Consonantal codes in 32 Turkic languages

A catalogue built without genealogy, and what a young family teaches

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

Abstract

We present a catalogue of 1,517 consonantal codes grouping 24,517 lexical forms from 32 Turkic 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-Turkic is used.

The procedure is right 216 times more often than chance (25.5% precision against 0.118% expected), against 35 for Austronesian recomputed with the same code. But this paper's work is to decompose that figure rather than celebrate it: two thirds of the difference live in the denominator, which depends on the family's word length and not on its history. Observed precision differs by only 1.9×, and the 216× is now exact on the chance side. What holds is that two families' chance ratios are not comparable to each other, and that every comparison is published decomposed.

Four results about the instrument follow. Morphological stripping pays when the rule is declared and worsens when discovered by resemblance: the automatic discoverer gains 6.3 points of grouping and 13.9 of null, and finds 365 prefixes in a family with no prefixes. The substitution matrix, aggregated over the family, loses the correspondence defining its primary division and recovers it only when restricted to the branch pair: from invisible to 6.1 times chance. The catalogue separates branches without knowing they exist — 297 sibling code pairs, the largest of which group Chuvash with Siberian and Oghuz with Western Kipchak. And maximum enrichment is a loanword again, here Russian, by a nameable mechanism: it is produced by the rare concept, not the rare code.

The instrument's cost is declared: 59% of Turkic forms share their skeleton, within their own language, with a form of another concept. The code `k·l` gathers ten concepts, among them *kül* "ash", *kel-* "come", *köl* "lake" and *kül-* "laugh".

1. What is sought, and why

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

This paper asks a different question: what the figures that one produced mean. For that the instrument must be applied to a family whose history *is* known, and Turkic serves for three reasons: known and shallow time depth, well-established diagnostic correspondences to check against, and strongly agglutinative morphology, which is where the previous work's declared limit had to hurt.

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.

And for this family in particular: nothing that follows bears on the Altaic hypothesis, neither supporting nor refuting it. It never leaves the family, and where the material touches the correspondences that debate uses — `z` `r`, `š` `l`, the pronominal `-z` — the distribution is described without assuming direction of change or morphemic status.

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 Turkic

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. In Turkic it matters especially to know that displaced sibilants and affricates are written **s** (š, č, ś, ʧ → s), that the velars split across k, q, g and x, and that y is written j.

Written	as	Written	as
s š ś ş ɕ z ž ʒ and every affricate (č ʧ ts dž…)	s	n ŋ ɲ ɳ ɴ	n
l ɬ ɬ̵ ɭ ʟ	l	r ɾ ɻ ɻ̢ ɾ ʀ ʁ	r
x ç ɣ χ	x	g ɟ ɡ ɢ	g
h ħ ɦ ʔ ʕ	h	m ɱ	m · d ɗ → d

Thus Kazakh *bes*, Turkish *beš*, Uzbek *bɛʃ* and Khalaj *bi:eʃ*, all "five", give the same skeleton b·s.

This table has a cost worth seeing before the results, and in Turkic it is larger than in Austronesian. By merging every sibilant into s, the normalisation makes invisible one of the three correspondences defining the family's primary division: lambdacism, Chuvash *l* against Common Turkic *š*. Since š is written s, that pair is diluted among all the family's *s*-es and §5.3 cannot see it. It is declared here and taken up there.

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 `b·s|five`, never `b·s = five`: the code does not mean "five", it is the shape "five" repeatedly takes.

The concept requirement is not decorative. Without it, the skeleton groups by resemblance of form and gathers unrelated words; with it, the grouping has to get two things right at once.

2.3 The instrument's cost: collision

Here something must be said that Austronesian allowed us to keep in the background and Turkic does not.

59% of Turkic forms share their skeleton, within their own language, with a form of another concept. Not across languages: within a single one. It is the clean measure of what the skeleton fails to distinguish, with no genealogy involved, and it is high because Turkic words are short: 2.85 consonants on average. (The unit matters: it is 59% of forms; counted over triplets of language, skeleton and concept it is 55%, and those are two different populations.)

The clearest example sits in the appendix's reverse index. The code `k·l` appears with ten concepts, of which the appendix shows the four of widest reach: "ash", "come", "lake" and "laugh" — in Turkic *kül*, *kel-*, *köl* and *kül-*. Three are separable by the vowel, which is exactly what the skeleton does not record; and the fourth is better still, because **`kül`** "ash" and **`kül-`** "laugh" are homophones already in Old Turkic: not even a procedure preserving vowels would separate them. Part of the collision is a fact of the language, not a cost of the instrument.

Beside it sits a case that looks the same and is not. `k·n` appears with "day" and with "sun", but there *kün* is the same word used for both. That is not collision but colexification, and it is a fact of the language, not a defect of the instrument.

The code alone does not tell the two apart. It may gather one word used for two things or two words the instrument confuses, and one must look at the forms to know which. This warning governs the reading of the whole catalogue.

2.4 Arity

A code may have one consonant (unary), two (binary), three (ternary) or more. A unary supports no claim on its own: there are few consonants and any of them coincides by chance. A code of four or more, by contrast, is grounds for doubting one's own analysis, because it usually means the morphology has not been separated. In Turkic that suspicion proved right, and §3.2 resolves it.

Arity	Codes	
1 (unary)	185	12.2%
2 (binary)	693	45.7%
3 (ternary)	457	30.1%
4 or more	181	12.0%

The figures this paper defends are restricted to the 1,152 binaries and ternaries.

Figure 1: Distribution of codes by arity. 88% have three consonants or fewer — but before the stripping of §5.2 those with four or more were 29%.

3. Method

3.1 Data

Integrative Corpus, Turkic subset: 32 languages, 54,875 forms with transcription, assigned concept and skeleton. The sources are standardised wordlists — Lexibank, IDS and NEL — under CC-BY licences, and the concept is a Concepticon node.

Three clarifications about those figures, all from the phase-0 audit (`runs/fase0-higiene.md`):

The raw corpus holds 70,276 forms and we publish 54,875. The difference is 14,976 forms in Cyrillic script that the pipeline silently discards. We checked that the loss costs no coverage — zero concepts

exist only in Cyrillic: all have their Latin form — but the count published is that of the forms that enter, not of those that exist.

One doculect is excluded. `Crimean Tatar` and `CrimeanTatar` share their skeleton in 83% of their common concepts: they are the same lect transcribed twice under two glottocodes. The one contributing more forms is kept. `Tatar`/`KazanTatar` (64%) and the two Altai (59%) are not excluded: they come from different sources and contribute different data. The 80% threshold is a declared decision.

The family has a single macroarea, so this paper cannot have the geographic-anomaly section the Austronesian one had.

3.2 Stripping the citation form, and why it comes first

This is the work's most important methodological finding and it belongs early.

