Consonantal codes in 27 Uralic languages

The family where the instrument can finally do what it promises

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

Abstract

We present a catalogue of 1,486 consonantal codes grouping 13,470 lexical forms from 27 Uralic languages. A code is a consonant skeleton shared by words for the same concept across distinct languages. No reconstruction, no expert cognacy, and no hypothesis about Proto-Uralic is used.

This is the series' third paper and the first in which the family has the structure the method needs: nine branches, six of them with sufficient data, against Turkic's two and Austronesian's six very unequal ones. That makes it possible, for the first time, to do what the protocol promised and could not test: computing substitutions per branch pair over thirty-six pairs instead of one.

Four results. The codes of widest reach — `n·m|name` with 20 languages and 378×, `v·r|blood`, `k·l|fish`, `m·n|I`, `t·n|THOU` — are the ones any Uralicist will recognise at once, obtained without consulting any reconstruction. The search for sibling codes — pairs for the same concept differing in one position — partitions languages into groups using no genealogical label, and its disjointness filter turns out to be indispensable: a quarter of the pairs share languages and therefore partition nothing. Geography here yields a monotonic gradient across all six buckets, which Turkic could not measure for want of range. And automatic affix stripping again makes the result worse — ten points of grouping and eighteen of null — confirming in a second agglutinative family what the Turkic paper found.

With three families measured by the same code, the chance ratio orders by time depth — Turkic 216×, Uralic 78×, Austronesian 37× — but observed precision does not. That is a datum, not a thesis. *(With the eight families the series now has, the ordering does not hold: see the note in §5.6.)*

1. What is sought, and why Uralic

Historical linguistics has a powerful and expensive instrument: the comparative method. It has been applied with enormous unevenness — Indo-European has dense reconstruction and expert cognacy, most families do not — so that any study leaning on the historical layer of the lexicon measures, without meaning to, the density of the comparative literature rather than the past of the languages.

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 Austronesian paper showed that a catalogue of consonantal codes remains; the Turkic paper asked what its figures mean.

This paper asks how far the instrument goes when nothing stands in its way. The Turkic paper declared three limits that turned out to belong to the family and not the method: its corpus distinguishes only two branches — one with a single language — so the rule of computing substitutions per branch pair was tested on one pair; the geographic analysis came out degenerate, with 467 of 496 pairs in one bucket; and the sibling-code search arrived at the end, with nothing to check it against.

Uralic lifts all three restrictions. Nine distributed branches, a geographic spread of thousands of kilometres without being an archipelago, and an established primary division.

> And here one must be precise, because it is easily misread. Those nine branches are not an input to the method: the catalogue is built exactly as in the other two families, without knowing they exist. What the branches allow is checking what the catalogue produced — and that is why Uralic is a good family for this paper: not because the instrument needs them, but because we need something to check it against.

1.1 The method, and where it is described

The procedure — what a code is, in what order it is built, what nulls it uses and what beating them authorises — is described once for the whole series in "The catalogue of consonantal codes", and is not repeated here. The essentials, so that what follows can be read:

The catalogue is built without genealogy. Codes come from a form's consonant skeleton, the concept it expresses, and a threshold of three distinct languages. No branch classification, no proto-forms, no cognate lists. The procedure does not know this family has branches, and genealogy enters at the end only to read an already-closed catalogue.

Nothing is reconstructed, no kinship is claimed, inheritance is not distinguished from borrowing, and no meaning is attributed to a code.

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 in Uralic

2.1 The consonant skeleton

The skeleton is a word's sequence of consonants, with the vowels removed — not lost: they stay in a separate column — and each consonant written with a canonical representative. Uralic uses 25 canonical symbols, of which 23 merge something. The full table is in `runs/normalizacion.md` and is not written by hand: it is derived by aligning each form with its skeleton.

What matters when reading the results:

Written	as	Consequence in the data
s š ś ʃ ʂ ɕ z ž ʒ and every affricate	`s`	‘s·l·m
eye‘ gathers *silmä*, *šalmə*, *śeľme*		
n ŋ ɲ ɳ	`n`	‘n·m
name‘ gathers *nimi*, *ɲim*, *nam*		
l ɬ ɬ ɬ	`l`	
k, q, x are kept separate		which is why §5.2 can see `k·l` / `q·l` / `x·l`
v and w are kept separate		which is why it can see `v·r` / `w·r`

That **k**, **q** and **x** are not merged is the table's most productive decision in this family, and it is worth declaring before the results exploit it: it is what lets the Ob-Ugric and Samoyedic velar change appear as three distinct codes instead of one (§5.2). In Turkic, the opposite decision about sibilants destroyed a diagnostic correspondence.

2.2 A code is a skeleton recurring 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 The instrument’s cost: collision

59% of Uralic forms share their skeleton, within their own language, with a form of another concept. It is the clean measure of what the skeleton fails to distinguish, with no genealogy involved.

The clearest case is `k·l`, appearing with ”fish” and with ”tongue”. Finnish alone shows it: *kala* and *kieli*. Two words no description would confuse, and the skeleton writes them alike because the difference lies entirely in the vowels, which is what it discards. Likewise `t·l` with ”fire” (*tuli*) and ”winter” (*talvi*), and `n·l` with ”four” (*neljä*) and ”arrow” (*nuoli*).

Four pairs of distinct words under one code among those the appendix shows. It is not a defect more data would fix: it is the arithmetic of reducing a word to two or three consonants, and in Uralic it bites despite its words being the longest of the series.

2.4 Arity

Arity	Codes	
1 (unary)	166	11.3%
2 (binary)	652	44.2%
3 (ternary)	523	35.5%
4 or more	133	9.0%

The figures this paper defends are restricted to the 1,186 binaries and ternaries.

Figure 1: Distribution of codes by arity. 91% have three consonants or fewer; Uralic is the series’ longest-word family, at 3.27 consonants per form.

