary · 39 languages · enrichment 16× · emergent `m·2`

```
Akei matˢoi     Avok mʷaʧi         Buninga: Moso masoe  Fortsenal matˢoi
Koro mʷese      Mafea masoa        Mbwenelang mʷose     Nisvai mʷaʧi
Orkon mɨso      Piamatsina matˢoe  Ske mso              Sye mosi
Tesmbol mʷose   Unua mat͡ʃo         V'ënen Taut məsi     Vunapu matˢoi
```

`f·t|stone` · binary · 38 languages · enrichment 50× · emergent `1·t`

```
Ambae fat    Amblong fatu     Bontok fiato     Ctl: Nalemba: Nar fat
Futuna fatu  Ifira–Mele fatu  Leo Tuvalu fatu  Marino fato
Narango fat  Polonombauk fat  Samoan fatu      South Efate fat
Takuu fatu   Thao B96 fatu    Tsou T63 fatu    Uvea fatu
```

`n·b·n|night` · ternary · 37 languages · enrichment 72× · emergent `n·1·n`

```
Ambae ᵐboiŋ    Bieria iᵐboŋ     Ctl: Nalemba: Nar ᵐbʷoŋ  Ghanongga mbɔŋi
Kusaghe mboŋi  Longgu mboŋi     Maii iᵐboŋ               N: Havai: Lombaha ᵐboŋi
Narango ᵐbon   Nese ᵐbun        Pangkumu aᵐboŋ           S: Raga: Avavanva ᵐboŋi
Talise mbɔŋi   W: Walaha ᵐboŋi  camuki ᵐboen             ʰMoavekɛ ᵐboen
```

`r·n|RAIN (PRECIPITATION)` · binary · 36 languages · enrichment 14× · emergent `3·n`

```
Antambohoaka orana  Arosi raŋi      Ayta Mag–antsi uran  Basay TsYM91 uran
Bikol uran          Cuyonon urán    Ibanag uran          Iranun uran
Kapampangan uran    Kinaray–a uran  Masbatenyo uran      Merina_Malagasy orana
Sakalava orana      Sihanaka orana  Tanala orony         Tsimihety oragna
```

Animals

`m·n|bird` gathers 172 languages at 36× and is the widest binary in the section. It is worth comparing with `m·t|snake` (95 languages, 23×), which shares its skeleton with `m·t|eye` (246, 53×) and `m·t|BE DEAD OR DIE` (204, 47×): these are separate sets — each with its concept — but *snake*'s enrichment is the lowest of the three, and the figure warns that `m·t` is too common to support anything on its own.

And a second 'snake' teaches more. `n·m·t|snake` (35 languages, 72×) is the same etymon with a nominal prefix that the data itself brings hyphenated: *na-mat*, *nə-mat*, *na-mate*, *n-mat*. The catalogue records it as a distinct code because it does not know na– is morphology — the same phenomenon §5.4

describes for 'black', but here in a noun and with the boundary visible in the transcription. Anyone wanting to see at a glance why segmentation is missing will find this block clearer than any explanation.

`m·n|bird` · binary · 172 languages · enrichment 36×

```
'Are'are manu    Arop man                Bauro amanu    Ctl: Nalemba: Nar man
Hiw mɔn          Kove manu               Lihir mon      Mangseng mɛn
Moeneᵐbeŋ meni   Nakanamanga: Sesa manu  Nulnesa ma:nu  Rapa Nui manu
South Efate man  Tarpia man              Tongan manu    Vurës mʷon
```

`k·t|louse` · binary · 135 languages · enrichment 65× · emergent `2·t`

```
Agta kutu     Apma kut       Basay TsYM91 kutu  Buol kutu
Futuna kutu   Ibanag kutu    Itbayaten koto     Kankanaey koto
Lelepa kuutu  Manide kuto    Maranao koto       Namakir kit
Penrhyn kutu  Sikaiana kutu  Tagbanwa kutu      Uvea kutu
```

