Consonant Codes in 207 Indo-European Languages

A catalogue built blind

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

Abstract

We present a catalogue of 10,116 consonant codes grouping 94,152 lexical forms from 207 Indo-European languages. A code is a consonant skeleton shared by words for the same concept across distinct languages.

This is the fifth family in the series and the first with a dense comparative literature. The catalogue was built without consulting it: no reconstruction, no cognate lists, no etymology, and no comparison of any result against two centuries of the comparative method. That is this work's principal design decision, and §1.1 explains why it is taken and what it buys.

The figures: precision 8.73 % against a chance rate of 0.084 %, that is 104×, the second value of the series after Turkic. Collision is 52.5 % and mean arity 3.35, the highest, tied with Pama-Nyungan. The catalogue covers 1,946 concepts, more than twice any preceding family.

Four results. `n·m|name`, with 103 languages, is the widest-reaching code in the whole worked corpus. The 61 branch pairs give the sharpest alternation concentrations of the series. Geography is monotonic across all six buckets despite the family having languages transplanted to other continents. And automatic affix discovery again worsens the advantage over chance — the fourth consecutive family.

1. What is being looked for, and why Indo-European blind

Historical linguistics has a powerful and expensive instrument: the comparative method. It has been applied with enormous unevenness, and Indo-European is the extreme of that unevenness — the family with the densest reconstruction, the most reviewed expert cognacy and the deepest literature in existence. Any study resting on that layer will measure, among other things, the density of the literature.

The series this work belongs to asks whether something useful can be done by giving up reconstruction: not improving or replacing it, but doing without it and seeing what remains. The four preceding families — Turkic, Uralic, Austronesian, Pama-Nyungan — were chosen partly because their comparative literature is less dense, so that giving it up cost little.

Here it costs everything. Which is why it is done.

1.1 The commitment: not looking at the answer

This catalogue was built without consulting Indo-European reconstruction. Specifically:

- No protoform is used, neither to build the catalogue nor to read it.

- No cognate list is used. The `iecor` source, which contributes 25,731 forms to this corpus, carries expert cognacy judgements and they have not been read: from it we take forms and glosses, as from any other source.
- No sound law is cited, neither by name nor by content.
- No result is compared against the family's established classification beyond the branch labels the corpus already carries as metadata, used exactly as in the four preceding families: to read an already-closed catalogue, never to build it.
- The single reconstructed lect the corpus contains is excluded from the catalogue (§3.1).

Why. An instrument that claims little is useful precisely where there is no literature. But to know what it is worth, one must apply it where there is — and the only way for the test to mean anything is not to look at the answer while taking it. In the other four families we did not look because there was nothing to look at; here we do not look although there is, and that is what turns the exercise into a test.

The comparison is deferred, deliberately. What this paper publishes is the catalogue and its measures. Contrasting it with Indo-European historical linguistics is a different, later piece of work, and anyone can do it with these files in front of them. We do not do it and we do not do it now.

A warning to the reader who knows the family. Several tables in §5 — §5.1 in particular — show alternations between segments concentrated at specific cuts. Anyone who knows Indo-European historical phonology will name them immediately. We do not name them, and that is the design and not an oversight. What the table says is where an alternation concentrates in this catalogue; what that concentration means is exactly the question we leave open.

1.2 What this work does not attempt

We do not reconstruct protoforms. We propose no ancestral form, not even implicitly.

We do not claim relatedness. Two words sharing a code does not mean they descend from a common one.

We do not distinguish inheritance from borrowing. Without protoforms this is impossible. A group of words sharing a code may be inheritance, borrowing, convergence or coincidence. The catalogue does not choose — and in this family, with two millennia of documented contact among its branches, that limit weighs more than in any other (§6).

We do not attribute meaning to a code. We write `n·m|name` and never `n·m = name`.

And we do not compare Indo-European with other families: the comparisons in §5.6 are about the instrument's behaviour, not about the languages.

Where this procedure sits in the automatic-comparison literature — Dolgopolsky, Turchin/Peiros/Gell-Mann, ASJP — is covered in the method paper.

2. What a code is, and what it costs here

2.1 The consonant skeleton

The skeleton is a word's consonant sequence, with vowels removed — not lost: they are kept in a separate column — and each consonant written with a canonical representative. Indo-European uses 35 canonical symbols, 31 of which merge something: it is the family with the largest canonical

inventory of the series, and also the one that merges most. The full table is in `runs/normalizacion.md` and is derived by aligning each form with its skeleton, not written by hand.

The 35 symbols, ordered by frequency:

```
s   r   n   t   k   l   d   m   p   b   g   v   j   f   z   x   h   w
c   ð   ʋ   θ   β   ʈ   ɋ   ɖ   ɥ   pf  ɸ   ʍ   ɢ   ɯ   tθ  z’  b’
```

Two mergers must be kept in mind when reading §5, because they erase the most:

- **`s`** gathers every sibilant and affricate — `ʃ`, `ʒ`, `ts`, `tʃ`, `dʒ`, `ɕ`, `ʂ` and a dozen more, over 149,209 positions. It is the inventory's most frequent symbol and the one that merges most.
- **`n`** gathers every nasal but the labial — `ŋ`, `ɲ`, `nʲ`, `ɳ` and variants.

Any alternation that depends on distinguishing within those two groups is invisible to this instrument, and it is worth saying so before the reader wonders why it does not appear.

2.2 A code is a skeleton that recurs across languages

A code is a skeleton appearing, for the same concept, in three or more distinct languages. It is written `n·m|name`, never `n·m = name`: the code does not mean ”name”, it is the shape ”name” repeatedly takes.

2.3 What the instrument costs: collision

52.5 % of Indo-European forms share a skeleton, within their own language, with a form of another concept. This is the clean measure of what the skeleton fails to distinguish. In the series it sits in the middle: below Turkic and Uralic (59 %), above Austronesian (46 %) and well above Pama-Nyungan (34 %).

That it is 52.5 % at an arity of 3.35 — the highest of the series, tied — says something about the inventory: the sibilant merger of §2.1 takes back much of what word length should have gained.

The appendix's reverse index shows it. `m·r` appears with ”sea”, ”die”, ”snake” and ”ant”; `m·s` with ”meat”, ”fly”, ”man” and ”mouse”; `s·t·r` with ”star”, ”four” and ”old”. Four unrelated concepts under two consonants is what the instrument costs, and one must carry that.

2.4 Arity

Arity	Codes	
1 (unary)	705	7.0 %
2 (binary)	3,645	36.1 %
3 (ternary)	3,800	37.6 %
4 or more	1,947	19.3 %

The figures this paper defends are restricted to the 7,454 binary and ternary codes.

This matters here more than in any preceding paper, because four of the fifteen widest-reaching codes are unary: `d|two` with 79 languages, `n|one` with 71, `t|THOU` with 53, `n|NOT` with 43. A single consonant is evidence of nothing — it is the kind of coincidence a small inventory produces on its own — and so they are left out. They are listed in §4.2 marked, not hidden.

3. Method

3.1 Data

Integrative Corpus, Indo-European subset: 207 languages, 319,149 forms with transcription, assigned concept and skeleton. It is the largest family of the series by concepts (2,778) and the second by forms.

Six lects excluded, all declared in `ci/decisiones.py`, on two criteria:

1. One reconstructed lect, `Proto-Italic`. A study done without reconstruction cannot have a reconstruction as a datum. The rule is declared even though that entry would have arrived empty anyway: excluded by construction, not by luck.
2. Five duplicate records of the same doculect — same glottocode and high overlap of written forms, that is, the corpus holds the same lect twice from different sources. The record with more forms is kept: `Old Norse` over `Old Icelandic` (96 %), `Lower Sorbian` over `Sorbian: Lower` (87 %), `Ossetic` over `Ossetic: Iron` (70 %), `Breton` over `Breton_ST` (61 %), `Bihari` over `Maithili` (61 %).

And three exclusions phase 0 proposed and that are NOT applied, because what one declines to discard should also be stated:

- Nobody is excluded for closeness. The 80 % threshold set in the Turkic paper is crossed by no pair of distinct glottocodes: the highest is `Greek: Ancient` / `Greek: New Testament` at 79 %, followed by the two Norwegian standards at 73 %. Lowering the threshold here would mean two criteria across families.
- Nobody is excluded for tone marks. Phase 0 flagged Slovene (90 % of its forms), Lithuanian (14 %), Swedish (3 %) and English (0 %). These are pitch-accent marks, not lexical tone, and above all they do not reach the skeleton: it was checked that in the whole family there are *zero* skeletons with a digit or suprasegmental mark — `gla˩˥d` gives `g·l·d`. The check was calibrated for tonal families and over-warns here.
- Nobody is excluded for being far away. Five languages sit more than 8,000 km from the family's centre (Bislama, Tok Pisin, Jamaican Creole, Saramaccan, Negerhollands). Excluding them would require a judgement about their mode of formation, which is exactly the kind of judgement this work does not make. They are kept, and §5.5 shows the geographic analysis with and without them.

A quarter of the forms do not reach the catalogue, and this must be said loudly. Of 428,150 forms with a concept, 322,827 obtain a skeleton and 105,323 do not (24.6 %). The cause is not the quality of the words but an incomplete segmentation pass: they are perfectly legible forms — `topur`, `caill`, `bēl` in Old Irish — whose segment field is empty.

And it is not a distributed loss: it is one source. Counting by each form's actual source — the prefix of its own identifier, not the label on the language, which is where an earlier version of this table came from and was the wrong column:

Source	Forms with a concept	Without a skeleton	
`ids`	85,559	85,559	100 %
`kaikki`	147,921	18,353	12.4 %
`iec`	25,731	1,036	4.0 %
`lexibank`	128,289	310	0.2 %

Source	Forms with a concept	Without a skeleton	
nel	40,650	65	0.2 %

IDS does not lose much: it produces nothing. Not one of its 85,559 forms with an English gloss has a skeleton row — not empty, absent — and those 85,559 are 81 % of the family's entire loss. The other four sources together lose 5.7 % of theirs.

That rewrites the effect on coverage, and for the better: the languages that disappear entirely are the ones that exist only in IDS. Sanskrit (2,684 forms) and Middle English (1,995) are there and nowhere else, which is why they are lost completely. Old High German loses its 4,945 IDS forms but keeps 1,423 of 1,426 from lexibank; Old Irish loses 1,353 and keeps its 172 from iec. Hittite is left with 19 of 1,400 and Mycenaean Greek with 4 of 136, the latter through kaikki rather than IDS.