The sources do not cite verbs alike. `ids` gives them in the infinitive nearly always — *ичмакъ*, *ичм*[?]*к*, *ичмаг* "drink": 85 % of its verbs carry `–mAK`. `lexibank` and `nel` are mixed and carry it in only 7.4 % and 6.4 %, so *œlmæk* "die" sits beside *øl*.

(An earlier version attributed the infinitive to `nel` as well as `ids`. The split was computed by grouping forms on `lect.source_id`, the source of the language record, when the citation form is a property of whoever edited the form: one language gathers forms from several sources. With `form.source_id`, the right column, `ids` goes from 25,739 forms to 11,831 and its infinitive rate rises from 52 % to 85 %. The contract in `ci/esquema.py` exists so that confusion cannot be written again.) The Turkic suffix `–mAK` contributes two consonantal positions to the skeleton, so the same verb came out as `j·ʃ·m·q` in one language and `j·ʃ` in another, and formed no code.

That is 4,974 forms, 9.1% of the family, spread over 24 languages and 541 concepts. It is not the language's morphology: it is what the dictionary's editor added in citing it.

The stripping applies under a double condition — the written suffix *and* the skeleton's tail must both match — so that a root ending in `–mak` by accident survives, and it does not apply if the suffix is the whole word. The rule is declared per family in `ci/decisiones.py`, not discovered automatically: trimming endings that "look like suffixes" is the quick road to manufacturing the codes one wants to see.

3.3 Enrichment

A code appearing in many languages may be remarkable or trivial depending on what it is made of. Enrichment corrects for that: it is the observed coverage divided by the skeleton's frequency in the family. `b·s|five` covers 26 languages with an enrichment of 139×; a code of equal reach made of very frequent consonants would stay in the low single digits.

3.4 Nulls

Every grouping is compared against a permutation breaking the form–meaning relation by shuffling concepts within each language, which preserves each one's inventory. The real grouping covers 56.7% of the forms; the null, 13.7%. Both at a two-language threshold — the catalogue uses three (§4.1).

4. Results

4.1 The catalogue

1,517 codes over 793 of the 1,828 concepts, grouping 24,517 forms.

Magnitude	Value	Over what population
Forms inside a code	44.7% (24,517 of 54,875)	threshold of 3 languages, the catalogue's
Forms in some group	56.7% (null: 13.7%)	threshold of 2 languages, grouping test
Precision	25.5% (chance: 0.118%)	all pairs of forms from different languages
Times above chance	216× [216–216; chance is computed exactly]	idem
Binaries and ternaries	1,152 (75.9%)	of the 1,517 codes
Median enrichment	185×	of the 1,152 binaries and ternaries

Correction of 2026-09-09. This row read 26.2% (14,350 of 54,875), 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 44.7%. The error came from reading the TSV row count as if it were the population's, and it was in the Uralic paper too; the other six use the correct figure. No other row of this table moves, and no conclusion of the paper depends on this one.

Each row carries its population because there are three different ones, and conflating them was an error of this paper's draft: the catalogue's coverage (44.7%) and the grouping test's (56.7%) were cited as if they were the same. The first requires three languages per code; the second, two.

4.2 The codes of widest reach

Concept	Code	Languages	Enrichment
WHO	`k·m`	30	352×
ash	`k·l`	29	119×
long	`z·n`	28	471×
GRASP	`t·t`	28	278×
I	`m·n`	28	224×
see	`k·r`	27	168×
stone	`t·s`	27	102×

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. The median is four languages per code.

5. What this family teaches about the instrument

5.1 The headline figure is a ratio, and both halves must be published

Turkic scores 216 times above chance; Austronesian, recomputed with the same code and the same definition, 35. The temptation is to read that as ”the method works six times better on Turkic”. It would be false, and decomposing the ratio shows why:

	Turkic	Austronesian	Ratio
Observed precision	25.48%	13.40%	1.92×
Chance	0.118%	0.383%	3.35× (the denominator falls)
× chance	216	37	5.89×

Two thirds of the difference live in the denominator. And the denominator is no constant of the universe: it is the probability that any two forms from different languages share a concept, which depends on how specific the skeleton is in that family — that is, on word length.

One convenient explanation can be ruled out at once: it is not the conceptual space. Turkic has 1,828 concepts and Austronesian 1,823. Almost identical.

And the figure's stability must be stated. An earlier version of this paper published 225× with a range of [172–227] across three seeds, and attributed that spread to the data. It was ours: the chance rate was estimated with 80,000 draws, which at a base rate near 0.1 % is some eighty hits, so removing six forms from the 54,881 was enough to move the null by 14 %. Chance is now computed exactly, in closed form over the distribution of concepts and languages, and only the subsampling of large groups remains as a source of variation — which for this family is none: [216–216].

There was a second correction, of reproducibility. The query that loads a family did not order its rows, and PostgreSQL promises no order without `ORDER BY`: the same family loaded in a different order in every process, so the subsampling of large groups drew from a different list each time and the figures could not be reproduced — running twice gave different numbers. With the load ordered by a unique key, the figure is stable across successive runs and settled at 216× (25.48 % precision). Before ordering it hovered around 218×, which is what the previous version published: a single-draw value.

5.1-bis What we cannot conclude, and why it is said here

A draft of this paper argued that the difference measures time depth: Turkic is some two thousand years deep and Austronesian more than five thousand. The decomposition above does not allow that, and the uncontrolled candidates are better enumerated than dismissed by decree:

- Word length, which moves the denominator directly.
- Vowel harmony. In Turkic the vowel is largely predictable from the root, so discarding it costs *less information* than in Austronesian. The instrument throws away precisely the dimension that is redundant in Turkic — a purely structural explanation for high precision, with no time involved.
- The axis of change. Turkic change is mostly vocalic and morphological; Austronesian change is massively consonantal, with mergers and losses. The instrument measures the axis Turkic moves least.
- Intrafamily borrowing, enormous in Turkic: Chagatai over Kipchak, Ottoman over Balkan, Soviet Russian reinjecting lexicon across Kazakh, Karakalpak and Nogai.

- The composition of the sample. 32 doculects, several sharing a skeleton at around 60 % and **`OldTurkic`**, from the eighth century (§3.1). Measured, all of that together is worth 2 % of the figure: 216× → 215×.
- The number of languages: 32 against 582.

The honest conclusion is smaller and more useful than the one we had: two families' chance ratios are not comparable to each other as if they measured the same thing. Each family is compared against its own null, and every cross-family comparison is published decomposed.