`n·m|MOSQUITO` · binary · 104 languages · enrichment 44×

```
Agutyanen namo     Axamb nam     Bughotu ñamu    Ctl-E: Tawet: Nav nam
Futuna-Aniwa namo  Kilivila nim  Lemerig nœm     Maewo na:mu
Misima nimu        Nalik namu    Nukuoro namu    Puluwatese -nómʷ
Samoan namu        Sobei namu    Tuamotuan namu  Vera'a nam
```

`m·t|snake` · binary · 95 languages · enrichment 23×

```
'Oyaoya mata       Anuki mota    Bebeli mɔtɔ   Bulu mata
Eton mʷat          Kove mota     Loniu mʷat    Malasanga mota
Minaveha mwata     Mota mʷata    Nggela umata  Saliba mwata
Sivisa Titan moat  Talise muata  Tigak muata   Vunapu mata
```

`t·l|egg` · binary · 47 languages · enrichment 18× · emergent `t·3`

```
Ambae tol              Ambel talo         Antanosy atoly  Biga tolo
Ctl-E: Tawet: Nav tol  Ifira-Mele atolu   Lengo tolu      Mahafaly atoly
Marino toli            Mota toliu         N: Leonda tol   Nggela tolu
S: Raga: Avavanva toli Sakao atœl-        Tunjung talo    Vao tol
```

`k·l|dog` · binary · 35 languages · enrichment 16× · emergent `2·3`

```
Ambrym kuli     Apma kuli     Baki kuli    Bieria kuliu
Daa kaka kuli   Futuna kulii  Lamenu kuli  Leo Tuvalu kulii
Nahavaq kilu    Nembao kuli   Niue kulii   Pukapuka kulī
Sikaiana kulii  Tanna kuli    Uvea kulii   Vaeakau-Taumako kuli
```

Plants and food

Drink yields `m·n` (81) and `n·m` (69). Note these are the same two consonants in reverse order. The catalogue treats them as distinct codes, and rightly: order is part of the code. But the coincidence is conspicuous and deserves examination — it may be metathesis, or they may be two words.

`b·n·g|fruit` (133×) shows the usual contrast with neighbouring verbs: concrete concepts tend to high enrichment, very general ones (`n·n|eat`, 9×) to enrichments grazing chance.

`m·n|drink` · binary · 81 languages · enrichment 17×

```
Ambrym -mün            Antanosy mino  Babuza Ts82 man o  Bierebo -munia
Ctl-E: Tawet: Nav min  Gimán imin     Lelepa munu        Maii -mun
Na'ahai -mʉn           Nasarian -mɪn  Nisvai -mun        Pangkumu -min
Sa imini               Sori mwiŋ      Tape -mɪn          Uripiv: Durpif -mini
```

n·m|drink · binary · 69 languages · enrichment 28×

```
'Auhelawa numa    As -nim          Ayta Mag-indi inum  Buhutu nom
Hawaiian inu-mia  Itawis inúm      Kinaray-a inəm      Maori inu-mia
Mindiri ŋa-mi     Mpotovoro -num   Numbami -numu       Ramoaaina inim
Sambal inom       Surigaonon inom  Tigak inum          Wampar nom
```

k·n|eat · binary · 69 languages · enrichment 25× · emergent 2·n

```
Akolet ken               Arop a-kani              Axamb -kan     Ayta Mag-indi ɨkan
Bolaang Mongondow kaan   Fifti: Khatbol -kan      Itbayaten kan  Kiribati kana
Maguindanaon kan         Maleu kan                Maranao kan    Na'ahai -kan
Nitita -kan              Rano: Chinambong ku-ain  Sambal kan     Surigaonon kaon
```