The mean arity of Uralic forms is 3.27 consonants, the highest of the series’ three families (Turkic 2.85, Austronesian 2.79). That matters for reading §5.6.

3. Method

3.1 Data

Integrative Corpus, Uralic subset: 27 languages, 80,032 forms with transcription, assigned concept and skeleton. Sources: Lexibank, IDS and NorthEuraLex, under CC-BY and CC-BY-SA licences.

The branches, which are what this family contributes:

Branch	Catalogued forms	Branch	Forms
Finnic	2,819	Samoyedic	374
Saami	1,581	Khantyic	285
Permian	1,379	Mansic	276
Mordvin	532	Hungaric	145
Mari	434		

Nine branches, the largest at 36% of the forms. It is the most balanced distribution of the three families: in Turkic one branch had a single language; in Austronesian, Malayo-Polynesian took 95%.

One exclusion, and of the kind that justifies phase 0. The corpus had a language called `Xincheng Beigeng`, with 1,008 forms, marked as Finnic. It is not Uralic: it is Kam-Sui (Tai-Kadai), from southern China. Three independent checks flagged it without anyone looking for it — it sits 6,446 km from the family's centre (23.8° N / 108.6° E, Guangxi), it marks tone in digits when no other Uralic language does, and its forms are *θaːm*^{453} "three" and *ðɐm*^{214} "water". Excluded and declared in `ci/decisiones.py`.

The raw corpus holds 99,255 forms and we publish 80,032. The difference is 18,113 forms without a skeleton — spread evenly across all languages, around 28% of each — plus the 1,008 excluded. It is not a script problem: Uralic has 25 Cyrillic forms in total. Those forms simply never got segmented into IPA. We checked that zero concepts exist only in forms without a skeleton, so the loss costs no coverage, but the published count is that of the forms that enter.

A single macroarea, hence no geographic-anomaly section.

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, preserving each one's inventory. The real grouping covers 33.6% of the forms; the null, 5.0%.

3.4 One population for the whole paper

Every report under `runs/` declares its population in its header, and they are identical:

```
POBLACIÓN · Uralic: 27 lenguas · 80,032 formas · 1,843 conceptos · 9 ramas ·
            despiece=sí(0) · excluidos=1 · umbral=3 lenguas
```

It is a mechanical check (`analysis/tools/verifica_paper.py`, check 7) and it exists because the Turkic paper had four sections computed over different populations without anyone noticing until review.

4. Results

4.1 The catalogue

1,486 codes over 867 of the 1,843 concepts, grouping 13,470 forms.

Magnitude	Value	Over what population
Forms inside a code	16.8% (13,470 of 80,032)	threshold of 3 languages, the catalogue's
Forms in some group	33.6% (null: 5.0%)	threshold of 2 languages, grouping test
Precision	6.50% (chance: 0.083%)	pairs of forms from different languages
Times above chance	78× [78–79; chance is computed exactly]	idem
Binaries and ternaries	1,186 (79.8%)	of the 1,486 codes

Correction of 2026-09-09. The coverage row read 9.8% (7,825 of 80,032), and that figure was not what its name promises: it counted distinct (language, spelling) pairs — the rows of the per-form TSV, which deduplicates — rather than forms. The catalogue's coverage over forms is 16.8%. The same error was in the Turkic paper; the other six use the correct figure. It is still the lowest coverage of the three families, which is what this paragraph uses.

Each row carries its population because there are three different ones. The catalogue's coverage (16.8%) is the lowest of the three families, with an explanation that is not a defect: with 1,843 concepts and only 27 languages, most concepts do not reach three languages with the same form.

4.2 The codes of widest reach are the known etyma

Concept	Code	Languages	Enrichment	Branches it appears in
name	`n·m`	20	378×	Finnic, Saami, Permian, Khantyic, Mansic, Samoyedic, Mari
blood	`v·r`	19	343×	Finnic, Saami, Permian, Mordvin, Mari, Hungaric
fish	`k·l`	19	174×	Finnic, Saami, Mordvin, Mari, Khantyic, Mansic, Samoyedic
I	`m·n`	18	194×	Finnic, Saami, Permian, Mordvin, Khantyic, Samoyedic
four	`n·l`	17	204×	Finnic, Saami, Permian, Mordvin, Mari
egg	`m·n`	17	183×	Finnic, Saami, Permian, Mordvin, Mari
sea	`m·r`	16	239×	Finnic, Saami, Permian, Mordvin, Samoyedic
fire	`t·l`	15	155×	Finnic, Saami, Permian, Mordvin, Mari

This is the catalogue's output, no more. The branch column describes how each code's languages are distributed; it does not claim that distribution has a historical cause, which is a different question and not this method's.

Figure 2: The twenty binary and ternary codes of widest reach, by number of languages.

Figure 3: Distribution of reach: how many codes are shared by how many languages.

A Uralicist will recognise several of these codes at once, and that is all the check needed: the procedure saw no reconstruction, consulted no etymological dictionary, and grouped forms by their skeleton.

5. What this family makes possible

5.1 Thirty-six branch pairs, instead of one

The Turkic paper established that the substitution matrix must be computed per branch pair, because aggregated over the family it erases precisely what defines the genealogy: there, Chuvash rhotacism went from 1.4 times chance to 6.1 when restricted. But Turkic had only one possible cut, with a single language on one side. The rule was stated in the plural and tested in the singular.

Uralic gives thirty-six pairs with data. The family's primary cut — Samoyedic against the rest — yields this:

Segment	Main	Cases	Share crossing	Share within	Diagnostic
t	s	283 of 970	29%	9%	3.3×
c	t	26 of 44	59%	22%	2.7×
r	s	210 of 759	28%	16%	1.8×
ŋ	n	26 of 95	27%	0%	only when crossing
x	k	93 of 276	34%	26%	1.3×
q	k	82 of 299	27%	28%	1.0×
b	p	64 of 229	28%	63%	0.4×
z	s	58 of 188	31%	77%	0.4×

The deciding column is the last, and it reads better here than in Turkic because there is contrast in both directions. **t** **s** at 3.3× is diagnostic of the cut: when two languages on the same side differ in that position, t corresponds to s 9% of the time; crossing the cut, 29%. By contrast **b** **p** and **z** **s** come out below 1.0, that is, they are *more* frequent within a branch than across: they are internal alternations, and the column exposes them without anyone needing to know Uralic.