That is: the bias towards the ancient is real, has a localised cause, and is a property neither of the method nor of the lexicon.

And it is a declared decision, not an outstanding fault. IDS ships its forms in orthography and its distribution includes no profiles for converting them to IPA — the CLDF standard for that, etc/orthography/, one per language — so segmenting them would mean writing 274 profiles. Doing that badly does not lose the form: it manufactures a false skeleton that enters the catalogue looking like data. The decision (2026-09-04) was not to work those languages; it is written with its figures in docs/DECISION-IDS.md, declared in ci/decisiones.py, and reported on every run of the QA suite. This catalogue describes worst those languages whose only testimony is IDS, and §6 returns to it.

Seven macroareas, by colonial transplant (§5.5).

3.2 Enrichment

The code's coverage divided by its base frequency; the full definition, and the warning that it is not comparable across families, are in the method paper.

3.3 Nulls

Every grouping is compared against a permutation shuffling concepts within each language. Real grouping covers 43.9 % of forms; the null, 11.8 %.

The chance rate for precision (§4.1) is not sampled: it is computed exactly, in closed form over the distribution of concepts and languages. Earlier papers estimated it by drawing and published intervals that measured their own sampling; the correction is documented in docs/ARQUITECTURA.md.

3.4 One population for the whole paper

The seven reports in runs/ that declare it all print the same line:

```
POPULATION · Indo-European: 207 languages · 319,149 forms ·
             2,778 concepts · 12 branches · segmentation=yes(0) ·
             excluded=6 · impossible=34 · threshold=3 languages
```

This is checked mechanically (analysis/tools/verifica_paper.py, check 7).

4. Results

4.1 The catalogue

10,116 codes over 1,946 of the 2,778 concepts, grouping 94,152 forms.

Quantity	Value	Over which population
Forms inside a code	29.5 % (94,152 of 319,149)	3-language threshold, the catalogue's
Forms in any group	43.9 % (null: 11.8 %)	2-language threshold, grouping test
Precision	8.73 % (chance: 0.084 %)	pairs of forms from different languages
Times over chance	104× [104–105; chance is computed exactly]	same
Binary and ternary	7,454 (73.7 %)	of the 10,116 codes

Each row carries its population because there are three distinct ones.

The 104× is the second of the series, and §5.6 insists on what the Turkic paper established: it is a quotient, and its denominator depends on the family. Indo-European's chance rate (0.084 %) is the lowest of the five, and it is so for a reason of corpus rather than of languages — 2,778 concepts spread across the forms mean two random forms very rarely coincide in concept. With more concepts the denominator falls and the quotient rises without anything having improved.

Figure 1: Distribution of codes by arity.

4.2 The widest-reaching codes

Concept	Code	Languages	Enrichment	Branches
name	`n·m`	103	379×	6
three	`t·r`	81	130×	10
two	*d*	*79*	*128*×	*9*
nose	`n·s`	79	150×	6
one	*n*	*71*	*77*×	*8*
sea	`m·r`	62	124×	6
THOU	*t*	*53*	*220*×	*9*
star	`s·t·r`	52	120×	7
salt	`s·l`	48	62×	6
feather	`p·r`	47	90×	3
meat	`m·s`	46	72×	7

The rows in italics are unary and count towards no figure in this paper (§2.4): a single-consonant code gathers languages by coincidence far too easily. They are shown because hiding them would be worse, and because their presence at the head of the list is itself a datum about the instrument.

`n·m|name`, with 103 languages, is the widest-reaching code in the whole worked corpus — ahead of Pama-Nyungan's `m·r|hand` (69) and of any Austronesian code.

The branch column counts how many branches each code appears in; it does not claim that distribution has a historical cause, which is a different question and not this method's.

Figure 2: The twenty widest-reaching binary and ternary codes.

5. What this family teaches

5.1 Sixty-one branch pairs

The Turkic paper established that the substitution matrix must be computed per branch pair, since aggregating over the family dilutes what distinguishes the groups. Turkic could test this on one cut, Uralic on thirty-six, Pama-Nyungan on ninety-six. Here there are sixty-one, and they carry the largest samples of the whole series.

One cut, Germanic ↔ Romance:

Segment	Main	Cases	Share when crossing	Share when not	Diagnostic
h	k	5,799 of 12,343	47 %	8 %	6.2×
k	h	5,799 of 22,418	26 %	3 %	7.8×
θ	t	1,050 of 2,597	40 %	20 %	2.0×
f	p	4,104 of 12,001	34 %	15 %	2.3×
p	f	4,104 of 13,160	31 %	13 %	2.5×
m	l	5,627 of 15,580	36 %	6 %	5.9×

And another, Indo-Iranian ↔ Romance:

Segment	Main	Cases	Share when crossing	Share when not	Diagnostic
θ	s	586 of 877	67 %	29 %	2.3×
l	r	8,747 of 21,622	40 %	17 %	2.4×
ʈ	s	921 of 2,733	34 %	8 %	4.3×
q	k	194 of 484	40 %	20 %	2.0×

The last column compares how much more concentrated an alternation is when crossing the cut than within a branch.

What it is and what it is not. It is a measure of where an alternation concentrates, computed on the catalogue. It is not a sound correspondence: a correspondence requires environment, direction and regularity, and none of the three is present here. And — §1.1 — we do not name them. Anyone who knows this family's historical phonology will see in these two tables things we do not say; saying them would be consulting the answer, which is exactly what this work has committed not to do. The files are published.

Three caveats. The tables are two cuts of sixty-one and are trimmed to the highest-share rows; the rest is in `runs/fase6-por-rama.md`. The "cases" are pairs of forms, not independent observations: the figure points at where to look, it is not a test. And §2.1 recalls that any alternation internal to the sibilants or the nasals is invisible here by construction.

5.2 Sibling codes

The search for pairs of codes for the same concept differing in a single position (`analysis/fases/codigos_hermanos.py`) yields 3,647 binary and ternary pairs here, by far the highest count of the series.

	Indo-European	Pama-Nyungan	Uralic	Turkic
Pairs	3,647	653	153	357
Overlap in language	827 (23 %)	35 (5 %)	39 (25 %)	140 (39 %)

77 % are disjoint: no language appears in both codes, so each pair splits the languages into two groups without exception. Of the 827 that overlap, the classifier separates 425 as "another form" — the language has two words for the concept, or the same root with different inflection, and neither is an error — and 402 as a transcription artifact, two sources writing the same thing two ways.

The largest case is `name`: `n·m` with 103 languages against `n·n` with 10 — and the ten are mostly from a different branch from the hundred and three. The catalogue puts both halves on the table, separated, without knowing that is what they are.

5.3 Automatic segmentation makes things worse, for the fourth time

Step	In a group	Null	Advantage	κ	Codes of 4+ consonants
Whole word	43.9 %	11.8 %	32.1	0.364	38.8 %
+ reduplication	44.4 %	—	—	—	38.0 %
+ discovered affixes	58.9 %	43.7 %	15.2	0.270	7.7 %

The κ column is the Hubert-Arabie correction, `(observed − null) /(1 − null)`: it normalises by the attainable ceiling and so allows comparison across families where an advantage in points does not — ten points over a 5% null and ten over a 44% null are not worth the same. Here $\Delta\kappa = -0.094$. All eight families in the series give the same sign: 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, Austronesian −0.151.

(This list had six families and the figures of the time. It was rebuilt with all eight on 2026-09-09 from `pipeline/tools/tabla_serie.py`. The magnitude does NOT grow with the raw gain, which is what this paragraph used to say: Sino-Tibetan, with the shortest word of the eight, falls less than Indo-European.)

Fifteen points of signal, thirty-two of noise. The advantage over chance falls to less than half.

This is the fourth consecutive family where it happens — Turkic, Uralic, Pama-Nyungan and this one — and the four have very different morphologies. By now the finding belongs to no family: discovering affixes by shape similarity inflates the null more than the signal, because the same trimming that brings related words together brings unrelated ones together too.

Here the discoverer finds 2,029 prefixes and 1,778 suffixes across 198 languages. That the prefix count exceeds the suffix count in a largely suffixing family says again what it said in Pama-Nyungan: the discoverer is finding the consonant inventory, not the morphology.

The published catalogue uses no automatic segmentation.

5.4 The substitution closure produces nothing, and that makes three

Transitively closing the substitutions that survive the validation half gave three clean classes in Austronesian and three in Pama-Nyungan. In Turkic and Uralic it gave none. Here it gives none either: at a $\geq 2\times$ threshold no publishable group remains, and this appendix is published without an emergent-code column.

With five families measured, the tally is three to two, and the candidate explanation is the one in §5.4 of the Pama-Nyungan paper: there the closure worked because the normalisation preserved the dimension on which that family's inventory is organised — place of articulation, with five distinct coronal series. Here §2.1 does the opposite: it merges every sibilant into `s` and nearly every nasal into `n`, which are precisely the most populated dimensions of the Indo-European inventory. Where the normalisation erases the dimension the family uses, the closure finds nothing.

That is not an excuse: it is a testable prediction. A normalisation preserving the sibilants ought to make classes appear here, and might worsen other figures. We have not tried it.

5.5 Geography, with transplanted languages inside

Shared codes	Pairs	Mean distance	÷ chance
1	4,680	3,649 km	1.09
2	2,823	3,433 km	1.02
3–5	3,024	3,384 km	1.01
6–10	1,293	3,157 km	0.94
11–20	1,041	2,497 km	0.74
21+	1,893	1,637 km	0.49
two random languages	40,000	3,355 km	1.00

Monotonic across all six buckets, and none falls below a thousand pairs.

And here is a check the other families did not need. This family has languages twelve thousand kilometres from its centre by colonial transplant — a fact about ships, not about diversification — and those languages share many codes with the ones they came from while sitting very far away, which ought to flatten the gradient. Of the catalogue's 207 languages, 204 have coordinates and 198 of those are in Eurasia; repeating the computation with only those 198:

Shared codes	÷ chance (all)	÷ chance (Eurasia only)
1	1.09	1.12
2	1.02	1.05
3–5	1.01	1.00
6–10	0.94	0.86
11–20	0.74	0.64
21+	0.49	0.42

The gradient becomes steeper when they are removed, as expected, but does not change shape. That is: the transplanted languages flatten the result without producing it. The table with all languages is the one published, because excluding them would require a judgement about their mode of formation that this work does not make.