5.2 Declared stripping pays; discovered stripping does not

Turkic is agglutinative, so morphology had to hurt here. It does, in a way visible only when the null column is published:

Step	In group	Null	Advantage	κ	Codes of 4+ consonants
Raw word	54.4%	12.5%	41.9	0.479	29.1%
+ declared **`-mAK`** stripping	56.7%	13.7%	43.0	0.498	23.8%
+ discovered affixes	63.0%	27.6%	35.4	0.489	4.0%

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 $\Delta\kappa = -0.009$. 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`, which derives them in one pass so two papers cannot disagree. 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. §5.1 of the Sino-Tibetan paper says so.)

The declared stripping — the **`-mAK`** rule of §3.2 — raises the advantage over chance. Automatic affix discovery lowers it: it gains 6.3 points of signal and 13.9 of noise. By the criterion the procedure itself declares — that grouping rise *above the null*, not that arity fall — the automatic phase fails, however striking its quaternary column.

And there is a second, typological proof needing no number: the automatic discoverer finds 365 prefixes in a family that has no prefixes. It has not discovered morphology: it has discovered the consonant inventory. That is exactly the diagnosis the Austronesian paper had already made of its own detector, and which this work nearly repeated without citing it.

What the published catalogue uses is the middle row: declared stripping, no discovered affixes.

The lesson for the protocol is not "stripping pays in agglutinative families" but something narrower and more reliable: what is declared and justified pays; what is discovered by resemblance manufactures the morphology one wanted to see.

5.3 The diagnostic correspondence drowns if the family is averaged

Turkic divides between Oghur, of which only Chuvash survives, and Common Turkic, everything else. And it has known diagnostic correspondences: Chuvash *r* ↔ Common Turkic *z* (PT $*r_2$) and Chuvash *l* ↔ Common Turkic *š* (PT $*l_2$).

On the names. The literature calls these *rhotacism* and *lambdacism*, or *zetacism* and *sigmatism*, depending on which side is taken as primary — and that choice takes a position in a debate this work does not enter. Here they are written as correspondences, z r and š l, with no direction assumed.

Aggregating substitutions over the whole family, the first does not appear: 1.4 times chance, far below b↔p (31.9×) or d↔t (19.5×). The reason is arithmetical: Oghur has one language and Common Turkic thirty-one, so pairs crossing the division are 6.7% of the total and their signal is buried by the remaining 93.3%.

Restricted to the branch pair, it appears. The deciding column is the last: how much more concentrated the correspondence is when crossing the cut than when not.

Segment	Main	Cases	Share crossing	Share within	Diagnostic
f	v	26 of 30	87%	0%	only when crossing
b	p	2,887 of 3,385	85%	43%	2.0×
d	t	862 of 1,088	79%	85%	0.9×
w	ʊ	230 of 409	56%	0%	only when crossing
z	**r**	550 of 1,006	55%	9%	6.1×
ð	s	32 of 64	50%	9%	5.3×
j	s	1,602 of 3,305	48%	71%	0.7×
ɢ	x	64 of 136	47%	7%	6.6×

z r goes from invisible to 6.1×, and the table shows why it is diagnostic and d↔t is not: when two Common Turkic languages differ in that position, z corresponds to r only 9% of the time; when the cut is the Oghur one, 56%.

Three more rows deserve comment, and two count against us:

- **w** ʊ and **f** **v** do not exist without crossing the cut (0% share within). By the concentration criterion this very section declares, they are the cleanest in the table, cleaner than z r. They correspond to Chuvash labial prothesis (*vut* 'fire' Common Turkic *ot*).
- **ɢ** **x** gives 6.6×, above the rhotacism, and we do not know what it is. With a one-language branch, any transcription peculiarity of that language comes out as diagnostic. We say so rather than celebrating only the row that suited us.
- **ð** **s** (5.3×) is a Bashkir innovation, not an Oghur one: it measures a language of the large side against Chuvash. The "Diagnostic" column does not distinguish that, and that is its limit.

Lambdacism appears not even restricted, and our first diagnosis of why was wrong. We blamed the normalisation of §2.1, which writes both s and š as s. But the main cause is another, and one example shows it: the etymon is PT *bēĺ(k), whose Oghur reflex is Chuvash *pillĕk* → p·l·k, against Common Turkic *beš* → b·s. Different lengths, and the procedure only compares skeletons of equal length.

And a warning about "cases". With one language on one side and thirty-one on the other, each Chuvash form generates up to 31 pairs: the 1,006 cases of z r are not 1,006 independent observations but a few dozen etyma counted many times. The figure says where to look; it is not a test.

5.4 The catalogue separates branches without knowing they exist

This is the result we took longest to see, because it sat in the TSV and in no report.

Searching the catalogue for pairs of codes of the same concept differing in a single position yields 297, of which 184 are disjoint in languages — no language is in both — and only those can be proposed as

an isogloss. The substitutions generating the disjoint ones are `j s` (38 pairs), `d t` (21), `q x` (13), `k q` (12), `k x` (10), `g k` (8).

And the list of the 113 that overlap is worth reading beside it, because the two do not coincide: there `k q` (17), `b p` (13) and `g k` (13) lead. A substitution that dominates the second list and not the first is almost certainly a transcription convention rather than a fact of the language.

The disjoint ones of widest reach are recognisable branch cuts:

Concept			What it separates
one	`b·r` (25 lgs)	`p·r` (7)	Chuvash + Siberian: *per*, *pur*, *pir* against *bir*, *ber*
head	`b·s` (24)	`p·s` (6)	the same: *pas*, *paš* against *baš*, *bas*
five	`b·s` (25)	`p·s` (7)	Siberian, without Chuvash (which says *pillĕk*, §5.3)
four	`t·r·t` (16)	`d·r·t` (14)	Oghuz + Western Kipchak: *dört* against *tört*
see	`k·r` (27)	`g·r` (6)	Oghuz *k-* > *g-*

The seven languages of `p·r|one` are Chuvash, Khakas, Shor, both Altai, Baraba Tatar and West Yugur: the initial devoicing that groups Oghur with Siberian, which no tree gives as such. The fourteen of `d·r·t|four` are Azerbaijani, Gagauz, Kumyk, Karaim, Crimean Tatar, Kazan Tatar and Bashkir: Oghuz plus Western Kipchak.

The catalogue does not know those branches exist. It separates them anyway, because it separates forms.

This qualifies §5.3, where `d↔t` comes out at 0.9× and reads as "a general voicing alternation, not a branch correspondence". With the code pairs in view the correct reading is different: **`d`↔`t`** does not concentrate in the only cut the corpus admits — Oghur against Common Turkic — but it does separate a coherent group. The corpus has only two branch labels (§3.1) and the Oghuz cut could never be tested.

The protocol lesson is that this search should be a standard phase. It comes free from the catalogue, it is conditioned by concept — which is what the aggregated matrix lacks — and it needs no branch labels.