That capacity to rule out is what Turkic could not show, because with a one-language branch there was no "within" to compare against.

And a warning inherited from Turkic: the "cases" are pairs of forms, not independent observations.

5.2 Sibling codes: what the catalogue partitions, and what only looks like it

The Turkic paper found that searching the catalogue for pairs of codes for the same concept differing in a single position yields groups of languages partitioned without any genealogical label. Here it is a declared phase (`analysis/fases/codigos_hermanos.py`) and carries a filter it lacked there.

The filter. For a pair to partition languages into two groups it must be disjoint: if a language appears in both codes, it partitions nothing. Obvious put that way, and it was not when the first version of this section was written, which showed five examples of which three failed it.

Of the 153 pairs of binaries and ternaries: 114 disjoint and 39 overlapping (25%).

Overlapping does not mean wrong, and classifying them all as error was the other failure of the first version. A language having two forms for a concept is normal:

Cause	Pairs	What it is
Another form	24	the language has another word — Finnish "cover" = *kattaa* and *peittää* — or the same stem with another suffix — Finnish "here" = *täällä* and *tässä*, Komi "near" = *matın* and *matiꞑ*. Facts of the language
Transcription	15	the same word written two ways by two sources: Estonian *nahk* / *naxk*, *täht* / *tæxt*, Khanty *kuːl* / *qul*. This one is an artefact

The two separate by edit distance. What cannot be separated without a morphological analyser is the first category from within: in a family with fifteen cases, two forms of the same stem with different suffixes are constant, and calling them all ”synonymy” would be a lie. They are left together and it is said.

The disjoint pairs of widest reach, with the partition they produce:

Concept	Code A	Code B	Group A	Group B
name	`l·m` (4)	`n·m` (20)	Mordvin, Mari	Finnic, Saami, Permian, Khantyic and others
FIR	`k·s` (9)	`k·z` (8)	Saami, Finnic, Mari	all of Permian, Mordvin, two Finnic
nose	`n·n` (12)	`n·r` (5)	Finnic, Saami	all of Permian, Mari
NEST	`p·s` (8)	`p·z` (8)	Saami, Finnic	Finnic, Mordvin, Permian

Two of the four put all of Permian on one side. That is the observation, and this paper goes no further: the catalogue partitioned those languages that way, knowing no classification. What the partition means is a question for someone with the family’s historical phonology to hand.

One case worth telling, because it is the instrument splitting rather than gathering. `t·n|THOU` has ten languages and not one Finnic: Finnish says *sinä* and lands in `s·n`, where it shares a code with `s·n|VEIN`. In a single entry sit the gain and the cost: the catalogue isolates a group and at the same time merges two words.

What must be withdrawn from the first version. It claimed that `k·l/q·l/x·l` for ”fish” splits the family into three blocks. False: Khanty and Mansi have forms in all three, and the classifier marks them as transcription, which is what they are — *kuːl*, *qul*, *χul* are the same Khanty word written by three sources. Likewise `v·r/w·r` for ”blood”, with Estonian, Finnish and Hungarian in both.

The protocol lesson, corrected. The search works, but not without the filter: a quarter of the pairs partition nothing, and they are the most striking ones. And the breakdown of those 39 is a result in itself: fifteen are different transcriptions of the same word, a measure of how many incompatible sources live in the corpus.

5.3 Automatic stripping again makes the result worse

The Turkic paper found that discovering affixes automatically raises grouping and raises the null more. Uralic, also agglutinative and also exclusively suffixing, confirms it:

Step	In group	Null	Advantage	κ	Codes of 4+ consonants
Whole word	33.6%	5.0%	28.7	0.301	36.1%
+ discovered affixes	43.7%	23.3%	20.4	0.266	7.3%

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 $\Delta\kappa = -0.035$. 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.)

Ten points of signal, eighteen of noise. The advantage over chance falls from 28.7 to 20.4, however spectacular the quaternary column. And the same typological proof as in Turkic: the discoverer finds 367 prefixes in a family that has no prefixes. It has not discovered morphology, it has discovered the consonant inventory.

That this happens identically in two unrelated agglutinative families, with different corpora and sources, is what turns the Turkic finding into a property of the procedure rather than an anecdote about one family.

The published catalogue does not use automatic stripping.

5.4 The emergent classes do not appear here either

Transitively closing the substitutions that hold up in the validation half puts all twenty segments reaching the threshold into a single class. Not even the `{d, t}` that Turkic detached.

That is why Appendix A carries no emergent-code column and the TSVs carry it empty. Two consecutive families in which this phase produces nothing usable, and it is worth saying: phase 6 of the protocol worked in Austronesian and has not worked since. Either Austronesian was a fortunate case, or the procedure needs something it does not have.

5.5 Geography, here, works

In Austronesian, languages sharing more codes were closer together. In Turkic it could not be known: with 32 languages of a young family, 467 of the 496 pairs fell into one bucket and the design had no dynamic range.

Uralic gives the complete relation, with all six buckets populated:

Codes shared	Pairs	Mean distance	÷ chance
1	2	3,512 km	2.34
2	6	2,675 km	1.79
3–5	35	2,303 km	1.54
6–10	59	2,094 km	1.40
11–20	85	1,640 km	1.09
21+	162	1,066 km	0.71
two random languages	40,000	1,498 km	1.00

Figure 4: Reach against enrichment. Each point is a binary or ternary code; the vertical axis is logarithmic.

Monotonic across all six buckets, from 2.34 to 0.71. And by groups, compactness also decays with reach: 0.48 for codes of 3–5 languages, 0.64 for 6–10, 0.76 for 11–25.