s·n|salt · binary · 51 languages · enrichment 30× · emergent 2·n

```
Alangan asín   Ayta Mag-indi asin  Batangan asín     Bikol asín
Butuanon asin  Chuukese san        Ibanag asin       Inibaloi asin
Itawis asin    Kalinga asín        Kapampangan asin  Masbatenyo asin
Romblon asín   Sangir asiŋ         Tadyawan asín     Tontemboan asiŋ
```

m·m|CHEW · binary · 51 languages · enrichment 35×

```
Ansus ɛmama   Bali-Vitu mama  Bikol mama    Gedaged mam
Isneg mama    Kalinga mu'ma   Kwaio meme-a  Lunga Lunga  mamai
Namakir mama  Numfor mam      Roro momo     Sikaiana maamaa
Tetum mam:a   Torau mama      Vitu mama     Windesi Wandamen mam
```

b·n·g|fruit · ternary · 41 languages · enrichment 133× · emergent 1·n·g

```
Atta bunga:               Bantoanon bunga  Binukid 'bunga         Capisano bunga
Guiangan bungo            Ifugao bu:nga    Inati búnga            Isinay bunga
Kakiduge:n Ilongo bunga   Kalinga búnga    Koronadal Blaan bɨngɨ  Masbatenyo bunga
Pangasinan bunga          Sambal 'bonga    Sinauna bunga          Tagalog bunga
```

r·n|leaf · binary · 40 languages · enrichment 14× · emergent 3·n

```
Akei rau-na         Arop rau-na             Axamb raun    Doura rauna
Lou rei-n           Malasanga rau-na        Mor ranu      Nasarian a-raon
Pangkumu raon       Rano: Chinambong roin   Sissano ro-ñ  Tape ron
To'ambaita rea-na   Unua a-raon             Vunapu rauna  Wusi rauna
```

n·n|eat · binary · 36 languages · enrichment 9×

```
Amblong anian    Atchin: Orap an-ɛn   Baluan ŋan     Fortsenal aniani
Kandas i_uaŋon   Loniu ñani           Mafea anan     Merei anian
Morouas aniani   Namakir ŋan          Navut anani    Nisvai -nɪn
Sihanaka inana   Torau aniani         Ubir i|ani|n   Vangunu ŋaŋau
```

Verbs of action and state

The section where regular variation shows best.

Die yields `m·t` (204) and `m·s` (47): the palatalisation `*t > s` of §6.3. *Dig* yields `k·l` (77) and `x·l` (49): the spirantisation `*k > x`. *Hear* yields `r·n` (108) and `l·n` (47): the liquids again. Three distinct correspondences, each visible in one concept and confirmed in the global count.

`m·t·r|sleep` (90×) and `m·t·k|fear` (94×) are ternaries beginning with `m·t`, like *eye* and *die*. The third position is what separates them, and it is a good reminder of why binaries discriminate less (§2.3).

`m·t|BE DEAD OR DIE` · binary · 204 languages · enrichment 48×

'Oyaoya mate	Aria mete	Basay TsYM91 matai	Chamorro matai
Haku mete	Kankanaey matey	Lelepa mate	Malango mate
Mengen mate	N: Havai: Lombaha mate	Nitita -mat	Pukapuka mate
Sarangani Blaan mati	Sursurunga mat	Tikopia mate	Vaeakau-Taumako mate

`t·n|cry` · binary · 109 languages · enrichment 27×

Akei tani	Anejom tañ	Bariai taŋ	Fortsenal tani
Kove taŋi	Levei e|teŋ	Malalamai taŋ	Marquesan tani
N: Leonda tæŋ	Nggela taŋi	Penrhyn tangi	Roinji u-taŋ
Sobei e-tan	Tangoa taŋi	Uruava tani	Vurës teŋ