5.5 Maximum enrichment is a loanword again

In Austronesian the highest-enrichment code turned out to be a Gallicism. It repeats, in the same shape:

Code	Concept	Languages	Enrichment	Forms	What it is
`n·s·r·t·r·k`	BANYAN	3	13,720×	*nusrat-terek*	maximum of the whole catalogue
`f·r·m·r`	FARMER	3	8,232×	*fermer*	Russian *фермер*
`p·l·m`	PALM TREE	3	7,840×	*palma*	Russian *пальма* · max among binaries and ternaries
`m·n·t`	COIN	4	7,840×	*manat*, *mänet*	Russian *монета*

Superlatives carry their population: 13,720× is the maximum of the catalogue; 7,840× only of binaries and ternaries, which is the population this paper defends (§2.4).

It confirms in a second family what the previous paper established in the first: high enrichment does not indicate historical depth, it indicates a rare sequence recurring, and a rare sequence recurring among neighbours is what a recent loan produces.

Figure 4: Reach against enrichment. Each point is a binary or ternary code; the vertical axis is logarithmic. The codes of extreme enrichment are few and of short reach — and, as the table above shows, recent loans whose concept is attested in three or four languages.

And the mechanism shows better here than there. These concepts — ”banyan”, ”farmer”, ”palm tree”, ”coin” — are attested in only three or four languages of the corpus. Since enrichment is coverage ÷ base frequency, and the coverage of a code present in *all* the languages attesting its concept is 1.000 by construction, extreme enrichment is produced by the rare concept, not the rare code. It is an artefact of the denominator, and it is better said than discovered.

5.6 Geography says less, it does not say nothing

A draft of this paper claimed that in Turkic there is no relation between sharing codes and being close. That was false: we were using only the first of the analysis's two parts.

The first part, by language pairs, does not discriminate. With 32 languages of a young family, every pair shares many codes: 467 of the 496 possible pairs fall in the ”21+” bucket, and the buckets Austronesian needed are empty. From a design with no dynamic range nothing can be concluded, least of all a zero.

The second part, by groups, does discriminate, and gives a monotonic gradient:

Group size	Groups	Dispersion ÷ chance
3–5 languages	763	0.64
6–10	224	0.72
11–25	149	0.85
26–60	14	0.98

Turkic codes do form groups more compact than expected, and compactness fades as reach grows — the same pattern as in Austronesian, where the range ran from 0.40 to 0.65. The difference is one of intensity, not of existence: in Turkic the signal is weaker.

A Turcological explanation we had not considered, and which probably weighs more than any ”mode of diversification”: the modern coordinates of several of these languages are the product of twentieth-century deportations — Karaim in Lithuania, Crimean Tatar deported to Uzbekistan, Karachay-Balkar in 1943 — not of diversification. The geographic proxy is poisoned in this family for concrete historical reasons.

5.7 The emergent classes do not appear, and that too is published

Transitively closing the substitutions that hold up in the validation half yields nothing useful: at any workable threshold, fifteen of the seventeen segments reaching the threshold fall into a single group, and only `{d, t}` detaches. (The full Turkic alphabet is 25 symbols, §2.1; seventeen is how many reach the minimum pair count.)

That is why Appendix A carries no emergent-code column, and the published TSVs carry it empty. Transplanting the Austronesian classes would have been convenient and false: the value of a class is relative to the family that produced it. We say so because it was done wrong: during part of this work, the catalogue generator wrote the Austronesian classes into that column.

6. Limits

The instrument does not distinguish collision from colexification, and in a family with 59 % collision this is the limit that weighs most (§2.3).

Lambdacism is invisible, and not for the reason we assumed: the main cause is that its etyma give skeletons of different lengths (*pillĕk* against *beš*) and the procedure only compares skeletons of equal length. The sibilant merger of §2.1 aggravates this, but is not the cause (§5.3).

There are no zero reflexes. This is a limit of the substitution matrix, not of the whole catalogue: the catalogue groups by skeleton identity, where length is not compared.

A single macroarea, so there is no geographic-anomaly list.

The corpus carries two branch labels — Oghur and Common Turkic — where the reference classification distinguishes Oghuz, Kipchak, Karluk, Siberian and Arghu besides Oghur. So §5.3 did not test "branch pairs" in the plural: it tested the only cut the corpus admits. The results of §5.4 suggest there would be more with complete labels.

Segmentation solves the citation form, not the morphology. Removing the `-mAK` infinitive removes what the dictionary added. Real segmentation remains unresolved, and §5.2 shows that discovering it automatically makes the result worse.

The sample mixes doculects recorded centuries apart. It includes `OldTurkic`, from the eighth century, alongside contemporary varieties, and no measure of lexical similarity models that; nor do we.

What is not claimed here is that some doculects are not "real languages". This work's unit is the doculect with a wordlist, and deciding which of them count as one language is a judgement foreign to the method — the same kind of judgement §1 commits to not making. What *is* a data problem is a duplicate record, and for that there is a declared mechanical rule: same glottocode, or more than 80 % identical skeletons. The close pairs of §3.1 do not cross it, and their effect on the headline figure is measured: 2 %.

We do not calibrate against expert cognacy, though it exists. Savelyev and Robbeets (2020) published a Turkic phylogeny with expert cognacy over a sample much like ours, distributed through Lexibank — the source this pipeline already reads — and it is not ingested into the Integrative Corpus. The Austronesian paper did calibrate (against IE-CoR and ABVD); this one does not, and that is a gap, not an impossibility.

A single family. Everything this paper says about what the figures mean rests on two points. And `docs/PERFIL-FAMILIAS.md` already profiles sixty families with the same metric, where Nakh-Daghestanian — far deeper than Turkic — ties with it. That comparison remains to be made.

7. Conclusion

The catalogue of Turkic codes groups 26.2 % of forms into codes of three or more languages, with a precision 216 times chance, and reproduces without looking for it the vocabulary Turkology knows: the numerals, the pronouns, the body parts.

But what this work contributes is not that catalogue. It is what the family teaches about the instrument, and nearly all of it consists of corrections to what we believed:

That the headline figure is a quotient, and that two thirds of the difference between families comes from the denominator — so two families' ratios over chance are not comparable to each other. That segmentation pays when it is declared and hurts when it is discovered, and that a discoverer finding 365 prefixes in a prefixless family has found the inventory, not the morphology. That the substitution

matrix must be computed per branch pair, because aggregated it erases exactly what defines the genealogy. That geography says less, not nothing. And that maximum enrichment flags borrowings by a nameable mechanism: the rare concept produces it.

And a result we were not looking for and were slow to see, because it was in the data and in no report: the catalogue separates branches without knowing they exist. Three hundred and fifty-seven sibling code pairs, and the largest group Chuvash with Siberian and Oghuz with western Kipchak. It learned that from no tree.

An instrument that claims little can be applied where there is no literature. But to know what it claims, one must apply it where there is — and be willing for the answer to be that it claimed too much.