The small buckets hold two and six pairs, so their extreme values sustain nothing on their own; what sustains the reading is the monotonicity across the six, and that the three large buckets — 59, 85 and 162 pairs — already show it.

This closes the question the Turkic paper left open. There we wrote that geography ”says less, it does not say nothing”, unable to tell whether the signal was weak or the design was bad. With Uralic it is clear it was the design: in a family with range of reach, the relation appears cleanly.

5.6 Three families, and what three points allow

With the series' three families measured by the same code and the same definition (`ci/metricas.py`):

Family	Lgs	Arity	Precision	Chance	× chance	Conventional depth
Turkic	32	2.85	25.52%	0.118%	216×	2,000 years
Uralic	27	3.27	6.50%	0.083%	78×	4,000–6,000
Austronesian	582	2.78	13.45%	0.363%	37×	5,200

The chance ratio orders by depth: 216 > 78 > 37, exactly from shallowest to deepest. Observed precision does not: Austronesian (13.45%) exceeds Uralic (6.50%) while being deeper.

The Turkic paper first argued that the ratio measured time, and its reviews overturned that because two thirds of the difference from Austronesian came from the denominator. With the third point the situation is more interesting and still unresolved: the ordering appears in the ratio and not in the numerator, which is exactly what one would expect if the denominator were doing real work — correcting for skeleton specificity in each family — rather than merely adding noise.

But three points are still not a curve, and at least three variables covary with depth and are uncontrolled: the number of languages (32, 27, 582), arity (2.85, 3.27, 2.78) and source type. The honest course is to leave it there: the ordering is a fact about three points, and `docs/PERFIL-FAMILIAS.md` holds sixty families profiled with the same metric where it can be checked. That is the next task, and we have not done it.

Note of 2026-09-09. The series now has eight families measured with the same code (`pipeline/tools/tabla_serie.py`), and the three-point ordering does not survive them: Afro-Asiatic gives 44× and Sino-Tibetan 36×, below Uralic, without being shallower. The chance ratio does not order by depth. The paragraph is left as it stood, with its figures corrected, because it was what three points allowed one to say; what eight allow one to say is that it does not hold.

6. Limits

The instrument does not tell apart etyma that converge in the skeleton. Four pairs among the appendix's hundred codes: *kala/*kele, *tule/*talwe, *neljä/*ńōle, *kolme/*kylmä (§2.3).

No zero reflexes. The procedure of §5.1 compares equal-length skeletons, so it cannot see any correspondence involving segment loss. In Uralic that leaves out any correspondence in which a segment is lost.

A single macroarea, hence no geographic-anomaly list.

Coverage is low: 16.8% of forms enter a code.

18% of the corpus's forms never reach the catalogue for lack of IPA segmentation (§3.1). It costs no concepts, but it is 18% we have not looked at.

We did not calibrate against expert cognacy, though it exists. Uralic has published cognate sets and is among the best-reconstructed families in the world. As in Turkic, this is a shortcoming, not an impossibility.

The emergent classes have gone two families without appearing (§5.4), and we do not know whether the problem is the phase or was the families.

7. Conclusion

The Uralic code catalogue groups 13,470 forms into 1,486 codes with a precision 78 times above chance, and its codes of widest reach are the ones any Uralicist recognises at once. It sought none of them.

What this family contributes to the series is having been able to test what Turkic could only state. The rule of computing substitutions per branch pair was tested there on one cut and here on thirty-six, and here its capacity to rule out is also visible: `b` `p` and `z` `s` come out below 1.0 and are marked as internal alternations without anyone needing to know Uralic. Geography, which in Turkic could not be measured for want of range, here gives a monotonic gradient across all six buckets. And the sibling-code search, which in Turkic arrived late, is here a standard phase and produces the work's best result: PU *kala split into three codes reproducing the division between western Uralic and the Ob-Ugric–Samoyedic block, without the procedure knowing that division exists.

And two confirmations that count because they are in a second family and not the same one: automatic affix stripping again worsens the advantage over chance, and again finds hundreds of prefixes where there are none; and high enrichment still does not mean antiquity.

An instrument that claims little can be applied where there is no literature. But it is worth applying it first where the family has the structure the instrument needs — because only there does one learn whether what the instrument fails to find is its own fault or the data's.

References

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

Dellert, J., Daneyko, T., Münch, A. *et al.* (2020). NorthEuraLex: a wide-coverage lexical database of Northern Eurasia. *Language Resources and Evaluation* 54.

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.

Häkkinen, J. (2009). Kantauralin ajoitus ja paikannus: perustelut puntarissa. *Suomalais-Ugrilaisen Seuran Aikakauskirja* 92.

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

Janhunen, J. (2009). Proto-Uralic — what, where and when? *Suomalais-Ugrilaisen Seuran Toimituksia* 258.

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).

Sammallahti, P. (1988). Historical phonology of the Uralic languages. In D. Sinor (ed.), *The Uralic Languages*. Brill.

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/uralic/output/catalogo-uralic.tsv` — 1,486 rows
Per-form detail	`families/uralic/output/catalogo-uralic-lenguas.tsv` — 7,886 rows
Hygiene (§3.1)	`analysis/fases/fase0_higiene.py` → `runs/fase0-higiene.md`
Data decisions (§3.1)	`ci/decisiones.py`
Data loading, population, warnings	`ci/familia.py`
Headline figures (§4.1, §5.6)	`ci/metricas.py`
Construction	`analysis/fases/catalogo_familia.py`
Grouping and nulls (§3.3, §4.1)	`analysis/fases/agrupamiento_familia.py` → `runs/agrupamiento.md`
Stripping (§5.3)	`analysis/fases/fase2_despiece.py` → `runs/fase2-despiece.md`
Aggregated substitutions (§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`
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`
Cross-family contamination audit	`analysis/tools/audita_familias.py`

The five reports under **`runs/`** declare the same population (§3.4), and `verifica_paper.py` fails if they differ.