`r·n|hear` · binary · 108 languages · enrichment 39× · emergent `3·n`

'Are'are rono	Anuta roŋo	Butmas-Tur roŋ	Faghani roŋo
Kwara'ae rɔŋɔa	Lemerig rœŋ	Marino roŋo	Morouas ronea
Nahavaq -roŋ	Nggela roŋo	Piamatsina roŋo-i	Sa ɹoŋo
Ske -rŋo	Tawaroga roŋo	Tutuba rŋoa	Vunapu ronoa

`m·r|BE ALIVE` · binary · 77 languages · enrichment 35× · emergent `m·3`

'Are'are emauri	Anuta ma-uri	Axamb -maur	Emae mauri
Ifira-Mele mouri	Langalanga mouri	Longgu mauri	Mandri -maur
Motu mauri	Nasarian -maor	Nggela mauri	Nombotkote -maor
Port Sandwich -maur	Sa mur	Tawaroga mæri	Unua -maur

`k·l|dig` · binary · 77 languages · enrichment 38× · emergent `2·3`

Akolet keyl	Ambae keli	Avava ki-il	Bantik kali
Bola keli	Kandas i_kil	Lamenu -kil	Lwepe kel
Maranao kali	Niue keli	Palawan Batak kali	Ramoaaina kil
Satawalese kkel	Tagbanwa kali	Tungag / Tungak / kel	Western Fijian keli

`t·r|stand` · binary · 62 languages · enrichment 30× · emergent `t·3`

Akei turi	Aore turu	Fortsenal turi	Label tur
Malo turu	Morouas turi	Na'ahai -tur	Nduke turu
Neverver -tur	Patpatar tur	Roviana turu	Sursurunga tur
Tasmate turi	Tolai tur	Uripiv: Durpif -tur	Waropen otaro

m·n|laugh · binary · 58 languages · enrichment 13×

Akei mana	Amblong mana	Apma man	Baetora mana
Ctl: Nalemba: Nar mana	Hoanya Ts82 munē	Maewo mana	Marino man
N: Havai: Lombaha mana	Narango man	Orkon mæn	Port Sandwich -man
Sa man	Tasmate mana	Uliveo -man	Vunapu mana

m·t·r|sleep · ternary · 54 languages · enrichment 91× · emergent m·t·3

Ambae matur	Antanosy matory	Bara matory	Ctl: Nalemba: Nar matir
Fifti: Khatbol -matur	Lelepa maturu	Malo maturu	Merina_Malagasy matòry
N: Leonda matur	Namakir matir	Neverver -matur	Papitalai metir
S: Raga: Avavanva maturu	Sivisa Titan matir	Tanala matory	Uripiv: Durpif -matur

l·l|vomit · binary · 52 languages · enrichment 23× · emergent 3·3

Akei lulua	Arop i-lulu	Bali-Vitu lualua	Ctl: Nalemba: Nar lulua
Lamenu -lulua	Löyöp lolo	Malalamai lulua	Mpotovoro -lulue
Na'ahai i-lulu	Nasarian -lalu	Neverver -lialu	Nombotkote -lualua
Siviti -lolo	Tasmate lulua	Uvea lulua	Vera'a luluo

x·l|dig · binary · 49 languages · enrichment 106× · emergent 2·3

Baetora xil	Ctl: Nalemba: Nas xəl	Koro ɣil	Lemerig ɣil
Maewo xili	Marino xili	Mwotlap ɣil	Na'ahai i-ɣil
Navwien 162-87 -ɣəl	Neverver -xil	Nombotkote -ɣil	S: Raga: Avavanva xeli
Siviti -xil	Tangoa xeli	Tirax -xeli	V'ënen Taut -xəli

m·m|PISS · binary · 48 languages · enrichment 159×

'Are'are mimi	Asumboa mimi	Bughotu mimi	Ghari mimi
Kwaio mimi	Lau mimi	Luangiua mimi	Maori mimi
Mbaelelea mimi	Mbirao mimi	Nggela mimi	Roviana mimi
Simbo mimi	Tanema mimi	Tongan mimi	Vangunu mimi