5.6 Five families

With the five measured by the same code and the same definition (`ci/metricas.py`), and with chance computed exactly:

Family	Lgs	Arity	Collision	Precision	Chance	× chance	Concepts
Turkic	32	2.85	59 %	25.52 %	0.118 %	216×	1,828
Indo-European	207	3.35	52 %	8.73 %	0.084 %	104×	2,778
Uralic	27	3.27	59 %	6.50 %	0.083 %	78×	1,843
Austronesian	582	2.78	46 %	13.45 %	0.363 %	37×	1,823
Pama-Nyungan	154	3.35	34 %	5.87 %	0.298 %	20×	1,160

The concepts column is added because with five families one can now see what two could not: chance falls when the corpus has more concepts, and the quotient rises for that reason. Indo-European and Uralic have almost identical chance rates (0.084 % and 0.083 %) with precisions differing by 1.3×, and the quotient separates them into 104× against 78×. Ranking families by that column remains a mistake, and now it is easier to see why.

Arity predicts collision, with an exception that explains itself. Austronesian 2.78 → 46 %, Turkic 2.85 → 59 %, Uralic 3.27 → 59 %, Pama-Nyungan 3.35 → 34 %. Indo-European has the highest arity, tied, and a collision of 52 %, well above Pama-Nyungan at the same word length. The difference is §2.1: here the normalisation merges sibilants and nasals, there it preserved five coronal series. Word length gives room; the normalisation decides how much of that room is used.

6. Limits

The main limit is the corpus, not the method — and it has a single culprit. A quarter of the forms do not reach the catalogue, and 81 % of that loss is one source: IDS contributes 85,559 forms with an English gloss and produces not one skeleton (§3.1). The other four sources together lose 5.7 % of theirs.

That hits the ancient languages because they are the ones most dependent on that source: Sanskrit and Middle English exist only in IDS and are lost entirely; Old High German and Old Irish lose their IDS share and keep the rest. In a family whose interest lies largely in its time depth, that is not a detail. This catalogue describes worst those languages whose only testimony is IDS, no figure in §4 or §5 corrects that, and it is repairable: it is not a property of the method but a missing ingestion pass.

Collision is 52.5 %, and §2.1 explains why it is high despite the arity: the sibilant merger.

Inheritance and borrowing are not distinguished, and here that weighs double. These languages have been in documented contact with one another and with outside languages for two millennia, and they also share a common learned layer. A high-reaching code may be any of those things.

There are no zero reflexes. The §5.1 procedure compares skeletons of equal length, so it cannot see any alternation involving the loss of a segment — and in this family that is a central phenomenon.

19.3 % of codes have four or more consonants, and §5.3 shows automatic segmentation is not the way out. Morphological segmentation remains unresolved.

Unary codes head the reach list. Four of the fifteen widest-reaching codes have a single consonant. They are outside the defended figures, but their presence says that in a large family the instrument produces coincidence easily.

We do not calibrate against expert cognacy, and here that is a decision rather than a gap. The `iecor` source carries it, in the same corpus, unread. Calibrating against it is exactly the work §1.1 defers — but whoever wants to do it has both files.

A single family, again. Everything this paper says about what the figures mean rests on five points.

7. Conclusion

The Indo-European catalogue groups 29.5 % of forms into 10,116 codes over 1,946 concepts, with a precision 104 times chance. It is the widest catalogue of the series in concepts and holds the widest-reaching code of the whole worked corpus, `n·m|name` in 103 languages.

What this work contributes is not that catalogue, and it is not a result about Indo-European either. It is a catalogue built without looking at the answer, in the one family where the answer is written down. Everything in these pages can now be contrasted with two centuries of the comparative method, and that contrast will say something none of the four preceding families could: not whether the instrument is right — that we already know how to measure — but what kind of thing it finds and what kind of thing escapes it, against an independent yardstick.

Meanwhile, two results that do not depend on that comparison. Automatic segmentation worsens the advantage over chance for the fourth consecutive time, across four different morphologies; it is by now a property of the procedure. And the substitution closure produces nothing for the third time out of five, with a candidate explanation the catalogue itself suggests and which is testable: it works where the normalisation preserves the dimension on which the family's inventory is organised, and here it erases it.

An instrument that claims little can be applied where there is no literature. Applying it where there is, and not reading it, is the only way to find out how much of what it fails to find is its own fault.

References

Brown, C. H., Holman, E. W., Wichmann, S. and Velupillai, V. (2008). Automated classification of the world's languages. *Sprachtypologie und Universalienforschung* 61.

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

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

Heggarty, P., Anderson, C., Scarborough, M. *et al.* (2023). Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. *Science* 381. [Source `iecor` of the forms; its cognacy judgements have not been used — §1.1.]

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

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.

List, J.-M., Cysouw, M. and Forkel, R. (2016). Concepticon: a resource for the linking of concept lists. *LREC*.

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

List, J.-M., Greenhill, S. J. and Gray, R. D. (2017). The potential of automatic word comparison for historical linguistics. *PLOS ONE* 12(1).

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.

Data and reproducibility

Catalogue	`families/indo-european/output/catalogo-indo-european.tsv` — 10,116 rows
Per-form detail	`families/indo-european/output/catalogo-indo-european-lenguas.tsv` — 65,503 rows
Hygiene (§3.1)	`analysis/fases/fase0_higiene.py` → `runs/fase0-higiene.md`
Declared exclusions (§3.1)	`ci/decisiones.py`
Loading, population and warnings	`ci/familia.py`
Headline figures (§4.1, §5.6)	`ci/metricas.py` — exact chance
Construction	`analysis/fases/catalogo_familia.py`
Grouping and nulls (§3.3, §4.1)	`analysis/fases/agrupamiento_familia.py` → `runs/agrupamiento.md`
Segmentation (§5.3)	`analysis/fases/fase2_despiece.py` → `runs/fase2-despiece.md`
Substitution closure (§5.4)	`analysis/fases/fase6_clases_emergentes.py` → `runs/fase6-clases.md`
Substitutions per branch pair (§5.1)	`analysis/fases/fase6_por_rama.py` → `runs/fase6-por-rama.md`
Sibling codes (§5.2)	`analysis/fases/codigos_hermanos.py` → `runs/codigos-hermanos.md`
Codes and geography (§5.5)	`analysis/fases/codigos_vs_parentesco.py` → `runs/codigos-parentesco.md`
Normalisation table (§2.1)	`analysis/fases/tabla_normalizacion.py` → `runs/normalizacion.md`
Appendix A	`analysis/fases/anexo_codigos.py` + `analysis/fases/empalma_anexo.py`
Cross-family mixing audit	`analysis/tools/audita_familias.py`

Data sources: Lexibank (List, Forkel, Greenhill et al. 2022, CC-BY-4.0), IDS (Key & Comrie, CC-BY-4.0), IE-CoR (Heggarty et al. 2023, CC-BY-4.0 — forms only), Wiktionary/kaikki (CC-BY-SA-3.0), Concepticon and Glottolog (CC-BY-4.0).

Appendix A · One hundred annotated codes

The catalogue has 10,116 codes and the complete data is in `families/indo-european/output/`. What follows is a readable sample: the hundred widest-reaching binary and ternary codes, each with up to sixteen languages and their words.

Two things about the presentation. Languages are spread along the alphabetical list, not taken from its head; long names are truncated and forms are not altered. And the thematic grouping is a reading order declared by hand, not a semantic classification.

This appendix carries no emergent-code column. The transitive closure of substitutions produces no clean classes in this family (§5.4), and transplanting the Austronesian or Pama-Nyungan ones would be convenient and false: a class's value is relative to the family that produced it.

And what governs the whole reading: this appendix describes what is in each block — how many languages, how they are distributed, what varies and what does not — and says nothing about what any form descends from, nor names any alternation. That is not rhetorical caution: it is the design decision of §1.1. Anyone who knows the family will see things here that we do not say, and the files are published so that whoever wants to can say them.

One warning to carry throughout: 52.5 % of forms share a skeleton, within their own language, with a form of another concept. A code may gather the same word used for two things, or two words the instrument does not distinguish, and one must look at the forms to know which.

Reverse index: codes recurring across several concepts

This table is longer than in any preceding family, and the cause is not arity — 3.35 consonants per form, the highest — but the normalisation: §2.1 merges every sibilant into `s` and nearly every nasal into `n`, the two most populated zones of the inventory. Word length gives room and the normalisation spends it.

Three cases worth looking at slowly:

`m·r` appears with "sea" (62 languages), "die" (38), "snake" (22) and "ant" (21). `m·s`, with "meat" (46), "fly" (22), "man" (21) and "mouse" (21). And `s·t·r`, with "star" (52), "four" (31) and "old" (22). Four apparently unrelated concepts under two consonants is exactly the cost §2.3 declares, and in this family it is paid four times in the table's first screenful.

Code	Concepts it appears with, and their enrichment
`m·r`	sea (62, 124×) · die (38, 74×) · snake (22, 42×) · ant (21, 46×)
`m·s`	meat (46, 72×) · fly_n (22, 40×) · mouse (21, 147×) · man (21, 32×)
`s·t·r`	star (52, 120×) · four (31, 75×) · old (22, 55×)
`s·l`	salt (48, 61×) · sun (32, 40×) · year (27, 33×)
`k·r·n`	horn (33, 110×) · root (22, 73×) · meat (20, 66×)
`m·n`	man (31, 70×) · moon (30, 68×) · hand (26, 58×)
`s·n`	woman (28, 29×) · snow (23, 27×) · chest (20, 23×)

Numerals

The sharpest block of the appendix, and the one that best shows what a code does and does not do.

"Three" has two codes — `t·r` with 81 languages and `t·r·s` with 22 — "four" has three — `s·t·r` (31), `s·r` (27), `k·t·r` (20) — and "five" has two, `p·n·s` (33) and `p·s` (29). In all three cases the codes for one

concept differ in a single position or in one consonant present or absent, which is the raw material of §5.2.

And in all three, the forms within each code resemble one another closely and those of the neighbouring code little. The catalogue separates; what it separates, it does not say.

Note also what is not here: ”one” and ”two” head the reach list of the whole catalogue, with 71 and 79 languages, and do not appear in this block because their codes are unary — `n` and `d` — and fall outside by §2.4. It is the best available reminder of why that rule exists.