References

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

Clauson, G. (1972). *An Etymological Dictionary of Pre-Thirteenth-Century Turkish*. Oxford University Press.

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

Erdal, M. (2004). *A Grammar of Old Turkic*. Brill.

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

Holman, E. W., Wichmann, S., Brown, C. H. *et al.* (2008). Explorations in automated language classification. *Folia Linguistica* 42.

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

Johanson, L. and Csató, É. Á. (eds.) (1998). *The Turkic Languages*. Routledge.

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., 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).

Róna-Tas, A. and Berta, Á. (2011). *West Old Turkic: Turkic Loanwords in Hungarian*. Harrassowitz. — [Oghur rhotacism and lambdacism]

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/turkic/output/catalogo-turkic.tsv` — 1,517 rows
Per-form detail	`families/turkic/output/catalogo-turkic-lenguas.tsv` — 14,383 rows
Hygiene (§3.1)	`analysis/fases/fase0_higiene.py` → `runs/fase0-higiene.md`
Data decisions (§3.1, §3.2)	`ci/decisiones.py`
Construction	`analysis/fases/catalogo_familia.py`
Grouping and nulls (§3.4, §4.1)	`analysis/fases/agrupamiento_familia.py` → `runs/agrupamiento.md`
Morphological stripping (§5.2)	`analysis/fases/fase2_despiece.py` → `runs/fase2-despiece.md`
Aggregated substitutions (§5.3, §5.6)	`analysis/fases/fase6_clases_emergentes.py` → `runs/fase6-clases.md`
Substitutions by branch (§5.3)	`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`
Appendix A	`analysis/fases/anexo_codigos.py`

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

Appendix A · One hundred commented codes

The catalogue holds 1,517 codes and the full data are in `families/turkic/output/`. What follows is a readable sample: the hundred binary and ternary codes of widest reach, each with up to sixteen languages and their words.

Three notes on presentation. The languages are spread across the alphabetical list, not taken from its head. Long names are trimmed; the forms are not altered, except for the dots with which lexibank separates segments in 296 Turkish forms (`b.i.r` reads `bir`), which are the source's punctuation. And the thematic grouping is a declared reading order, not a semantic classification.

This appendix carries no emergent-code column, and that absence is a result. In Austronesian the substitutions closed into three clean classes, written beside each code. In Turkic they do not: at any useful threshold the transitive closure puts sixteen of the eighteen segments into a single group (§5.6). A class of sixteen segments informs nothing, so the column is absent. Transplanting the Austronesian classes here would have been convenient and false: the value of a class is relative to the family that produced it.

Reverse index: codes recurring across concepts

In Turkic this is the most important table in the appendix — and it is worth reading before the blocks, because it explains the rest.

`k·l` appears with four concepts: "ash", "come", "lake" and "laugh". This is neither a catalogue error nor a coincidence: in Turkic they are *kül*, *kel-*, *köl* and *kül-*. Four distinct words whose only difference is the vowel, which is exactly what the skeleton does not record. Likewise `k·n`, covering "day", "sun" and "wide" (*kün*, *kün*, *keŋ*) — with one qualification: "day" and "sun" are the same word, and that pair is genuine colexification, not collision.

That is the distinction to hold throughout the appendix: one code may gather the same word used for two things, or two words the instrument cannot tell apart, and the code alone does not say which. The 53% collision declared in §2.3 is this, counted.

Code	Concepts it appears with, and their enrichment
`k·l`	ash (29, 119×) · come (26, 104×) · lake (23, 95×) · laugh (23, 95×)
`t·r`	stand (27, 86×) · skin (26, 83×) · narrow (20, 68×) · sit (16, 51×)
`s·r`	yellow (19, 55×) · suck (18, 60×) · rotten (18, 58×) · ask (request) (15, 47×)
`s·n`	thou (27, 86×) · count (20, 66×) · true (18, 57×)
`t·s`	stone (27, 101×) · tooth (23, 92×) · breast (17, 66×)
`k·n`	day (25, 150×) · wide (20, 124×) · sun (17, 102×)
`b·r`	one (25, 98×) · give (22, 86×) · go (22, 86×)
`j·t`	say (18, 143×) · lie down (17, 144×) · swallow (15, 123×)

Numerals

The numerals show Turkic at its sharpest and explain why this family scores 225 times above chance.

The vowel varies everywhere and the skeleton does not move. "One" is *bir* in Azerbaijani, Kazakh and Uyghur, *ber* in Bashkir and Tatar, *bĭr* in Kazakh, *bi:r* in Dolgan and Sakha: six distinct vowels over `b·r`. "Five" runs through *beš*, *bes*, *bɛʃ*, *biʃ*, *bäš* and *bi:eʃ* without leaving `b·s`.

`t·r·t|four` is the block's only ternary, and deserves attention because its first and last positions are the same consonant: *tört*, *tørt*, *tyørt*, *tort*. A ternary with `t…t` has a smaller space of possibilities than

one with three distinct consonants, and therefore a slightly cheaper enrichment — 129× against the 139× of `b·s`, which is binary. Enrichment should not be read without looking at what the code is made of.

`b·s|five` · binary · 25 languages · enrichment 127×

Azerbaijani beš	Bashkir beš	Dolgan bies	Gagauz bulaʃɯq
KarachayBalkar beš	Karaim beš	KazanTatar biš	Khalaj bi:eʃ, be:eʃ
Kumyk beš	NorthAltai beš	Northern Uzbek bɛʃ	Sakha bies
Tatar biʃ	Turkish beʃ	Turkmen bäš	Uyghur bäš

`b·r|one` · binary · 25 languages · enrichment 98×

Azerbaijani bir	Bashkir ber	Dolgan bi:r	Gagauz bir
KarachayBalkar bir	Karaim bir	KazanTatar ber	Ki-Akmuz 1 bir
Nogai bir	Northern Uzbek bir	OldTurkic bir	Salar bir
Tatar ber	Turkish bir	Turkmen bir	Uyghur bir

`t·r·t|four` · ternary · 16 languages · enrichment 124×

BarabaTatar tört	Dolgan tyørt	KK-Halqabad 1 tört	KarachayBalkar tört
Kazakh tört	Khakas tört	Khalaj tø:ørt, ty:ørt	Ki-Akmuz 1 tort
North Altay tørt	NorthAltai tört	Northern Uzbek tort	OldTurkic tört
Sakha tyørt	Shor tört	Uyghur tört	Uzbek tọrt

Body parts

`s·s|hair` gathers the trace of a diagnostic correspondence without declaring it. The forms run from *sač* and *šaš* to *čač* and *cac*, and among them sits Chuvash *śüś*. The instrument groups them because the canonical normalisation writes all those sibilants as `s`; what it cannot do is say that *śüś* against *sač* is the same change that separates Oghur from Common Turkic. That is what §5.3 says, and only when restricted to the branch pair.