m·s|BE DEAD OR DIE · binary · 47 languages · enrichment 21× · emergent m·2

Anejom mas	Atchin: Orap -mat͡s	Babuza Ts82 macha	Fifti: Khatbol i-mas
Kusaie misɛ	Mandri -mæʧ	Motu mase	Nahavaq -mɛs
Navwien 162-87 -mæʧ	Nisvai -mat͡s	Pangkumu -met͡ʃ	Port Sandwich -maʧ
Siviti -mes	Tesmbol -mes	Unua -mit͡ʃ	Whitesands -imis

l·n|hear · binary · 47 languages · enrichment 20× · emergent 3·n

Amara loŋo	Arop a-loŋo	Bariai -loŋo	Biliau loŋ
Buma leŋi	Hiw ^{g}ʟoŋ	Kapingamarangi longo	Kwaio loŋa-a
Lihir lan	Malasanga i-luŋ	Mangap-Mbula -leŋ	Minyaifuin lɔŋɔ
Sarangani Blaan liŋɨ	Takia -loŋ	Tungag / Tungak / loŋo	Vaeakau-Taumako longo

m·t|die · binary · 43 languages · enrichment 56×

Alor m'ɑtɛɛ	Antaisaka maty	Antandroy mate	Anuta mate
Bara maty	Betsimisaraka mate	Kusaghe mate	Lote mɛte
Malango mate	Merina maty	Mikea mate	Nggela mate
Sihanaka maty	Simbo mate	Tanala maty	Ughele mate

m·t·k|fear · ternary · 42 languages · enrichment 94× · emergent m·t·2

Ambae mataku	Angavae –matak	Bantik mataku	Bieria –matak
Ctl: Nalemba: Nas matak	Futuna mataku	Ifira–Mele mataku	Leo Tuvalu mataku
Maori mataku	Namakir matak	Nombotkote –mɛtak	Nukumanu mataku
Rapa Nui mataku	South Efate mtak	Tikopia mataku	Uvea mataku

t·n|burn · binary · 40 languages · enrichment 10×

Amara tin	Anuta tungia	Bellona tuungia	Chamorro tunu
Gedaged tun|i	Iban tunu	Kapingamarangi tungi	Lunga Lunga tunia
Minaveha tunu	Navut taon	Nukeria ttuni	Nyelâyu thînî
Sengseng tun	Takuu tuni	Tiang tan	Tuamotuan tuŋi

k·t|see · binary · 39 languages · enrichment 18× · emergent 2·t

Anuki kita	Ayta Ambala akit	Ayta Mag–indi akit	Dawawa kita
Futuna kite	Inati kítæ	Mamanwa kita	Manobo kita
Nukeria kite	Pazih F69 kita–	Rapa Nui kite	Saliba kita
Sikaiana kite	Tagalog ki:ta	Tuamotuan kite	Uvea kitea

s·l|burn · binary · 35 languages · enrichment 25× · emergent 2·3

Akei sulia	Aore sulia	Avava –sulai	Gimán slai
Lou suli	Mafea sulia	Marovo sulua	Morouas sulia
Narango sula	Navut sulu	Neve'ei suli	Roviana sulu–a
Siviti –sul	Tangoa suli	Tiale sului	Tutuba sulia

l·m|come · binary · 35 languages · enrichment 16× · emergent 3·m

'Auhelawa laoma	Buhutu laoma	Ghanongga lame	Halia lama
Langalanga lamae	Longgu lamai	Mbaengguu lea_mai	Mumeng lam
Neverver –lɛm	Puyuma aɭamo	Sa'a lae_mai	Saliba lao–ma
Suau laoma	Tanema loma	Tubetube laoma	Vaghua lama

n·r|cry · binary · 35 languages · enrichment 24× · emergent n·3

'Are'are nara	Baetora ŋara	Ctl: Nalemba: Nar ŋara	Fifti: Khatbol –ŋar
Maewo ŋara	Marino ŋar	Mpotovoro –ŋar	N: Havai: Lombaha ŋara
Nduke ŋire	Neverver –ŋar	Roria ŋar	Sakao ŋor
Siviti –ŋaɾ	Talise ŋara	Tutuba nara	Valpei ŋara

Objects and culture

House with n·m (62 languages, 26×) and b·l·j (42, 271×) is the clearest case of two different words for one concept, not two variants of one. The figures say so: the second is a rare ternary with ten times the enrichment.