Data sources: Lexibank (List, Forkel, Greenhill et al. 2022, CC-BY-4.0), IDS (Key & Comrie, CC-BY-4.0), NorthEuraLex (Dellert et al. 2020, CC-BY-SA-4.0), Concepticon and Glottolog (CC-BY-4.0).

Appendix A · One hundred annotated codes

The catalogue has 1,486 codes and the complete data is in `families/uralic/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. As in Turkic, the transitive closure of substitutions produces no clean classes in this family, and transplanting the Austronesian ones would be convenient and false: a class's value is relative to the family that produced it.

One warning that governs the reading, and which weighs less in Uralic than in Turkic without disappearing: 59 % 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

In Uralic this table is less dramatic than in Turkic, and the reason is arity: 3.28 consonants per form against 2.85. With one more consonant on average there is more room to distinguish, and codes spread across fewer concepts.

The case worth looking at is `k·l`, which appears with "fish" and with "tongue (organ)". It is neither collision nor colexification but two distinct etyma converging in the skeleton: PU *kala "fish" and PU *kele "tongue". The instrument joins them because it sees only `k` and `l`; Finnish separates them effortlessly — *kala* against *kieli* — because it keeps the vowels. It is the cleanest illustration of the cost, and there is no need to go further.

Code	Concepts it appears with, and their enrichment
`k·l`	fish (19, 170×) · language (13, 121×) · tongue (10, 89×)
`p·l`	half (15, 153×) · side (11, 112×) · ear (10, 98×)

Numerals

Only two numerals make the cut, and one should read no more into that than is there: the appendix's cut is the widest-reaching codes, not a sample of the numeral lexicon.

What the two blocks do show is the usual thing: the vowel varies and the skeleton does not. "Four" runs through *neljä*, *njeallje*, *ńil'*, *ńel'*, *nel* without leaving `n·l`. And `n·l` also appears with "arrow" — two different words under one code, which is the cost §2.3 declares.

`n·l|four` · binary · 17 languages · enrichment 204×

Erzya Mordvin nʸilʸe	Estonian neli	Hill Mari nɯl	Inari Sami neʎi
Khanty nʸal	Komi ɲoʎ	Komi–Permyak ɲoʎ	Livonian nɛːʎɑ
Mansi niːla	Mari 'nəl	Moksha ɲilʲe	North Karelian nɛlːæ
Northern Saami ɲe̮æʎːe	Olonets Karelian nɛlːi	Skolt Sami nɛʎːɐ	Udmurt nʸɪlʸ

`k·l·m|three` · ternary · 10 languages · enrichment 239×

```
Erzya Mordvin kolmo     Estonian kolm        Finnish kolme  Inari Sami kulmɐ
Kildin Sami koːlːm      Livonian kuolm       Moksha kolma   North Karelian kɔlmɛ
Olonets Karelian kɔlmɛ  Southern Sami kulmɛ
```

Body parts

The most solid block of the appendix, and where Uralic most resembles itself.

v·r|blood (19 languages, 343×) runs through *veri*, *vïr*, *vur*, *vər* in six branches. And here is something the block does not show and §5.2 does: a second code exists, w·r, with five languages — but Estonian, Finnish and Hungarian are in both, and their forms are *veri* and *weñi*. It is not a split of languages: it is two sources writing the same word.

s·l·m|eye gathers *silmæ*, *sʲelʲmʲe*, *ʧalme*, *ʧaːʎːm* — the initial sibilant variation that §2.1 merges into s. And n·n|nose has a sibling in the catalogue, n·r, with the Permic languages and Mari: two different words for "nose", which the catalogue separates (§5.2).

v·r|blood · binary · 19 languages · enrichment 343×

```
Erzya Mordvin verʲ    Estonian veri          Hill Mari vɯr     Hungarian veːr
Inari Sami vorːɐ      Ingrian veri           Komi vir          Komi–Permyak vir
Livonian vɛr          Lule Sami vɑrːɑ        Mari vyr          North Karelian vɛri
Northern Saami vɑrːɑ  Olonets Karelian vɛri  Skolt Sami vərːɐ  Udmurt vir
```

s·l·m|eye · ternary · 14 languages · enrichment 352×

```
Erzya Mordvin sʲelʲmʲe  Estonian silm          Finnish silmæ
Inari Sami t͡ʃɑlme      Ingrian silmæ          Kildin Sami t͡ʃaːʎːm
Livonian siːlmɑ         Lule Sami t͡ʃɑlmːɛ     Moksha sʲelʲmʲe
North Karelian silmæ    Olonets Karelian silmʏ Skolt Sami t͡ʃelʲmːʲə
Southern Sami t͡ʃɛlmiɛ  Veps silʲm
```

s·n|VEIN · binary · 13 languages · enrichment 98×

```
Erzya Mordvin san  Finnish suo̯ni        Inari Sami suonɐ       Kildin Sami suːnː
Komi sən           Komi–Permyak sən      Livonian suːoŋ         Mari ʃøn
Moksha san         North Karelian sʊɔni  Olonets Karelian sʊɔni  Skolt Sami suənːɐ
Veps sɔnʲ
```

m·k·s|liver · ternary · 12 languages · enrichment 501×

```
Erzya Mordvin mɑksɔ   Estonian maks          Finnish maksa          Hill Mari mokʃ
Ingrian mɑksɑ         Livonian mɑksɑː         Mari 'mokš             Moksha maksa
North Karelian mɑksɑ  Olonets Karelian mɑksʊ  Southern Sami mʉɛksiɛ  Veps mɑks
```

n·n|nose · binary · 12 languages · enrichment 165×

```
Estonian nina          Finnish nenæ       Inari Sami ŋune       Ingrian nenæ
Kildin Sami ŋuŋː       Livonian nɑnɑː     North Karelian nɛnæ   Northern Saami ŋunni
Olonets Karelian nɛnæ  Skolt Sami ŋuːnːʲə  Southern Sami ŋʉɛniɛ  Veps nɛnɑ
```