Binaries dominate and they are short: `b·s|head` (*baš*/*bas*), `t·l|tongue` (*til*/*tel*), `k·z|eye` (*köz*/*küz*), `t·s|tooth`. It is the mean arity of 2.85 made into a list. And it shows where the instrument's cost comes from in this family: with two consonants there is little room to go wrong, but also little room to tell apart.

`s·s|hair` · binary · 27 languages · enrichment 134×

Azerbaijani sač	BarabaTatar cac	Chuvash śüś	Gagauz sač
KK-Halqabad 1 sas	Karaim sač	KazanTatar čäč	Khakas sas
Ki-Akmuz 1 čač	Nogai šaš	North Altay tɕatɕ	OldTurkic sač
Shor šaš	Tofa čeˤš	Turkmen sač	Uzbek sɔč

`t·r|skin` · binary · 26 languages · enrichment 83×

BarabaTatar täri	Bashkir tire	CrimeanTatar teri	Dolgan tiri:
KarachayBalkar teri	Kazakh terĭ	KazanTatar tire	Khalaj tɛri
Kumyk teri	Nogai teri	NorthAltai tere	Northern Uzbek tɛri
Sakha tiri:	Shor tere	Tatar tire	Uzbek teri

`b·s|head` · binary · 25 languages · enrichment 127×

Azerbaijani baš	Bashkir baš	Dolgan bas	Gagauz baš
KarachayBalkar baš	Karaim baš	KazanTatar baš	Khalaj baːʃ
Kumyk baš	North Altay baʃ	NorthAltai baš	OldTurkic baš
Sakha bas	Tofa baˁš	Turkish baʃ	Uyghur baš

t·l|tongue · binary · 24 languages · enrichment 143×

BarabaTatar til	Bashkir tel	Dolgan tïl	KK-Halqabad 1 til
Karaim til	Kazakh til	Khakas tɘl	Khalaj til
Kumyk til	Nogai til	NorthAltai til	Northern Uzbek til
Sakha tïl	Shor til	Uyghur til	Uzbek til

t·s|tooth · binary · 23 languages · enrichment 92×

Bashkir teš	CrimeanTatar tiš	Dolgan tiːs	KarachayBalkar tiš
Karaim tis	KazanTatar teš	Khakas tɘs	Ki-Akmuz 1 tiš
Kumyk tiš	Nogai tis	NorthAltai tiš	Northern Uzbek tiʃ
Sakha tiːs	Salar tič	Tatar teʃ	Uzbek tiš

q·n|blood · binary · 22 languages · enrichment 393×

BarabaTatar ḳan	Bashkir qan	CrimeanTatar ḳan	KarachayBalkar qan
Kazakh qan	KazanTatar ḳan	Ki-Akmuz 1 qan	Kumyk qan
North Altay qan	NorthAltai qan	Northern Uzbek qɒn	Salar qan
Shor qan	Tatar qɑn	Uyghur qan	Uzbek ḳɔn

r·n|PUS · binary · 21 languages · enrichment 389×

Azerbaijani irin	BarabaTatar äräŋ	Bashkir eren	CrimeanTatar irin
KK-Halqabad 1 iriŋ	KarachayBalkar irin	Karaim irin	KazanTatar eren
Khakas ɘrɘŋ	Ki-Akmuz 1 iriŋ	Nogai irin	NorthAltai iriŋ
Sakha iriŋe	Tofa iriŋ	Turkish irin	Turkmen iriŋ

q·l|ARM OR HAND · binary · 20 languages · enrichment 152×

BarabaTatar ḳol	Bashkir qul	CrimeanTatar ḳol	KK-Halqabad 1 qol
Kazakh qol	KazanTatar ḳul	Khalaj qol	Ki-Akmuz 1 qol
Nogai qol	North Altay qol	NorthAltai qol	OldTurkic ḳol
Shor qol	Tofa ḳol	Uyghur qol	Uzbek ḳọl

q·l·q|ear · ternary · 20 languages · enrichment 301×

BarabaTatar ḳolaḳ	Bashkir qolaq	CrimeanTatar ḳulaḳ	KK-Halqabad 1 qulaq
Kazakh qulaq	KazanTatar ḳolaḳ	Khalaj qulaːq	Ki-Akmuz 1 qulaq
Nogai qulaq	North Altay qɯlaq	NorthAltai qulaq	Northern Uzbek qulɒq
Tofa ḳulaḳ	Tuvan qulaq	Uyghur qulaq	Uzbek ḳulɔḳ

b·r·n|nose · ternary · 20 languages · enrichment 528×

Azerbaijani burun	CrimeanTatar burun	Gagauz burnu	KarachayBalkar burun
Kazakh burun	KazanTatar borïn	Khalaj burɯn	Ki-Akmuz 1 burun
Nogai burïn	North Altay burun	NorthAltai burun	Northern Uzbek burun
Tatar borɤn	Turkish burn	Turkmen burun	Uyghur burun

q·n·t|wing · ternary · 20 languages · enrichment 347×

BarabaTatar ḳanat	Bashkir qanat	CrimeanTatar ḳanat	KK-Halqabad 1 qanat
Kazakh qanat	KazanTatar ḳanat	Ki-Akmuz 1 qanat	Kumyk qanat
North Altay qanat	NorthAltai qanat	Northern Uzbek qanɒt	OldTurkic ḳanat
Shor qanat	Tatar qanat	Uyghur qanat	Uzbek ḳanɔt

t·z|knee · binary · 19 languages · enrichment 136×

BarabaTatar tizi	CrimeanTatar tiz	KK-Halqabad 1 tizʲe	Karaim tyz, tiz, kiz
Kazakh tizʲe	KazanTatar tez	Ki-Akmuz 1 tize	Kumyk tiz
Nogai tiz	North Altay tize	NorthAltai tize	OldTurkic tiz
Shor tize	Tatar tez	Uyghur tiz	Uzbek tiz

k·z|eye · binary · 18 languages · enrichment 303×

Azerbaijani köz	BarabaTatar köz	CrimeanTatar köz	KK-Halqabad 1 köz
KarachayBalkar köz	Karaim köz	Kazakh köz	KazanTatar küz
Ki-Akmuz 1 köz	Nogai köz	Northern Uzbek koz	OldTurkic köz
Salar koz	Tatar kʉz	Uyghur köz	Uzbek ḳoz

t·s|BREAST · binary · 17 languages · enrichment 66×

BarabaTatar tüš	Bashkir tyʃ	Dolgan tyøs	KK-Halqabad 1 ? tøs
KarachayBalkar töš	Karaim töš	Kazakh tös	KazanTatar tüš
Kumyk töš	North Altay tøʃ	NorthAltai töš	OldTurkic töš
Sakha tyøs	Shor töš	Tatar tʉʃ	Tofa töš