`t·r|three` · binary · 81 languages · enrichment 129×

```
Albanian tre        Belarusian try        Bulgarian tri          Elfdalian tri
Gaelic: Manx three  Greek: Italiot treí   Italian tre            Khowar troi
Macedonian tri      Middle Welsh tri      Norwegian: Bokmål tre  Old Polish trzy
Polabian tåri       Saṇu-viri: Wâmâ tr´a  Slovene: Early Mo tri  Tocharian B trai
```

`p·n·s|five` · ternary · 33 languages · enrichment 179×

```
Avestan: Younger paṇca  Balochi: Sistani panǰ  Gawarbati pants
Judeo-Tat penž          Khwarazmian panc       Kurdish C.: Jafi panj
Kurdish S.: Qorve panj  Mazanderani panj       Northern Kurdish pɛnt͡ʃ
Pali pañca              Parthian panž          Persian panǧ
Polish pɪɛnt͡ɕ           Sarikoli pindz         T-Agalyk panʒ
Vedic: Early páñca
```

`s·t·r|four` · ternary · 31 languages · enrichment 75×

```
Belarusian čatyry    Bulgarian t͡ʃɛtiri       Czech t͡ʃtɪr̥ɪ          Latgalian četri
Lower Sorbian styri  Macedonian: Suho četiri  Old Czech čtyři       Old Polish cztery
Polabian citĕr       Romani štar              Rusyn čotwɍɪ          Serbo-Croat četiri
Slovak štyri         Slovene: Early Mo štiri  Sorbian: Upper štyri  Tocharian B śtwer
```

`p·s|five` · binary · 29 languages · enrichment 49×

```
Albanian pesə     Albanian 'pesə         Albanian: Gheg pesë  Assamese pām̐s
Bengali pãt͡ʃɔ     Bihari pańc            Dras põš             Kalaṣa-alâ: Nišei pŭč
Kashmiri paantsh  Kâta-vari: Ktivi p´uč  Lower Sorbian pěś    Marathi pāca
Old Polish pięć   Russian pât'           Sorbian: Upper pjeć  Tocharian B piś
```

`s·r|four` · binary · 27 languages · enrichment 38×

```
Assamese sāri  Bakhtiari čār          Bengali cār           Bihari cār
Delvari t͡ʃɒr   Gawarbati tsur        Hindi cāra            Judeo-Tat čor
Khowar čoor    Kurdish C.: Jafi čʷār  Kurdish N.: Bahdi čar Marathi cāra
Nepali cār     Northern Kurdish t͡ʃar  Pashai: North-Wes čor  T-Agalyk ʧɔr
```

`t·r·s|three` · ternary · 22 languages · enrichment 112×

```
Ancient Greek 'treis    Anglo-Norman treis    Catalan trɛs
Greek: Ancient treĩs    Greek: Cypriot treis  Greek: New Testam treĩs
Greek_Mod tris          Latgalian treis       Latvian trīs
Lithuanian trĭːs        Old Catalan tres      Old Occitan tres
Old Spanish tres        Oscan trís            Portuguese: Brazi três
Sardinian: Logudo tres
```

k·t·r|four · ternary · 20 languages · enrichment 210×

Albanian katər	Albanian 'katər	Albanian: Arbëres katër
Albanian: Gheg katër	Catalan quatre	Dalmatian: Veglio kụatri
French katr	Friulian cuatri	Italian quattro
Ladin cater	Latin kwat:uor	Lithuanian keturi
Old Catalan quatre	Old Spanish cuatro	Portuguese kuatɾu
Spanish cuatro		

Body parts

The largest block, headed by n·s|nose (79 languages).

Two concepts have two codes each, and the two cases differ from one another. ”Hand” splits between m·n (26 languages) and r·k (24) — two codes sharing not one consonant, so they are not siblings in the sense of §5.2 but two forms without resemblance. ”Knee” splits between k·n (25) and k·l·n (23) — which share both consonants and differ by one inserted in the middle.

p·p·k|navel, at 690×, is the block’s highest enrichment and its twenty-one languages are all from a single branch: Belarusian, Croatian, Czech, Kashubian, Macedonian, Old Church Slavonic, Old Polish. A high-reaching code may be confined to one group; reach does not say where.

d·n·t|tooth (521×) and j·z·k|tongue (538×) are the section’s two highest-enrichment ternaries, and in both the block’s forms resemble one another obviously at a glance.

n·s|nose · binary · 79 languages · enrichment 149×

Anglo-Norman nes	Catalan nas	Dutch_List nøs	Frisian noas
Istro_Romanian nɒs	Khwarazmian nāc	Lower Sorbian nos	Megleno_Romanian nas
Norwegian ne:sə	Old Czech nos	Old Polish nos	Pashai: North-Wes nasə
Russian nos	Serbo-Croatian nos	Sogdian nas	Ukrainian nìs

d·n·t|tooth · ternary · 35 languages · enrichment 521×

Anglo-Norman dent	Breton: Treger dant	Catalan dent
Ferrarese_Emilian dent	Gawarbati dant	Istro_Romanian d'int+e
Lanzo_Torinese_Pi d'ɛŋt	Megleno-Romanian dínti	Middle Breton dant
Old Breton dant	Old Occitan dent	Pali danta
Portuguese: Brazi dente	Rapallo_Ligurian d'ɛnt+e	Romanian dinte
Venice_Venetian d'eŋt+e		

s·r|head · binary · 34 languages · enrichment 44×

Avestan sāra-	Bakhtiari sar	Bhojpuri sir	Hawrami sara
Judeo-Tat ser	Kumzari sar	Kurdish N.: Bahdi ser	Middle Persian sar
Ossetian sər	Ossetic: Digor sær	Parthian sar	Persian sar
Punjabi sir	Selice Romani šéro	T-Agalyk sar	Wakhi sar

k·r·n|horn · ternary · 33 languages · enrichment 110×

Aromanian k'orn+u	Breton: Gwened korn	Catalan corn
English -corn	Franco-Provençal kourna	Greek κόρνα
Italian corno	Lanzo_Torinese_Pi k'ɔrn	Megleno_Romanian korn
Middle Cornish corn	Milanese còrnu	Old Spanish cuerno
Palermitan_Sicili k'ɔɾn+u	Rapallo_Ligurian k'ɔrn+u	Romanian corn
Venice_Venetian k'ɔrn+o		

k·r|heart · binary · 27 languages · enrichment 46×

Ancient Greek 'kear	Campidanese kˈɔr+u	Dalmatian: Veglio kur
French cœur	Friulian cûr	Gaelic: Manx cree
Irish croí	Italian cuore	Lanzo_Torinese_Pi kœr
Milanese cör	Neapolitan core	Palermitan_Sicili kˈɔɾ+i
Ravennate_Romagno kˈɔar	Sardinian: Logudo coro	Spanish cor
Venice_Venetian kˈɔr		

m·n|hand · binary · 26 languages · enrichment 58×

Anglo-Norman main	Aromanian mˈɨn+ə	Dalmatian: Veglio muṅ
Ferrarese_Emilian man	Italian mano	Lanzo_Torinese_Pi maŋ
Latin manu	Milanese màn	Old French main
Old Occitan man	Palermitan_Sicili mˈan+u	Provençal_Occitan maŋ
Ravennate_Romagno mˈãn	Saramaccan máun	Sardinian: Logudo manu
Spanish mano		

k·n|knee · binary · 25 languages · enrichment 52×

Danish knɛ	Dutch knie	Elfdalian kni	Faroese knæ
German Knie	Icelandic kʰ.n.iɛː	Middle Dutch cnie	Middle High Germa knie
Norwegian kneː	Norwegian: Nynors kne	Old English cnēo	Old High German kneo
Old Norse kné	Old Swedish knä	Saramaccan kiní	Tocharian B keni

b·n|bone · binary · 24 languages · enrichment 72×

Bislama bun	Danish beˀn	Elfdalian bien	English bon
Flemish been	Icelandic bein	Middle Dutch been	Middle High Germa bein
Norwegian beːn	Norwegian: Bokmål ben	Old English bān	Old Frisian bēn
Old Norse bein	Old Saxon bên	Saramaccan bónu	Swedish beːn

r·k|hand · binary · 24 languages · enrichment 70×

Belarusian ruka	Bulgarian rəka	Czech rʊka	Latgalian rūka
Lower Sorbian ruka	Macedonian raka	Old Church Slavon rąka	Old Czech ruka
Old Polish ręka	Polish rɛ̃ka	Rusyn ruká	Serbo-Croat ruka
Slovak ruka	Slovene roka	Slovene: Kostel roka	Sorbian: Upper ruka

k·l·n|knee · ternary · 23 languages · enrichment 109×

Belarusian kalena	Bulgarian kolʲano	Croatian kɔ̌ʌɛnɔ
Kashubian kòlano	Lower Sorbian kóleno	Macedonian: Suho koleno
Old Church Slavon kolěno	Old Novgorod kolěno	Old Polish kolano
Polish kɔlanɔ	Rusyn kǫlʹíno	Serbo-Croatian koleno
Slovene koleno	Slovene: Early Mo koleno	Sorbian: Upper koleno
T-Agalyk kaleno		

k·s|skin · binary · 23 languages · enrichment 38×

Armenian: Eastern kaši	Armenian_Mod kɑʃi	Bulgarian kɔʒə
Czech kuːʒɛ	Lower Sorbian kóža	Macedonian: Suho koža
Megleno_Romanian kˌˈo̯a.ʒ.+.ə	Old Church Slavon koža	Old Czech kóžě
Old Novgorod koža	Rusyn kóža	Serbo-Croat koža
Slovak koža	Slovene koža	Slovene: Kostel koža
Sorbian: Upper koža		

j·z·k|tongue · ternary · 23 languages · enrichment 538×

Belarusian jazɨk	Croatian jezik	Czech jazɪk	Lower Sorbian jězyk
Macedonian ǰazik	Old Church Slavon językŭ	Old Czech jazyk	Old Polish język
Polabian jǫzĕk	Polish jɛ̃zɨk	Rusyn jazźk	Serbo-Croat jezik
Slovak jazik	Slovene jezik	Slovene: Kostel jezik	Sorbian: Upper jazyk

h·r·n|horn · ternary · 22 languages · enrichment 278×

Danish hoʁˀn	Dutch hoorn	Dutch_List horn	Flemish hoorn
German horn	German: Bernese Horn	Icelandic h.ɔ.r.t͡n	Middle Dutch hoorn
Negerhollands horən	Norwegian huːrn	Norwegian: Nynors horn	Old Frisian hōrn
Old High German horn	Old Norse horn	Old Swedish horn	Swedish hoːrn

b·k|mouth · binary · 21 languages · enrichment 83×

Breton: Gwened beg	Campidanese b'uk:+a	Catalan bokə
Dalmatian: Veglio buka	French bucco-	Italian bok:a
Lanzo_Torinese_Pi b'uk+a	Milanese bùca	Old Catalan boca
Old Spanish boca	Provençal_Occitan b'uk+o	Rapallo_Ligurian b'uk:+a
Ravennate_Romagno b'ok+a	Sardinian: Logudo bucca	Spanish boka
Venice_Venetian b'ok+a		

p·p·k|navel · ternary · 21 languages · enrichment 690×

Belarusian pupɔk	Croatian pupak	Czech pupek
Kashubian pãpk	Macedonian papok	Macedonian: Suho papok
Old Church Slavon pąpŭkŭ	Old Polish pępek, pąpek	Polish pępek, pąpek
Russian pupok	Serbo-Croatian pupak	Slovak pupok
Slovene popek	Slovene: Kostel popek	Sorbian: Upper pupk
Ukrainian pupɔk		

s·n|chest · binary · 20 languages · enrichment 23×

Bakhtiari sine	Delvari sine	Hawrami sīna	Hindi सीना
Kashmiri siinə	Kumzari sīna	Kurdish C.: Jafi seng	Kurdish S.: Qorve sina
Mazanderani sine	Northern Pashto si'na	Palula siiná	Pashai: North-Wes sino
Persian sine	Persian: Tehran sīneh	Raji: Barzoki siṇa	Tati sine

Kinship and person

n·m|name, with 103 languages, is the widest-reaching code in the whole worked corpus — ahead of Pama-Nyungan's m·r|hand, which reached 69.