s·r·v|horn · ternary · 11 languages · enrichment 709×

Estonian sɑrv	Hungarian sarv	Inari Sami t͡ʃu̯ærvi
Ingrian sɑrvi	Kildin Sami t͡ʃuerːv	Lule Sami t͡ʃʊ̹̈ɔ̯ɑrvːɛ
North Karelian sɑrvi	Northern Saami t͡ʃo̯ɑrɛvi	Olonets Karelian sɑrvi
Skolt Sami t͡ʃuarʲvːʲə	Veps sɑrʲv	

p·l|ear · binary · 10 languages · enrichment 98×

Erzya Mordvin piʎe	Inari Sami peʎi	Khanty paɬ	Komi peʎ
Komi-Permyak peʎ	Mansi paʎ	Moksha pilʲe	Northern Saami pe̮æʎːi
Skolt Sami pɛʎːʲə	Udmurt peʎ		

k·l|tongue · binary · 10 languages · enrichment 89×

Erzya Mordvin keʎ	Estonian keel	Finnish kieli	Ingrian keːli
Livonian kɛːʎ	Moksha kalʲ	North Karelian kiɛli	Olonets Karelian kiɛli
Udmurt kɯl	Veps kɛlʲ		

Kinship and person

n·m|name, at 378×, is the highest enrichment in this section, and it deserves a look because it explains the mechanism. Twenty languages from seven branches say *nimi*, *nim*, *nam*, *ɲim*, *nomːɐ* — over two nasal consonants, which are among the most frequent in Uralic.

That enrichment does not come from n·m being a rare sequence but from the fact that twenty of the languages attesting ”name” say it the same way: coverage is nearly total, and enrichment is coverage divided by base frequency. No rarity is needed to reach that figure; it is enough for a word to be well distributed in a well-sampled corpus.

The pronouns come out adjacent: m·n|I (*minä*, *mon*, *mun*, *mænʸ*) and t·n|THOU (*ton*, *tun*, *tənə*). Note that **t·n** has no Finnic language: Finnish says *sinä* and sits in another code. The catalogue puts both halves on the table, separated, without knowing that is what they are.

n·m|name · binary · 21 languages · enrichment 391×

Estonian nimi	Finnish nimi	Forest Enets nim	Inari Sami nom:ɐ
Khanty neːm	Kildin Sami neːmː	Komi nim	Livonian niˀm
Lule Sami nɑmːɑ	Mansi nam	North Karelian nimi	Northern Saami nɑmmɑ
Olonets Karelian nimi	Skolt Sami nəmːɐ	Southern Sami numːɛ	Udmurt nʸim

m·n|I · binary · 18 languages · enrichment 193×

Erzya Mordvin mon	Estonian mina	Finnish minæ	Inari Sami mun
Khanty mæn	Kildin Sami munn	Livonian minɑː	Lule Sami mɔn
Nenets manʸ	Nganasan mənə	Northern Saami mun	Olonets Karelian minæ
Selkup mɑn	Skolt Sami monɐ	Southern Sami manːɛ	Udmurt mon

t·n|THOU · binary · 10 languages · enrichment 153×

Erzya Mordvin ton	Hill Mari tɯnʲ	Inari Sami tun	Kildin Sami toːnː
Moksha ton	Nganasan tənə	Northern Saami tɔn	Selkup tan
Skolt Sami tonɐ	Udmurt ton		

Nature and sky

`m·r|sea`, `t·l|fire`, `l·m|snow`. Three binaries of medium-high reach.

`m·r|sea` deserves a look: alongside Finnish *meri* and Estonian *meri* appear *mɔrʲɛ*, *morʲe*, *morʲa* in Enets, Komi-Permyak, Selkup, Erzya and Moksha. They are visibly two different provenances within a single code of reach 16, and §5.5 explains why the instrument cannot separate them.

`t·l` also appears with ”winter”: *tuli* and *talvi* in Finnish, two words that only the vowel distinguishes.

`m·r|sea` · binary · 17 languages · enrichment 249×

Erzya Mordvin morʲa	Estonian meri	Finnish meŕi	Forest Enets mɔrʲɛ
Inari Sami me:rɐ	Kildin Sami mierr	Komi–Permyak morʲe	Livonian mɛˀr
Lule Sami mɛr:ɑ	Moksha morʲa	North Karelian mɛri	Northern Saami meɑŕŕɑ
Olonets Karelian mɛri	Selkup morʲe	Skolt Sami miɐr:ɐ	Southern Sami meæruʌ

`t·l|fire` · binary · 16 languages · enrichment 164×

Erzya Mordvin tol	Estonian tuli	Finnish tuli	Hill Mari təl
Inari Sami tul:ɐ	Ingrian tuli	Kildin Sami to:l:	Livonian tuˀʎ
Mari 'tul	Moksha tol	North Karelian tʊli	Northern Saami tɔl:ɑ
Olonets Karelian tʊli	Skolt Sami tol:ɐ	Southern Sami tɔl:ɛ	Udmurt tɯl

`p·l|HALF` · binary · 15 languages · enrichment 153×

Erzya Mordvin pelʲ	Estonian pool	Finnish puoli	Hill Mari pel
Inari Sami pe:li	Kildin Sami pe:ʎ:	Livonian pu:oʎ	Mansi pal
Mari 'pel	Moksha palʲe	Nenets pʲe:ʎɑ	North Karelian pʊɔli
Northern Saami pe̮æl:i	Skolt Sami piɛl:ʲə	Veps pɔlʲ	

`l·m|snow` · binary · 12 languages · enrichment 175×

Estonian lumi	Finnish lumi	Hill Mari ləm	Ingrian lumi
Komi lɯm	Komi–Permyak lɯm	Livonian lum	Mari 'lum
North Karelian lʊmi	Olonets Karelian lʊmi	Udmurt lɯmɯ	Veps lʊmi