The same in *path*: s·l (56, 39×) and d·l·n (38, 208×).

d·g·m|NEEDLE reaches 564× with only 41 languages — and §7.1 flags it, alongside b·l·j|house and s·l|path, as a borrowing candidate: §6.2 explains why high enrichment points to the local before the old. A sequence of three infrequent consonants repeated for the same object across forty-one languages is all but impossible by chance.

t·l|ROPE · binary · 70 languages · enrichment 26× · emergent t·3

Agutyanen tali	Antanosy taly	Balangaw ta'le	Buol tali
Guiangan toli	Iranun tali	Kalinga tali	Kankanay ta'li
Maloh tali	Maranao tali	Nasarian a-tal	Pukapuka taula
Sangil 'taḷi	Subanon tali	Toga tel	Vezo taly

n·m|house · binary · 62 languages · enrichment 26×

'Are'are nima	Angavae ne-im^{w}ɔ	Axamb na-im	Dori'o nima
Kayupulau [Kajupu numa	Magori numa	Mpotovoro ne-ime	Namakir næim
Nese naime	Nombotkote ne-im^{w}ɔ	Port Sandwich na-im	Saliba numa
Suau numa	Tanimbili nom^{w}a	Tesmbol ne-ma	Vao ne-me

s·l|PATH OR ROAD · binary · 56 languages · enrichment 39× · emergent 2·3

Akei sala	Angavae a-sɛlɛ	Baetora sala	Ctl-E: Tawet: Nav sala
Fortsenal sala	Maewo sala	Manam jala	Mbirao sala
Narango sal	Navwien 162-87 sal	Port Vato sili	Rotuman sala
Ske sal	Tambotalo sala	Tomoip sal	Wailapa sala

t·p|THATCH OR ROOF · binary · 46 languages · enrichment 66× · emergent t·1

Alangan atop	Ayta Abellen atɨp	Babuyan atep	Bolaang Mongondow atop
Butuanon atop	Cuyonon atɨp	Hanunóo atup	Ilokano atep
Isinay atop	Itbayaten atep	Kankanaey atep	Maranao atep
Romblon atóp	Surigaonon atop	Tonsea atep	Western Bukidnon atep

b·l·j|house · ternary · 42 languages · enrichment 271× · emergent 1·3·j

Agutyanen balay	Atta balay	Bikol balay	Binukid 'balay
Cuyonon baláy	Guiangan bɨla^{o}y	Hiligaynon balay	Ilokano balay
Isneg balay	Kalagan balay	Kinaray-a balay	Manobo balɨy
Masbatenyo balay	Sinauna baláy	Subanun balay	Tagbanwa balay

k·j|stick · binary · 41 languages · enrichment 132× · emergent 2·j

Agta kʌyu	Babuyan kayu	Binukid 'kayu	Gaddang kayo
Inati káyo	Itawis kayu	Kaidipang kayu	Kalinga 'kayu
Maloh kaju	Manobo kayu	Manobo kayo	Nauna kɪy
S.W. Palawano kayo	Sarangani Blaan kayo	Tagbanwa kayu	Western Bukidnon kayu

d·g·m|NEEDLE (FOR SEWING) · ternary · 41 languages · enrichment 564×

Agta dʌgum	Atta da:gum	Bikol dagom	Butuanon dagom
Dumagat di'gum	Hanunóo dagum	Ilokano dagum	Isneg da:gum
Kagayanen dágəm	Kallahan dagum	Kankanay 'dagom	Manobo dagom
Manobo dagum	Mansaka dagɨm	Pangasinan dagum	Surigaonon dagom