It deserves a look for its enrichment, 379×, because it explains the mechanism. n and m are nasals, among the family's most frequent consonants, so the base frequency is high. The figure comes not from rarity but from the fact that one hundred and three of the languages attesting "name" say it with the same skeleton: coverage is nearly total. No rare sequence is needed to get there; it is enough for a word to be well distributed in a well-sampled corpus.

"Man" appears with two codes, m·n (31) and m·s (21), and m·s is also the code for "meat", "fly" and "mouse" in the reverse index. m·m|MOTHER gives 916× with twenty languages, the section's highest enrichment, over the least informative consonant sequence possible.

n·m|name · binary · 103 languages · enrichment 379×

Ancient Greek 'onoma	Balochi: Sistani nām	Dalmatian: Veglio nam
Frisian namme	German Name	Greek: Modern Std ónoma
Italian nome	Kumzari nām	Lari nom
Middle Dutch name	Northern Pashto num	Old Irish ainm
Palermitan_Sicili n'ɔm+i	Persian: Tehran nām	Sarikoli num
Tocharian B ñem		

m·n|man · binary · 31 languages · enrichment 70×

Bhojpuri manai	Campidanese 'om:+in+i	Dutch man
English man	Flemish man	German Mann
Jamaican Creole man	Luxembourgish Mann	Middle High Germa man
Norwegian mɑn	Norwegian: Nynors mann	Old High German man
Old Swedish man	Sardinian: Logudo omine	Spanish man
Tok Pisin man		

s·n|woman · binary · 28 languages · enrichment 29×

Avestan ǰani-	Avestan: Younger jaini	Bulgarian žena
Czech žena	Hawrami žanī	Kurdish C.: Jafi žın
Macedonian žena	Northern Kurdish ʒən	Old Czech žena
Old Novgorod žena	Parthian žan	Rusyn žoná
Serbo-Croatian žena	Slovak žena	Slovene: Early Mo žena
Sorbian: Upper žona		

m·s|man · binary · 21 languages · enrichment 32×

Bulgarian mɤʃ	Czech mʊʃ	Dras mu'šɑ
Gawri mīš	Kamviri m′oč	Khowar mooš
Kâta-vari: Ktivi m′uč	Macedonian maž	Macedonian: Suho maž
Old Church Slavon mąžĭ	Old Novgorod mužĭ	Old Polish mąż, mężczy(z)na
Palula míiš	Saṇu-viri: Wâmâ m′âć	Slovak muž
Slovene mɔɻ:ʒ		

m·m|MOTHER · binary · 20 languages · enrichment 925×

Albanian mëmë	Bislama mama	Breton mãm	Dutch mama
Greek_Mod ma'ma	Hindi अम्मा	Lithuanian mama	Lower Sorbian mama
Persian مام	Romanian mamă	Saramaccan mamá	Selice Romani mama
Slovene maɻ:ma	Swedish mamma	Tok Pisin mama	Welsh mam

Nature and sky

The most populated section after body parts, and the one that most splits a single concept across codes: ”night” goes with n·t (38) and n·s (33), ”star” with s·t·r (52) and s·t·l (25), ”smoke” with d·m (35) and f·m (27), ”moon” with l·n (37) and m·n (30), ”sky” with s·m·n (26) and n·b (23).

Five concepts with two codes each in a single block. In ”star” the two codes differ in a single position; in ”moon” and ”smoke” they share no consonant at all. The catalogue does not distinguish those two cases — §5.2 does, which is why it exists.

And here begins what must be read with §5.5 in front of you. p·r·l|APRIL gives 1131× with 23 languages and m·r·s|MARCH gives 645× with 22. They are month names, and they are the two highest

enrichments in the whole appendix. The April block gathers Albanian, Armenian, Bulgarian, Croatian, Dutch and German saying nearly the same thing. An enrichment of a thousand says nothing about a word's age: it says many languages say it the same way, and that can happen by very different routes this instrument does not separate.

m·r|sea · binary · 62 languages · enrichment 124×

```
Anglo-Norman mer       Breton: Gwened mor  Croatian more          Friulian mâr
Gothic marei           Latin mare          Megleno-Romanian mári  Middle Welsh mor
Old Catalan mar        Old French mer      Old Occitan mar        Old Welsh mor
Portuguese: Brazi mar  Rusyn móre          Serbo-Croatian more    Ukrainian more
```

s·t·r|star · ternary · 52 languages · enrichment 120×

```
Ancient Greek a'stēr    Armenian: Eastern astɬ  Avestan: Younger star
Delvari setɒre          Flemish ster            German Star
Greek: Italiot ástro    Greek_Mod 'astro        Judeo-Tat astara
Kurdish N.: Bahdi stēr  Mazanderani setare      Negerhollands stɛre
Old High German sterro  Pashto store            Portuguese astro
T-Agalyk istora
```

l·n|moon · binary · 37 languages · enrichment 87×

```
Anglo-Norman lune         Bulgarian luna           Catalan lluna
English luni-             French lyn               Italian luna
Lanzo_Torinese_Pi l'yŋ+a  Megleno_Romanian l'un+ə  Neapolitan l'un+ə
Old Church Slavon luna    Old Spanish luna         Polabian laună, laină
Rapallo_Ligurian l'yŋ:+a  Sardinian: Logudo luna   Slovene luna
Spanish luna
```

n·t|night · binary · 37 languages · enrichment 127×

```
Albanian nat             Albanian: Gheg natë      Anglo-Norman nuit
Campidanese n'ɔt:+i      English nait             Ferrarese_Emilian nɔ:t
Icelandic n.ouʰ.t        Ladin nuet               Norwegian nɑt
Norwegian: Nynors natt   Old Norse nótt           Palermitan_Sicili n'ɔt:+i
Portuguese: Brazi noite  Sardinian: Logudo notte  Standard Albanian natə
Tok Pisin nait
```

d·m|smoke · binary · 35 languages · enrichment 131×

```
Belarusian dym   Croatian dim    Dras duum            Istro_Romanian dim
Kamviri d´üm     Khotanese dumä  Lithuanian dû:meʲɪ̯ˑ  Old Church Slavon dymŭ
Old Polish dym   Palula dhuumíi  Polabian dåim        Rusyn dɯm
Serbo-Croat dim  Sinhalese duma  Slovene dim          Slovene: Kostel dim
```

n·s|night · binary · 33 languages · enrichment 62×

```
Belarusian noč     Breton: Gwened nos, nosezhiad  Czech noʦ           Kashubian noc
Lower Sorbian noc  Middle Welsh nos               Old Breton nos      Old Polish noc
Old Welsh nos      Polish nɔʦ                     Russian noʧ         Serbo-Croat noć
Slovak noc         Slovene: Early Mo noč          Sorbian: Upper nóc  Ukrainian nìč
```

s·l|sun · binary · 32 languages · enrichment 40×

Campidanese sˈɔl+i	Danish sol	Faroese sól	Icelandic s.ouː.l
Lanzo_Torinese_Pi sul	Latin soːl	Lithuanian saulė	Norwegian suːl
Norwegian: Nynors sol	Old Norse sól	Old Prussian saule	Old Swedish sol
Portuguese sɔl	Sardinian: Logudo sole	Spanish sol	Venice_Venetian sol

d·n|day · binary · 31 languages · enrichment 91×

Assamese din	Bengali din	Bihari din	Croatian dan
Hindi dina	Latvian diena	Macedonian den	Magahi din
Old Church Slavon dĭnĭ	Old Novgorod denĭ	Polabian dan	Russian dʲenʲ
Serbo-Croat dan	Slovak dɛɲ	Slovene: Early Mo dan	Ukrainian denˊ

m·n|moon · binary · 30 languages · enrichment 68×

Ancient Greek μήνη	Bislama mun	Dutch maan	English mun
Flemish maan	Hawrami māŋ	Jamaican Creole mun	Kurdish S.: Qorve māŋ
Middle High Germa mâne	Negerhollands man	Norwegian: Bokmål måne	Old English mōna
Old High German mâno	Old Saxon mâno	Polabian mon	Tocharian A mañ

t·r|earth · binary · 27 languages · enrichment 41×

Anglo-Norman terre	Campidanese tˈɛrː+a	Dalmatian tˈar+a
Franco-Provençal térra	French tɛr	Ladin tera
Latin terːa	Milanese tèra	Old Breton tir
Old French terre	Old Occitan terra	Portuguese tɛʀɐ
Provençal_Occitan tˈɛʁ+o	Rapallo_Ligurian tˈær+a	Sardinian: Logudo terra
Venice_Venetian tˈɛr+a		

f·m|smoke · binary · 27 languages · enrichment 222×

Anglo-Norman fumee	Aromanian fˈum+u	Catalan fum	French enfumer
Friulian fum	Ladin fum	Latin fuːmo	Megleno-Romanian fum
Milanese füm	Old Catalan fum	Old French fumee	Old Spanish fumo
Portuguese fumu	Provençal_Occitan fym	Ravennate_Romagno fom	Sardinian: Logudo fumu