`p·l|SIDE` · binary · 11 languages · enrichment 112×

Estonian pool	Finnish puoli	Kildin Sami pe:ʎ:
Mansi pal	Moksha palʲe	North Karelian pʊɔli
Northern Saami pe̮æl:i	Olonets Karelian pʊɔli	Southern Sami piɛliɛ
Udmurt pal	Veps pɔlʲ	

`t·l|WINTER` · binary · 11 languages · enrichment 117×

Erzya Mordvin tʲelʲe	Estonian talw	Finnish talwi	Hill Mari tel
Hungarian teel	Khanty tal	Livonian tɔ:lɑ	Mansi ta:l
Mari 'tele	Moksha tʲala	Udmurt tol	

`k·l·m|cold` · ternary · 10 languages · enrichment 239×

Erzya Mordvin keʎme	Estonian kylm	Finnish külmæ	Ingrian kylmæ
Kildin Sami ke:l:m	Livonian ki:lmɑ	Moksha kelʲmʲe	North Karelian kylmæ
Olonets Karelian kʏlmʏ	Southern Sami kalmɛ		

Animals

`b·t|louse`, `q·s|bird`, `q·r·t|worm` and `b·l·q|fish`: two binaries and two ternaries.

`k·l|fish` is the block that teaches most about what the appendix cannot show. Here one sees 19 languages saying *kala*, *kuolli*, *kul*, *kol*. What one does not see is that the catalogue also has `q·l` and `x·l`, with three languages each — and that Khanty and Mansi are in all three, with *kuːl*, *qul* and *χul*. They are three transcriptions of the same word spread across three codes, not three groups of languages. §5.2 details it, and it is the example from which the first version of this paper drew a false conclusion.

`k·l|fish` · binary · 19 languages · enrichment 170×

Erzya Mordvin kɑl	Estonian kala	Finnish kala	Hill Mari kol
Inari Sami kyeli	Ingrian kɑlɑ	Kildin Sami kuːʎː	Livonian kɑlɑː
Mansi kuːɬ	Mari 'kol	Moksha kal	North Karelian kɑlɑ
Northern Saami kʊ̯ɔlːi	Olonets Karelian kɑlɑ	Skolt Sami kuɛlːʲə	Southern Sami kʉɛliɛ

`m·n|egg` · binary · 17 languages · enrichment 182×

Estonian muna	Finnish muna	Forest Enets mɔna	Hill Mari mənə
Inari Sami mɑne	Ingrian munɑ	Khanty moŋ	Kildin Sami mannʼ
Lule Sami mɔnːɛ	Mansi muŋi	Mari 'muno	Nganasan mənu
North Karelian mʊnɑ	Northern Saami monni	Skolt Sami mɐːnːʲə	Southern Sami mʉnːiɛ

`k·k|CUCKOO` · binary · 10 languages · enrichment 211×

Erzya Mordvin kuko	Finnish kæki	Hill Mari kuku	Komi kək
Komi–Permyak kək	Mari kuku·	Moksha ku·ku	North Karelian kæki
Southern Sami kɨɛkɛ	Udmurt kikɨ		

`p·n|dog` · binary · 10 languages · enrichment 134×

Erzya Mordvin piɲe	Kildin Sami peːnːe	Komi pon	Komi–Permyak pon
Livonian piɲ	Moksha piɲe	Northern Saami pe̮æ̮nɑ	Skolt Sami piɐnːɒi
Southern Sami pɨɛɲɛ	Udmurt punɯ		

Plants and food

`s·l|salt` and `j·r|root`. Of the second, one should say what the list shows and nothing more: its ten languages are Finnic, Mordvin and one Khanty. There is no Saami or Permic language, however much the first version of this comment said there was — a reminder that the printed list is right below and can be counted.

`s·j·n|CHEW`, at 823×, is the block's highest enrichment: a ternary of infrequent consonants, which is the combination §3.2 predicts.

`s·j|TEA` · binary · 15 languages · enrichment 188×

Erzya Mordvin t͡ɕaj	Forest Enets t͡ʃaj	Hill Mari t͡ʃæj	Khanty ʃaj
Kildin Sami t͡ʃajː	Komi t͡ɕɑj	Komi–Permyak t͡ɕaj	Mansi sʲaːj
Mari t͡ʃɑj	Moksha t͡ɕaj	Nenets sʲɑj	North Karelian t͡ʃæijy
Olonets Karelian t͡ʃʊɑjʊ	Selkup t͡ɕoj	Udmurt t͡ɕɑj	

k·s|BRANCH · binary · 12 languages · enrichment 55×

Estonian oks	Finnish ɔksɑ	Hill Mari ukʃ	Inari Sami u̯æksi
Livonian oksɑ:	Lule Sami ůɔɑksɛ	Mari ukʃ	North Karelian ɔksɑ
Northern Saami o̯ɑksi	Olonets Karelian ɔksʊ	Southern Sami oæksiɛ	Veps ɔks

k·r|SKIN (OF FRUIT) · binary · 12 languages · enrichment 122×

Estonian ko:r	Finnish kuo̯ri	Inari Sami kor:ɐ	Khanty kar
Komi kor	Livonian ku:orʲ	North Karelian kʊɔri	Northern Saami kɑr:ɑ
Olonets Karelian kʊɔri	Skolt Sami kər:ɐ	Southern Sami kɨr:ɛ	Veps kɔrʲ

s·l|salt · binary · 12 languages · enrichment 159×

Erzya Mordvin sɑl	Estonian sool	Finnish suola	Ingrian so:lɑ
Khanty sal	Kildin Sami su:ʎ:	Livonian su:oʎ	Mansi saɬ:ə
Moksha sal	North Karelian sʊɔlɑ	Olonets Karelian sʊɔlʊ	Veps sɔl

j·r|root · binary · 10 languages · enrichment 249×

Erzya Mordvin jur	Estonian juur	Finnish ju:ri	Ingrian ju:ri
Khanty jo:r	Livonian ju:rʲ	Moksha jur	North Karelian ju:ri
Olonets Karelian ju:ri	Veps jʊrʲ		

Objects and culture

This is the section to read with §5.5 in front of you, because it contains the highest enrichments in the whole appendix and none is what it seems:

l·m·p|TORCH OR LAMP 2964× (*lampa*, *lamp*, *lamppu*) · k·n·g|BOOK 1154× (*kniga*, *kŋiga*) · s·m·k|BAG 933× (*sumka*) · r·s·t|CROSS 855× (*risti*, *rist*) · l·n·t|RIBBON 477× (*lenta*).