r·q|back · binary · 17 languages · enrichment 331×

Bashkir arqa	CrimeanTatar arḳa	KK-Halqabad 1 arqa	Kazakh arqa
KazanTatar arḳa	Khalaj arqa	Ki-Akmuz 1 arqa	Kumyk arqa
Nogai arqa	North Altay arqa	NorthAltai arḳa	Northern Uzbek ɒrqa
OldTurkic arḳa	Turkmen arqa	Uyghur arqa	Uzbek ɔrḳa

b·t|UPPER LEG (THIGH) · binary · 16 languages · enrichment 124×

Bashkir bot	CrimeanTatar but	Dolgan bu:t	Gagauz but
Karaim but	KazanTatar bot	Khalaj bu:ut	Kumyk but
Nogai ? but	North Altay but	NorthAltai but	Sakha ? bu:t
Tatar bot	Tofa but	Turkish but	Turkmen but

q·r·n|belly · ternary · 15 languages · enrichment 504×

CrimeanTatar ḳarïn	KK-Halqabad 1 ? qarɯn	KarachayBalkar qarɯn	Kazakh qarən
KazanTatar ḳarïn	Khalaj qa(:)rɯn	Ki-Akmuz 1 qɔrɨn	North Altay qarɯn
NorthAltai ḳarïn	Northern Uzbek qɒrin	OldTurkic ḳarïn	Shor qarnɯ
Uyghur qerin	Uzbek ḳɔrin	WestYugur qarɯn	

s·s·k|flower · ternary · 15 languages · enrichment 354×

Azerbaijani t͡ʃɪt͡ʃæk	BarabaTatar čičäk	Bashkir säskä	Chuvash t͡ɕɛ't͡ɕɛk
CrimeanTatar čiček	Gagauz čiček	KazanTatar čäčäk	Kumyk čeček
North Altay ʨeʥek	NorthAltai čeček	OldTurkic čeček	Tatar t͡ɕæt͡ɕkæ
Turkish tʃɪtʃɛk	Tuvan čeček	Uyghur ? ʨeʨɛk	

Kinship and person

The whole pronominal system appears in the catalogue, and it appears as a system:

m·n|I (*men*, *min*, *mɛn*) · s·n|THOU (*sen*, *sin*, *sän*) · b·z|WE (*biz, bez, bĭz*) · s·z|YOU (*siz*) · l·r|THEY (*olar*, *alar*, *ular*, *ïlar*).

Note the top pair against the bottom one: the singular ends in **n** and the plural in **z**. The catalogue does not know –z is a plural morpheme — it does not know what a morpheme is — but it puts it on the table, adjacent and with its figures, which is exactly what this instrument can do: not analyse, but leave the material laid out and adjacent for someone else to analyse.

n·d|THERE reaches 562×, the highest enrichment in this section. With 19 languages and two frequent consonants, that figure says more about how rare the sequence n·d is in Turkic than about the age of the demonstrative.

k·m|WHO · binary · 30 languages · enrichment 352×

Azerbaijani kim	BarabaTatar kim	Chuvash kam	Dolgan kim
KK–Halqabad 1 kim	Karaim kim	KazanTatar kem	Khalaj kim
Kumyk kim	Nogai kim	NorthAltai kem	OldTurkic kem
Salar kem	Tatar kem	Turkmen kim	Uzbek kim

m·n|I · binary · 28 languages · enrichment 224×

Azerbaijani men	BarabaTatar min	CrimeanTatar men	KK–Halqabad 1 men
Karaim men	Kazakh men	Khakas min	Ki–Akmuz 1 men
Nogai men	North Altay men	Northern Uzbek mɛn	Salar men
Tatar min	Tofa men	Tuvan men	Uzbek men

s·n|THOU · binary · 27 languages · enrichment 86×

Azerbaijani sen	BarabaTatar sin	Gagauz sän	KarachayBalkar sen
Karaim sen	KazanTatar sin	Khalaj sɛn	Ki–Akmuz 1 sen
Nogai sen	NorthAltai sen	Northern Uzbek sɛn	Salar sen
Tatar sin	Tofa sen	Turkmen sen	Uyghur sän

l·r|THEY · binary · 20 languages · enrichment 305×

Bashkir ular	CrimeanTatar olar	KK–Halqabad 1 olar	Karaim alar
KazanTatar alar	Khakas olar	Khalaj ollar	Ki–Akmuz 1 alar
Nogai olar	North Altay ɯlar	NorthAltai olor	Northern Uzbek ular
Shor ïlar	Turkmen olar	Tuvan olar	Uyghur ular

b·z|WE · binary · 20 languages · enrichment 208×

Azerbaijani biz	BarabaTatar biz	CrimeanTatar biz	Gagauz biz
KarachayBalkar biz	Karaim biz	Kazakh bĭz	KazanTatar bez
Ki–Akmuz 1 biz	Kumyk biz	Nogai biz	Northern Uzbek biz
Tatar bez	Turkish biz	Turkmen biz	Uyghur biz

n·d|THERE · binary · 19 languages · enrichment 562×

BarabaTatar anda	Bashkir unda	Chuvash unda	CrimeanTatar anda
KarachayBalkar anda	Karaim anda	KazanTatar anda	Khakas anda
Ki–Akmuz 1 anda	Kumyk onda	Nogai onda	NorthAltai anda
Shor anda	Tatar anda	Tofa ïnda	Tuvan ïnda

q·s·n|WHEN · ternary · 19 languages · enrichment 489×

BarabaTatar ḳačan	Bashkir qasan	KK–Halqabad 1 qasan	KarachayBalkar qaʦan
Kazakh qaʃan	KazanTatar ḳayčan	Ki–Akmuz 1 qaʧan	Kumyk qačan
Nogai qašan	North Altay qaʗan	NorthAltai qaʦan	Salar qaʗan
Shor qaʦan	Tuvan qaʒan	Uyghur qaʦan	Uzbek ḳačɔn

k·s|PERSON · binary · 18 languages · enrichment 85×

BarabaTatar kiši	Bashkir keše	Karaim kiši	Kazakh kĭsĭ
KazanTatar keše	Ki–Akmuz 1 kiši	Nogai kisi	North Altay kiʒi
OldTurkic kiši	Salar kiš	Shor kiži	Tofa kiši
Turkish kiši	Tuvan kiži	Uyghur kiši	Uzbek kiši

s·z|YOU · binary · 17 languages · enrichment 125×

Azerbaijani siz	CrimeanTatar siz	Gagauz siz	KK–Halqabad 1 siz
KarachayBalkar siz	Karaim siz	Kazakh səz	KazanTatar sez
Khalaj siz	Kumyk siz	Nogai siz	Northern Uzbek siz
OldTurkic siz	Tatar sez	Turkish siz	Turkmen siz