s·l|year · binary · 27 languages · enrichment 33×

Bakhtiari sāl	Balochi: Sistani sāl	Delvari sɒl	Hindi sāla
Judeo-Tat sal	Khowar sal	Kurdish C.: Jafi sāɬ	Kurdish N.: Bahdi sal
Lari sāl	Mazanderani sāl	Northern Kurdish sal	Persian sāl
Raji: Barzoki sạl	Sarikoli sul	Tati sāl	Wakhi sol

s·m·n|sky · ternary · 26 languages · enrichment 178×

Armenian ասմանի	Avestan asman-	Balochi: Sistani āsmān
Gawarbati asmaan	Hindi aːsmaːn	Khowar asmaan
Kurdish C.: Jafi āsmān	Magahi āsamān	Middle Persian āsmān
Northern Pashto ɑsˈmɑn	Parthian āsmān	Pashai: North-Wes aasmon
Persian āsmān	Sarikoli ismun	T-Agalyk ɔsmon
Wakhi osmon		

s·t·l|star · ternary · 25 languages · enrichment 150×

Anglo-Norman esteille	Catalan estel	Dalmatian: Veglio stala
Ferrarese_Emilian stˈɛl+a	Italian stelːa	Ladin steila
Latin steːlːa	Milanese stèla	Old Catalan estela
Old Occitan estella	Ossetian stʼalɨ	Ossetic: Digor stˈalu
Portuguese estela	Rapallo_Ligurian stˈelːa	Ravennate_Romagno stˈɛal+a
Venice_Venetian stˈel+a		

p·r·l|APRIL · ternary · 23 languages · enrichment 1116×

Albanian pɾiɬ	Armenian ɑpɾil	Armenian_Mod ɑpɾil	Bulgarian apriɬ
Croatian ǎpriːl	Dutch ɑprɪl	Dutch_List ɑprɪl	German April
Hindi əprɛːl	Icelandic apríl	Northern Pashto apˈril	Norwegian ɑpˈriːl
Ossetian aprelʲ	Russian ɐprʲeˑl	Slovene apriɬːl	Standard Albanian pɾiɬ

n·b|sky · binary · 23 languages · enrichment 284×

Belarusian neba	Bulgarian nebe	Croatian nebo
Hindi नभ	Macedonian nebo	Old Church Slavon nebo
Old Czech nebe	Old Polish niebo	Polabian nebü
Polish niebo	Rusyn nébo	Selice Romani nebo
Serbo-Croatian nebo	Slovak nebo	Slovene: Early Mo nebo
Slovene: Kostel nebo		

s·n|snow · binary · 23 languages · enrichment 27×

Armenian_Mod ȶ.ɪ̯u.n	Danish sne	Dutch snauw	Elfdalian sniɥo
English sno	German Schnee	German: Bernese Schnee	Kashmiri šiin
Luxembourgish Schnéi	Middle Dutch snee	Norwegian snøː	Norwegian: Bokmål snø
Old High German snio	Old Saxon snêo	Portuguese sinó	Sorbian: Upper sněh

m·r·s|MARCH · ternary · 22 languages · enrichment 645×

Albanian maɾs	Bengali mart͡ʃɔ	Breton møʀs	Catalan marcer
Czech marš	French mars	Icelandic mars	Italian mart͡so
Norwegian mɑrs	Persian mɒːɾs	Polish marsz	Romanian marț
Russian марш	Slovak marɛt͡s	Spanish marcha	Standard Albanian maɾs

s·t·n|stone · ternary · 22 languages · enrichment 109×

Bislama ston	Danish sten	Dutch steen
English ston	Flemish steen	Frisian stien
Icelandic s.t.ei.t͡n	Luxembourgish Steen	Middle High Germa stein
Norwegian: Bokmål stein	Norwegian: Nynors stein	Old Frisian stēn
Old High German stein	Old Norse steinn	Old Swedish sten
Swedish steːn		

v·n·t|wind · ternary · 22 languages · enrichment 303×

Anglo-Norman vent	Catalan vent	Ferrarese_Emilian vent
German Wind	Italian vɛnto	Ladin vent
Megleno-Romanian víntu	Megleno_Romanian vint	Old French vent
Old Occitan vent	Palermitan_Sicili vˈɛnt+u	Provençal_Occitan vˈɛnt
Rapallo_Ligurian vˈeŋt+u	Ravennate_Romagno vˈɛnt	Saramaccan vḗntu
Venice_Venetian vˈeŋt+o		

p·s·k|sand · ternary · 21 languages · enrichment 152×

Belarusian pâsok	Bulgarian pâs″k	Czech pi:sɛk
Lower Sorbian pěsk	Macedonian: Suho pesok	Megleno_Romanian pis'ok
Old Church Slavon pěsŭkŭ	Old Polish piasek	Polabian ṕosăk
Polish piasek	Rusyn p´isók	Serbo-Croatian pesak
Slovak piesok	Slovene: Early Mo pesek	Slovene: Kostel pesek
Sorbian: Upper pěsk		

Animals

p·r|feather (47 languages) heads the block, and ”feather” also has a second code, p·n (20).

m·x|fly (300×) and m·s|fly (40×) are two codes for ”fly” differing in one position, and m·s is the same skeleton as ”meat”, ”man” and ”mouse”. That is: in a single line of the reverse index a sibling pair and a four-concept collision coexist. Telling one from the other means looking at the forms, which is why the appendix prints them.

m·r|snake (22) and m·r|ant (21) share a skeleton with m·r|sea (62) and m·r|die (38). Four concepts, two consonants.

p·r|feather · binary · 47 languages · enrichment 90×

Bakhtiari par	Belarusian pâro	Czech pɛ:ro
Hindi pər	Khotanese pārrä	Kurdish S.: Qorve paṛ
Macedonian: Suho pero	Old Church Slavon pero	Pashai: North-Wes par
Polabian perü	Romani por	Selice Romani pór
Slovak pero	Slovene: Kostel pero	Tati par
Urdu pəɾ		

m·x|fly_N · binary · 27 languages · enrichment 300×

Belarusian muha	Bulgarian muha	Czech moucha
Greek miɣa	Greek: Cypriot moýgia	Greek: Modern Std mýga
Lower Sorbian mucha	Macedonian mùxa	Old Church Slavon muxa
Old Polish mucha	Polabian mauxo, maixo	Russian muha
Serbo-Croat muha	Slovak mucha	Slovene: Early Mo muha
Sorbian: Upper mucha		

l·s|louse · binary · 26 languages · enrichment 57×

Bislama laos	Danish lus	Dutch_List lœys	Elfdalian laus
Faroese lús	Frisian lûs	German Laus	Icelandic lu:s
Luxembourgish Laus	Middle Dutch luus	Negerhollands lus	Norwegian lʉ:s
Norwegian: Nynors lus	Old High German lūs	Old Norse lús	Swedish lu:s

r·b|fish · binary · 25 languages · enrichment 193×

Belarusian ryba	Bulgarian ribə	Czech rɪba	Istro_Romanian r'ib+'æ
Lower Sorbian ryba	Macedonian riba	Old Church Slavon ryba	Old Czech ryba
Old Polish ryba	Polish rɨba	Russian rɨba	Serbo-Croat riba
Serbo-Croatian riba	Slovene riba	Slovene: Early Mo riba	Sorbian: Upper ryba

m·s|fly_N · binary · 22 languages · enrichment 40×

Bengali mat͡ʃʰi	Bhojpuri māchī	French mouche	Kalaṣa-alâ: Nišei mâša
Kashmiri məčh	Kumzari mēš	Kurdish N.: Bahdi mēš	Ladin moscia
Latvian muša	Lithuanian mùsė	Magahi machī	Northern Kurdish meʃ
Northern Pashto mut͡ʃ	Old Prussian muso	Pashto mach	Saṇu-viri: Wâmâ mâi´oːs

m·r|snake · binary · 22 languages · enrichment 42×

Bakhtiari mār	Balochi: Sistani mār	Delvari mɒr	Judeo-Tat mar
Kumzari mār	Kurdish C.: Jafi mar	Kurdish S.: Qorve mār	Lari mār
Middle Persian mār	Northern Kurdish mar	Northern Pashto mɑr	Persian mār
Persian: Tehran mār	Raji: Barzoki maṛ	Tati mār	Y-Dugova 1 mɔr

m·s|MOUSE · binary · 21 languages · enrichment 147×

Belarusian mɨʂ	Croatian mîʃ	Czech myš	Danish mus
Dutch_List mœys	English maʊs	German Maus	Icelandic mus
Italian mouse	Latin mus	Persian muʃ	Polish mɨʂ
Russian mɨˑʃ	Slovene miɪʃ	Spanish mouse	Swedish mus

m·r|ant · binary · 21 languages · enrichment 46×

Avestan maoiri-	Avestan: Younger maoiri	Danish myre
Dutch mier	Faroese meyra	Flemish mier
Icelandic maur	Lari mur	Middle Dutch miere
Middle Persian mōr	Norwegian mæur	Norwegian: Bokmål maur
Old Norse maurr	Persian mur	Persian: Tehran murčeh
Swedish myra		

k·n|dog · binary · 20 languages · enrichment 40×

Albanian kʸen	Ancient Greek 'kuōn	Aromanian k'ɨn+e
Ferrarese_Emilian kan	Greek: Mycenaean ku-na-∗	Italian can
Lanzo_Torinese_Pi kaŋ	Latin kani	Neapolitan cane
Old Spanish can	Palermitan_Sicili k'an+i	Rapallo_Ligurian kaŋ
Romanian câine	Sardinian: Logudo cane	Sardinian: Nuoro cane
Spanish can		

p·n|feather · binary · 20 languages · enrichment 58×

Anglo-Norman penne	Aromanian p.ˈe̯a.n.+.ə	Campidanese p'inː+a
Ferrarese_Emilian p'ɛn+a	Italian pɛnːa	Khwarazmian pan
Latin penːa	Megleno_Romanian p.ˈe̯a.n.+.ə	Old French pene
Palermitan_Sicili p'inː+a	Pali paṇṇa	Portuguese penɐ
Rapallo_Ligurian p'eŋː+a	Ravennate_Romagno pˈɛn+a	Romanian panə
Sardinian: Nuoro pinna		

k·r·m|worm · ternary · 20 languages · enrichment 183×

Avestan kərəma-	Balochi: Sistani kerm	Delvari kerm
Hawrami kirm	Judeo-Tat kürm	Khwarazmian kirm
Kumzari kurm	Kurdish C.: Jafi kɪrm	Kurdish S.: Qorve kɪrm
Lari kerm	Middle Persian kirm	Northern Kurdish kʊrm
Persian: Tehran kerm	Romani kermo	T-Agalyk kirim
Y-Dugova 1 kirm		

Plants and food

`s·l|salt` (48 languages) and `s·l·t|salt` (20) are the same concept with and without a final consonant; `s·l` is also the code for ”sun” (32) and ”year” (27).