d·l·n|PATH OR ROAD · ternary · 38 languages · enrichment 208× · emergent d·3·n

Agutyanen dalan	Atta dalan	Batangan dalan	Buol dolan
Cuyonon dalán	Gaddang dallan	Ibanag dalan	Iraya dalan
Kallahan dalan	Maloh dalan	Masbatenyo dalan	Pangasinan dalan
Ratagnon dalan	Sangir daleŋ	Tadyawan dalan	Tiruray dolon

Qualities

And here the appendix closes with its own counter-example, which is the most useful thing in it.

The enrichments in this section are the lowest of the whole appendix: m·t|black 8×, n·n|yellow 10×, m·t|right 16×, m·r|left 26×. The exception is m·m·s|dry at 118×.

The cause is identified and systematic. Many Austronesian adjectives carry the stative prefix *ma-*, which puts an initial m on all of them. The code cannot distinguish that prefix from a root m (§5.4), so it groups them by what they share — the prefix — rather than by their root. The result is codes that gather many languages and say little, and enrichment detects it.

Read this section as the warning of where the method performs worst: on concepts whose morphology introduces shared material that is not the root. It is the strongest argument for resolving segmentation.

m·t|right · binary · 71 languages · enrichment 16×

Akei matua-	Aore matua-	Ctl-E: Tawet: Nav matua	Futuna-Aniwa matau
Kokota mautu	Lehali moto	Lonwolwol mætu	Maori matau
Mota matua	Narango matua	Pukapuka mātau	Simbo matua
Tangoa m̈atua-	Tikopia matau	Uliveo matu	Vurës moto

m·r|left · binary · 68 languages · enrichment 26× · emergent m·3

'Are'are mauri	Atchin: Orap mʷir	Buma mɔrɔ	Eton mour
Kwamera mour	Lelepa maur	Malo mar̃au	Nahavaq meir
Neverver mʷɛr	Piamatsina marau	Ririo mør	Sisingga məri
Sye mor	Tesmbol maɛɾ	Tutuba marau	Vano mʷaure

m·m·s|dry · ternary · 46 languages · enrichment 118× · emergent m·m·2

Atchin: Orap -mɛmɛs	Baetora mamasa	Burmbar -mamas	Kove mamasa
Löyöp mamas	Malo mamasa	Mbwenelang -memɛs	Mutu mamas
Nisvai -mimɪs	Numbami -mamasa	Patpatar mamasa	Rotuman mamasa
Sio mamɔsa	Talise mamača	Tesmbol -mɛmɛs	Wab mamos

m·t|black · binary · 39 languages · enrichment 8×

Ambae maeto	Angavae -mɛtɛ	Antanosy meity	Bara mety
Betsileo maity	Ctl: Nalemba: Nas maito	Maewo maeto	N: Leonda maeto
Nese -mot	Ninde -mite	Nombotkote i-mɛt	Piamatsina maeto
Tape -mit	Tesmbol -miɛt	V'ënen Taut i-mɪt	W: Walaha maeto

n·n|yellow · binary · 36 languages · enrichment 10×

Avok -iaŋiaŋ	Blablanga ŋɔŋe	Burmbar -ŋaŋ	Hiw ŋaŋo
Lehali neŋoi	Leleghia ŋɔŋe	Manam aŋoaŋo	Mengen 'ŋaŋo
Mwerlap aŋaŋ	Nauna aŋoan	Nggao ŋɔŋe	Nisvai -iaŋiaŋ
Port Sandwich -ŋaŋ	Tesmbol -iaŋiaŋ	Uliveo -iaŋiaŋ	Uripiv: Durpif ɛ-ŋaŋɔ