They are the five highest-enrichment codes in the appendix, and they are visibly material these languages share with Russian. The instrument cannot distinguish them from an inherited word — both are "many languages saying the same thing" — and that is what §5.5 is about. Lined up like this, they are the best available illustration that a high enrichment says nothing about a word's age.

k·r|bark shares a skeleton with k·r|SKIN (OF FRUIT), and they are the same word used for two things.

n·l|ARROW · binary · 16 languages · enrichment 199×

Erzya Mordvin nal	Estonian nool	Finnish nuoli	Hungarian ɲiil
Inari Sami ɲuolɐ	Khanty nʸul	Kildin Sami ɲu:l:	Livonian nu:oʎ
Mansi nʸol	Moksha nal	North Karelian nʊɔli	Northern Saami ɲuollɑ
Skolt Sami ɲuəl:ɐ	Southern Sami ɲuʌlɛ	Udmurt nʸël	Veps nɔl

k·r|bark · binary · 16 languages · enrichment 157×

Erzya Mordvin kerʲ	Estonian ko:r	Finnish kuoři	Inari Sami kor:ɐ
Ingrian ko:ri	Khanty ka:r	Kildin Sami ke:r:	Livonian ku:orʲ
Mansi kæ:r	Moksha ker	North Karelian kʊɔri	Olonets Karelian kʊɔri
Skolt Sami kər:ɐ	Southern Sami kɨr:ɛ	Udmurt kur	Veps kɔrʲ

l·n·t|RIBBON · ternary · 14 languages · enrichment 477×

Erzya Mordvin lʲenta	Estonian lʲint	Forest Enets ʎɛnt	Kildin Sami ʎaːn̥ːt
Komi ʎentɑ	Livonian liːntɑ	Mansi ʎeːnta	Mari ʎeˑntʲe
Moksha lʲenta	Nenets ʎeːntɑ	Olonets Karelian lɛntʊ	Selkup ʎenta
Skolt Sami lieɲtːɐ	Udmurt lʲentɑ		

s·l·m|KNOT · ternary · 13 languages · enrichment 339×

Erzya Mordvin sʲulmo	Estonian səlm	Finnish solmu	Inari Sami t͡ʃuolmɐ
Kildin Sami t͡ʃuːlːm	Livonian suoʎm	Lule Sami t͡ʃʊʊ̯̂lmːɑ	Moksha sʲulma
North Karelian sɔlmʊ	Olonets Karelian sɔlmi	Skolt Sami t͡ʃuəlmːɐ	Southern Sami t͡ʃuʌlmɛ
Veps sɔlʲm			

k·l|LANGUAGE · binary · 13 languages · enrichment 121×

Erzya Mordvin kelʲ	Estonian keel	Finnish kieli	Inari Sami kielɐ
Kildin Sami kiːlː	Livonian kɛːʎ	Moksha kalʲ	North Karelian kiɛli
Northern Saami kiɛlːɑ	Olonets Karelian kiɛli	Southern Sami kɨɛlɛ	Udmurt kɨl
Veps kɛlʲ			

k·n·g|BOOK · ternary · 12 languages · enrichment 1154×

Erzya Mordvin kinʸiga	Forest Enets kniga	Kildin Sami knīga	Komi kniga
Komi–Permyak kŋiga	Mari knʲigɑˑ	Moksha kŋiga	Nenets kniga
Nganasan kŋigɐ	Olonets Karelian kniːgʊ	Selkup kŋiga	Udmurt knʸiga

l·m·p|TORCH OR LAMP · ternary · 12 languages · enrichment 2964×

Erzya Mordvin lampa	Estonian lamp	Finnish lamppu	Hungarian laampa
Khanty lampa	Kildin Sami lammp	Komi lampa	Mansi lampa
Mari 'lampe	Northern Saami lampa	Selkup lampa	Udmurt lampa

s·m·k|BAG · ternary · 10 languages · enrichment 933×

Hill Mari sumkə	Kildin Sami sum̥ʲːk	Komi sumkɑ	Komi–Permyak sumka
Mansi sumka	Mari sumkɑˑ	Olonets Karelian sʊmkʊ	Selkup sumka
Udmurt sumkɑ	Veps sʊmk		

r·s·t|CROSS · ternary · 10 languages · enrichment 855×

Estonian rist	Finnish risti	Inari Sami ristɐ	Kildin Sami rɨsst
Livonian riʃt	North Karelian risti	Northern Saami ristɑ	Olonets Karelian ristʊ
Skolt Sami ristːɐ	Veps rist		

Qualities

k·l·m|cold and n·r|YOUNG. The first shares a skeleton with k·l·m|three: two unrelated words under one code, and that makes four such pairs in the appendix — k·l, t·l, n·l and k·l·m.

The insistence is the point. In a family with 59 % collision, two or three consonants are not enough to separate two words, and the reader must carry that into every block.

n·r|YOUNG · binary · 12 languages · enrichment 189×

Estonian noor	Finnish nuoři	Inari Sami nuoɾe	Kildin Sami nuːrː
Livonian nuːorʲ	Lule Sami nʊu̯͡ɔrːɑ	North Karelian nʊɔri	Northern Saami nuořřɑ
Olonets Karelian nʊɔri	Skolt Sami nuərːɐ	Southern Sami nuʌrɛ	Veps nɔrʲ