And here is the second half of what §5.5 explains: `s·k·r|SUGAR` 452×, `s·p|SOUP` 295×, `f·r·t|fruit` 262×, `b·r|BEER` 189×. Together with the months of the previous block and with `s·k·l|SCHOOL` (712×) of the next, they are the highest enrichments in the whole appendix — and they are, visibly, vocabulary these languages share by routes other than ordinary transmission. The instrument cannot distinguish them from an inherited word, because both are ”many languages saying the same thing”.

Lined up, they are the best illustration the series has produced that enrichment measures coverage, not age.

`s·l|salt` · binary · 48 languages · enrichment 61×

Anglo-Norman sel	Campidanese s'al+i	Czech su:l	
French sɛl	Kashubian sól	Latin sal	
Macedonian: Suho sol	Old Church Slavon solĭ	Old Novgorod solĭ	
Old Spanish sal	Polish sul	Ravennate_Romagno seal	
Sardinian: Logudo sale	Serbo-Croatian sol	Sorbian: Upper sól, sel	
Tocharian B salyiye			

`m·s|meat` · binary · 46 languages · enrichment 72×

Albanian miʃ	Albanian: Arbëres mishë	Armenian: Classic mis	Armenian_Mod mis
Bihari mā̃s	Czech maso	Hindi māṁsa	Macedonian meso
Marathi mãũs	Old Armenian մի̈ս	Old Novgorod męso	Palula mhaás
Romani mas	Serbo-Croat meso	Slovak mäso	Slovene: Kostel meso

`s·p|SOUP` · binary · 30 languages · enrichment 294×

Albanian sʊpə	Albanian 'supə	Armenian_Mod sup	Bulgarian supa
Croatian sûpa	Dutch_List soep	French soupe	Greek_Mod 'supa
Icelandic supʰɑ	Italian t͡ʃup:a	Negerhollands səp	Persian sup
Romanian supă	Serbo-Croatian supa	Standard Albanian sʊpə	Ukrainian sup

`f·r·t|fruit` · ternary · 25 languages · enrichment 262×

Albanian frutë	Albanian früt	Albanian: Standar frut	Anglo-Norman frut
Dutch fruit	Dutch_List fruit	Frisian fruit	Greek: Italiot froúttho
Greek_Mod 'fruto	Jamaican Creole frut	Ladin frut	Milanese früta
Neapolitan frutta	Old Occitan fruit	Old Spanish fruta	Portuguese: Brazi fruta

`g·r·s|grass` · ternary · 22 languages · enrichment 115×

Danish græs	Elfdalian gras	English gras	German Gras
German: Bernese Gras	Gothic gras	Jamaican Creole grās	Luxembourgish Gras
Negerhollands gras	Norwegian grɛs	Norwegian: Bokmål gress	Old English græs
Old Frisian gers	Old High German gras	Old Swedish gräs	Spanish grass

`k·r·n|root` · ternary · 22 languages · enrichment 73×

Belarusian koranʹ	Bulgarian kɔrɛn	Croatian kȏriɛn
Istro_Romanian k'oren	Lower Sorbian kórjeń	Macedonian koren
Megleno_Romanian k'orin	Old Church Slavon korenĭ	Old Polish korzeń
Polish korzeń	Russian korʲenʲ	Serbo-Croatian koren
Slovak koreň	Slovene koren	Slovene: Kostel koren
Sorbian: Upper korjeń		

s·k·r|SUGAR · ternary · 21 languages · enrichment 452×

Albanian še'kʸer	Armenian շաքար	Armenian šakar	Armenian_Mod šakʰar
Catalan sucre	Czech cukr	Danish sukker	Dutch_List suiker
English saccharo-	French sucre	Irish siúcra	Judeo-Tat šeker
Lower Sorbian cukor	Persian šekær	Portuguese açúcar	Swedish socker

l·s·t|leaf · ternary · 21 languages · enrichment 142×

Belarusian list	Bulgarian list	Croatian list
Czech lɪst	Macedonian list	Macedonian: Suho list
Old Church Slavon listŭ	Old Polish list	Polabian laist
Polish list	Rusyn líst	Serbo-Croat list
Serbo-Croatian list	Slovene list	Slovene: Early Mo list
Slovene: Kostel list		

s·m|seed · binary · 21 languages · enrichment 54×

Belarusian sʲɛmʲa	Bulgarian sɛmɛ	Czech siemě
German: Bernese Saame	Lower Sorbian semje	Macedonian seme
Macedonian: Suho seme	Old Church Slavon sěmę	Old Czech siemě
Old High German sâmo	Polish ɕɛmʲɛ̃	Rapallo_Ligurian s'em+i
Russian semâ	Slovene seme	Slovene: Early Mo seme
Slovene: Kostel seme		

k·r·n|meat · ternary · 20 languages · enrichment 66×

Aromanian k'arn+e	Catalan carn	Ferrarese_Emilian k'aran
French carné	Italian carne	Lanzo_Torinese_Pi karn
Latin karn	Megleno_Romanian k'arn+i	Neapolitan carne
Old Catalan carn	Old Spanish carne	Palermitan_Sicili k'aɾn+i
Rapallo_Ligurian k'arn+e	Ravennate_Romagno k'earn+a	Romanian carne
Spanish carne		

s·l·t|salt · ternary · 20 languages · enrichment 126×

Czech osolit	Danish salt	Elfdalian solt	English sɔlt
Gothic salt	Icelandic saḷt	Jamaican Creole sālt	Latvian sālīt
Norwegian: Bokmål salt	Norwegian: Nynors salt	Old English sealt	Old Frisian salt
Old Saxon salt	Old Swedish salt	Russian солить	Swedish salt

Verbs of action and state

The smallest block of the appendix: four codes, all binary, with reaches between 20 and 38 languages.

That verbs yield so little compared with nouns is constant across the five families of the series, and the candidate explanation is the usual one: the lists the corpus comes from cite verbs in inflected forms or with affixes that depend on the source, and §5.3 shows that removing them automatically makes the result worse. Part of what is missing here is morphology we do not know how to separate.

m·r|die · binary · 38 languages · enrichment 74×

Aromanian m'or+u	Bulgarian umrɤ	Catalan moɾ+'i
Friulian m.u.r.+.'iː	Kalaṣa-alâ: Nišei mrẽ-	Khotanese mīḍe
Kâta-vari: Ktivi mř´e-	Lanzo_Torinese_Pi m'œr+i	Macedonian umre
Megleno_Romanian mor	Middle Welsh marw	Old Catalan morir
Palula máara	Provençal_Occitan muɾ+'i	Vedic: Early mr̥-
Welsh marw		

m·l|grind · binary · 22 languages · enrichment 68×

Breton: Treger malã, mal-	Bulgarian melâ	Danish male
Faroese mala	Gaelic: Irish meil	Gaelic: Scottish meil
Icelandic mala	Irish meil	Macedonian: Suho mele
Middle Welsh malu	Norwegian: Bokmål male	Old Norse mala
Old Swedish mala	Seychelles Creole moule	Tocharian B mäl-
Welsh malu		

k·m|come · binary · 20 languages · enrichment 93×

Bislama kam	Danish kʌmə	Dutch 'koːmə	Dutch_List komen
English come	Faroese koma	Frisian komme	German kom
Jamaican Creole kɔm	Norwegian kɔmə	Norwegian: Bokmål komme	Norwegian: Nynors kome
Old Norse koma	Old Swedish koma	Swedish komːa	Tocharian A kum-

d·t|give · binary · 20 languages · enrichment 105×

Croatian dati	Czech daːt	Kurdish C.: Jafi adāt	Latgalian dūt
Lithuanian dûətʲɪ	Old Church Slavon dati	Old Czech dát	Old Novgorod dati
Polabian dot	Russian datʲ	Rusyn dátɪ	Serbo-Croat dati
Slovene dati	Slovene: Early Mo dati	Slovene: Kostel dati	T-Agalyk dot

Objects and culture

s·k·l|SCHOOL at 712× is the section's highest enrichment and the appendix's third. It is followed by t·n|TONE (MUSIC) (309×), k·r·s|CROSS (268×), k·n|CANOE (236×) and l·n|LINE (194×).

Five concepts of material or institutional culture, five very high enrichments, and no surprise once §5.5 has been read. k·n is also the code for "knee" and "dog".

h·s|house (195×) and k·r|bark (38×) are the two in the block that do not follow that pattern, and k·r shares a skeleton with "heart".

h·s|house · binary · 27 languages · enrichment 195×

Bislama haos	Czech house	Dutch house	English haʊs
Faroese hús	Frisian hûs	German: Bernese Huus	Icelandic hus
Middle Dutch huus	Negerhollands hus	Norwegian hʉːs	Norwegian: Nynors hus
Old Frisian hūs	Old High German hûs	Old Saxon hûs	Swedish house

k·r·s|CROSS · ternary · 25 languages · enrichment 268×

Belarusian krɨʑ	Bengali kruʃ	Croatian krîːʒ	Danish kɔʁs
Dutch_List kruis	English cross	Hindi kruːs	Icelandic kross
Italian croce	Neapolitan croce	Norwegian krʏs	Portuguese cruz
Romanian cruce	Slovak kriːʃ	Slovene kriɻːʒ	Ukrainian криж

l·n|LINE · binary · 25 languages · enrichment 195×

Albanian linjë	Albanian 'lʸiɲə	Dutch linie	Dutch_List lijn
French liɲ	German Linie	Icelandic linɑ	Irish lʲiːnʲə
Jamaican Creole laɪn	Negerhollands lin	Polish linia	Romani linia
Romanian linie	Saramaccan lín	Slovak ɬiːn̪ia	Welsh llin

s·k·l|SCHOOL · ternary · 25 languages · enrichment 712×

Albanian shkollë	Albanian 'školə	Bulgarian škóla	Catalan escola
Danish skole	English school	Icelandic skóli	Jamaican Creole skul
Latin schola	Negerhollands skol	Old High German skuola	Romani skoala
Romanian şcoală	Selice Romani iškola	Serbo-Croatian škola	Swedish skola

t·n|TONE (MUSIC) · binary · 24 languages · enrichment 313×

Albanian tɔn	Armenian tɔn	Breton tõːn	Bulgarian tɔn
Czech toːn	Danish 'toːnə	Dutch_List toːn	English təʊn
Irish t̪ˠɔn̪ˠ	Italian tono	Polish tɔn	Romanian ton
Slovene tɔɻːn	Spanish tono	Swedish tuːn	Ukrainian tɔn

k·r|bark · binary · 21 languages · enrichment 38×

Albanian 'kore	Belarusian kara	Bulgarian kora
Croatian kora	Istro_Romanian kˈor+a	Kashubian kòra, kóra
Macedonian kora	Old Church Slavon kora	Old Czech kóra
Old Polish kora	Russian kora	Rusyn korá
Serbo-Croat kora	Slovak kʊ̯ora	Slovene: Kostel kora
Ukrainian kora		

l·t·r|ALTAR · ternary · 20 languages · enrichment 939×

Albanian alʸ'tar	Bulgarian oltár	Czech oltář	Dutch altaar
English altar	Icelandic altari	Irish altóir	Italian altare
Old High German altâri	Portuguese altar	Romani altari	Romanian altar
Selice Romani oltári	Serbo-Croatian oltar	Spanish altar	Swedish altare

Qualities

The section where the splitting of one concept across several codes goes furthest.

”Full” has three codes: `p·r` (25 languages), `p·l·n` (23) and `f·l` (21). ”New” has two differing in a single position, `n·v` (37) and `n·w` (33). ”Dry” has two, `s·k` (33) and `s·x` (27), also differing in one. ”Green” has two with no consonant in common, `g·r·n` (23) and `z·l·n` (21). And ”good” has two, `b·n` (27) and `d·b·r` (21).

Five concepts split, and all five in a different way. It is the block that best shows why §5.2 exists as a separate section: the catalogue produces these pairs by the thousand — 3,647 in all — and sorting them into ”two ways of writing the same thing”, ”two different words” and ”the same root with different inflection” is work the catalogue does not do by itself.

`s·t·r|old` (22) shares a skeleton with ”star” and ”four”. It is the fourth time this warning appears in the appendix, and with 10,097 codes over 1,944 concepts it will not be the last.

`d·r|far` · binary · 42 languages · enrichment 81×

```
Assamese dūr  Avestan: Younger dūire  Bengali dūre       Bihari dur
Gawri dūr     Judeo-Tat duri          Khotanese durä     Kurdish N.: Bahdi dir
Magahi dūr    Mazanderani dir         Old Persian dūrai  Pali dūra
Persian dur   Romani dur              Sinhalese dura     Vedic: Early dūrá-
```

`n·v|new` · binary · 37 languages · enrichment 267×

```
Avestan nava-           Breton nevez             Croatian nov
Dalmatian: Veglio nuf   Franco-Provençal nuvo    Istro_Romanian nov
Kashubian nowi          Old Czech nový           Old Polish nowy
Polabian nüvě           Portuguese: Brazi novo   Ravennate_Romagno n'ɔav
Selice Romani névo      Seychelles Creole nouvo  Slovene: Early Mo nov
Venice_Venetian n'ov+o
```

`s·k|dry` · binary · 33 languages · enrichment 61×

```
Catalan assecar            English secco            French sɛk
Hindi suːkʰaː              Lanzo_Torinese_Pi sœk    Latin sikːo
Milanese sèch              Nepali sukkhā            Old Persian uška
Palermitan_Sicili s'ikːu   Palula šúku              Rapallo_Ligurian s'ekːu
Romani šuko                Sardinian: Logudo siccu  Selice Romani šúko
Spanish seko
```

`n·w|new` · binary · 33 languages · enrichment 355×

```
Aromanian now        Breton: Treger neweż   Catalan nɔuw
Dutch nieuw          Flemish nieuw          Lari now
Lower Sorbian nowy   Middle Cornish nowyth  Middle High Germa niuwe
Old Catalan nou      Old High German niuui  Pashto nawe
Persian: Tehran naw  Slovene nov            Sogdian nawi
Tocharian B ñuwe
```

`k·r·t|short` · ternary · 29 languages · enrichment 172×

```
Catalan curt               Dalmatian: Veglio kṳart  Dutch kort
Ferrarese_Emilian kurt     French courtaud          Italian corto
Kurdish N.: Bahdi kurt     Lanzo_Torinese_Pi kyrt   Milanese cürt
Norwegian kɔrt             Old Saxon kurt           Old Spanish corto
Palermitan_Sicili k'uɾtu   Provençal_Occitan kuɾt   Ravennate_Romagno kurt
Swedish kɔrt
```

`s·x|dry` · binary · 27 languages · enrichment 144×

```
Belarusian suhì    Breton sec'h           Bulgarian sux         Croatian suh
Czech sʊxiː        Lower Sorbian suchy    Macedonian: Suho sùx  Middle Breton sech
Middle Welsh sych  Old Czech suchý        Old Polish suchy      Serbo-Croat suh
Slovene suh        Slovene: Early Mo suh  Sorbian: Upper suchi  Welsh sych
```

b·n|good · binary · 27 languages · enrichment 78×

Anglo-Norman bon	Aromanian b'un+u	Catalan b'ɔn	Dalmatian: Veglio buṅ
English ben	French bɔn	Italian b'ɔn+o	Ladin bon
Latin bono	Megleno_Romanian bun	Milanese bùn	Old French bon
Provençal_Occitan boŋ	Rapallo_Ligurian buŋ	Saramaccan búnu	Sardinian: Nuoro bonu

p·r|full · binary · 25 languages · enrichment 46×

Bactrian poro	Bakhtiari por	Delvari por	Gawarbati por
Hindi puːraː	Judeo-Tat pur	Kamviri pâr′ea	Kurdish C.: Jafi par
Kâta-vari: Easter pur′a	Mazanderani per	Middle Persian purr	Palula purá
Parthian purr	Persian: Tehran por	Raji: Barzoki pör	Tati pör

p·l·n|full · ternary · 23 languages · enrichment 186×

Anglo-Norman plein	Aromanian pl'in+u	Bulgarian pɤlɛn	English plene
French plɛn	Ladin plen	Latin pleːno	Macedonian: Suho poln
Old Church Slavon plŭnŭ	Old Czech plny	Old Occitan plen	Old Polish pełny
Provençal_Occitan pleŋ	Romanian plin	Slovene pɔɻːln	Slovene: Early Mo puln

g·r·n|green · ternary · 23 languages · enrichment 137×

Bislama grin	Danish gʁœnˀ	Elfdalian gryön	Frisian grien
German Green	Jamaican Creole grin	Luxembourgish gréng	Negerhollands grun
Norwegian gɾœn	Norwegian: Bokmål grønn	Old English grēne	Old Frisian grēne
Old Norse grœnn	Old Swedish grön	Spanish green	Swedish grøːn

g·r·m|hot · ternary · 23 languages · enrichment 282×

Assamese gôrôm	Avestan garəma-	Avestan: Younger garəma
Bhojpuri garam	Bihari garam	Hindi garama
Judeo-Tat germ	Kumzari garm	Kurdish C.: Jafi garm
Kurdish N.: Bahdi germ	Lari garm	Magahi garam
Mazanderani garm	Middle Persian garm	Parthian garm
Persian: Tehran garm		

l·n·g|long · ternary · 23 languages · enrichment 350×

Aromanian l'ung+u	Campidanese l'oŋg+u	Dalmatian luŋg
Elfdalian laungg	Ferrarese_Emilian luŋg	Italian lungo
Lanzo_Torinese_Pi luŋg	Old English lang	Old Frisian long
Old High German lang	Palermitan_Sicili l'ɔŋg+u	Rapallo_Ligurian l'uŋg+u
Romani lungo	Romanian luŋg	Sardinian: Logudo longu
Sardinian: Nuoro longu		

s·t·r|old · ternary · 22 languages · enrichment 55×

Belarusian stary	Bulgarian star	Croatian star	Danish stor
Kashubian stôri	Lower Sorbian stary	Macedonian: Suho star	Old Church Slavon starŭ
Old Polish stary	Polabian storě	Polish staɾɨ	Serbo-Croatian star
Slovak stariː	Slovene star	Slovene: Kostel star	Sorbian: Upper stary

f·l|full · binary · 21 languages · enrichment 56×

Danish fulˀ	Elfdalian full	English fʊl	French full
German fol	Jamaican Creole fʊl	Luxembourgish voll	Negerhollands ful
Norwegian fʉl	Norwegian: Bokmål full	Old English full	Old Frisian ful
Old High German fol	Old Swedish full	Polish ful	Swedish ful:

d·b·r|good · ternary · 21 languages · enrichment 391×

Belarusian dobry	Bulgarian dobɤr	Croatian dobar
Czech dobri:	Lower Sorbian dobry	Macedonian dobar
Macedonian: Suho dobar	Old Czech dobrý	Old Polish dobry
Polabian dübrě	Serbo-Croat dobar	Serbo-Croatian dobar
Slovak dobrý	Slovene: Early Mo dober	Slovene: Kostel dober
Sorbian: Upper dobry		

z·l·n|green · ternary · 21 languages · enrichment 560×

Belarusian zâlëny	Bulgarian zɛlɛn	Croatian zelen
Czech zɛlɛni:	Kashubian zelony	Lower Sorbian zeleny
Macedonian zelen	Old Church Slavon zelenŭ	Old Czech zelený
Romani zeleno	Serbo-Croat zelen	Serbo-Croatian zelen
Slovak zelený	Slovene: Early Mo zelen	Slovene: Kostel zelen
Sorbian: Upper zeleny		

l·s|smooth · binary · 20 languages · enrichment 47×

Albanian: Arbëres lish	Ancient Greek 'leios	Catalan llis
French lis	Greek: Ancient leĩos	Greek: Modern Std leíos
Greek: New Testam leĩos	Greek_Mod 'lios	Ladin lize
Milanese lìs	Neapolitan liscio	Old Catalan llis
Old Spanish liso	Portuguese liʃu	Portuguese: Brazi liso
Spanish liso		

Other

f·g·r|IDOL · ternary · 20 languages · enrichment 1838×

Albanian figʊɾə	Breton fi:gɔʁ	Bulgarian figura	Croatian figŭ:ra
Danish fi'gu:ˀʁ	French fi'gy:ʁ	Italian figura	Latvian figu:rɑ
Northern Kurdish fəgur	Norwegian fi:'gʉ:r	Polish figuɾa	Romanian figurə
Slovak figu:ra	Slovene figu:ra	Standard Albanian figʊɾə	Swedish fi:gər