b·l|CHILD (DESCENDANT) · binary · 15 languages · enrichment 71×

Bashkir bala	CrimeanTatar bala	KK–Halqabad 1 bala	KarachayBalkar bala
Karaim bala	Kazakh bala	KazanTatar bala	Khalaj bala
Ki–Akmuz 1 bala	Kumyk bala	Nogai bala	North Altay bala
NorthAltai bala	Uyghur bala	Uzbek bɔla	

m·n·d|HERE · ternary · 15 languages · enrichment 919×

BarabaTatar mïnda	CrimeanTatar mïnda	Kazakh mŭnda	KazanTatar monda
Khakas mïnda	Ki–Akmuz 1 mïnda	Kumyk munda	Nogai munda
North Altay mɯnda	NorthAltai mïnda	Shor mïnda	Tatar monda
Tofa mïnda	Tuvan mïnda	WestYugur mïnda	

Nature and sky

k·n with ”day” and with ”sun” is colexification, not collision — the distinction the reverse index asks to be held. In Turkic *kün* means both; it is one word used for two concepts, not two words confused by the instrument. That both enter the catalogue separately is correct: they are two Concepticon nodes, and the catalogue indexes by concept.

t·t·n|smoke (*tütün*) is one of the appendix’s few ternaries with a repeated consonant, and q·m|sand (*qum*) one of the shortest possible binaries. Between the two lies almost the whole arity range of the family.

k·l|ash · binary · 29 languages · enrichment 119×

Azerbaijani kyl	Bashkir köl	CrimeanTatar kül	Gagauz kyl
KarachayBalkar kyl	Kazakh kyl	KazanTatar köl	Khalaj kil
Kumyk kul	North Altay kyl	Northern Uzbek kul	OldTurkic kül
Salar kul	Tatar kœl	Turkmen kyl	Uzbek kul

t·s|stone · binary · 27 languages · enrichment 101×

BarabaTatar taš	Bashkir taš	Dolgan ta:s	KK-Halqabad 1 tas
KarachayBalkar taš	Kazakh tas	Khakas tas	Khalaj ta:aʃ
Kumyk taš	North Altay taʃ	NorthAltai taš	OldTurkic taš
Salar taš	Shor taš	Turkish taʃ	Uzbek tɔš

b·l·t|cloud · ternary · 25 languages · enrichment 318×

Azerbaijani bulut	Bashkir bolot	Dolgan bïlït	Gagauz bulut
KarachayBalkar bulut	Karaim bulut	KazanTatar bolït	Khalaj bulɯt
Kumyk bulut	NorthAltai bulut	Northern Uzbek bulut	Sakha bïlït
Tatar bolɤt	Turkish bulut	Turkmen bulut	Uyghur bulut

k·n|day · binary · 25 languages · enrichment 150×

BarabaTatar kOn	Bashkir kön	CrimeanTatar kün	Dolgan kyn
KarachayBalkar kyn	Karaim kyn, kin	KazanTatar kön	Khakas kyn
Ki-Akmuz 1 kyn	North Altay kyn	NorthAltai ? kyn	OldTurkic kün
Sakha kyn	Shor kyn	Tatar kœn	Uzbek kun

k·l|lake · binary · 23 languages · enrichment 95×

BarabaTatar köl	Bashkir kyl	Chuvash kylə	KK-Halqabad 1 köl
KarachayBalkar köl	KazanTatar kül	Khakas köl	Kumyk köl
Nogai köl	North Altay køl	Northern Uzbek kol	OldTurkic köl
Shor köl	Tatar kʉl	Uyghur köl	Uzbek kọl

q·r|snow · binary · 22 languages · enrichment 132×

BarabaTatar ḳar	Bashkir qar	CrimeanTatar ḳar	KarachayBalkar qar
Kazakh qar	KazanTatar ḳar	Ki-Akmuz 1 qar	Kumyk qar
North Altay qar	NorthAltai qar	Northern Uzbek qɒr	Salar qar
Shor qar	Tatar qɑr	Uyghur qar	Uzbek ḳɔr

t·m·n|fog · ternary · 21 languages · enrichment 368×

BarabaTatar tOman	Bashkir toman	CrimeanTatar tuman	Dolgan tuman
Karaim tuman	Kazakh tuman	KazanTatar toman	Kumyk tuman
Nogai tuman	NorthAltai tuman	OldTurkic tuman	Sakha tuman
Salar tuman	Tofa tuman	Tuvan tuman	Uyghur tuman

t·s·k|HOLE · ternary · 18 languages · enrichment 664×

BarabaTatar tĕšik	Bashkir tišek	CrimeanTatar tešik	KK-Halqabad 1 tesik
KarachayBalkar tešik	Karaim tešik	Kazakh tesık	KazanTatar tišek
Kumyk tešik	Nogai tesik	North Altay teʒik	NorthAltai težik
Northern Uzbek tɛʃik	Shor težik	Tatar tiʃek	Uyghur töšük

x·s|WOOD · binary · 18 languages · enrichment 159×

```
Azerbaijani aɣač    BarabaTatar aɣač     Bashkir aɣas  CrimeanTatar aɣač
KK-Halqabad 1 aɣaš  KarachayBalkar aɣač  Karaim aɣac   Kazakh aɣaš
Khakas aɣas         Kumyk aɣač           Nogai aɣaš    North Altay aɣaʃ
NorthAltai aɣaš     OldTurkic ïɣač       Salar aɣač    Shor aɣaš
```

t·n|night · binary · 18 languages · enrichment 105×

```
BarabaTatar tön  Bashkir tön     Dolgan ty:n         KK-Halqabad 1 tyn
Kazakh tyn       KazanTatar tön  Khakas tyn          Ki-Akmuz 1 tyn
North Altay tyn  NorthAltai tyn  Northern Uzbek tun  OldTurkic tün
Sakha ty:n       Shor tyn        Tatar tœn           Uyghur tyn
```

t·t·n|smoke · ternary · 17 languages · enrichment 389×

```
BarabaTatar tötin    Bashkir tötön         CrimeanTatar tütün        Gagauz tytyn
KK-Halqabad 1 tytin  KarachayBalkar tytyn  Karaim tytyn, tytsy etc.  Kazakh tytyn
KazanTatar töten     Khalaj tityn          Ki-Akmuz 1 tytyn          Kumyk tutun
Nogai tytin          Northern Uzbek tutun  OldTurkic tütün           Tatar tœten
```

k·n|sun · binary · 17 languages · enrichment 102×

```
Azerbaijani kün  Dolgan kyn       KK-Halqabad 1 kyn  KarachayBalkar kyn
Kazakh kyn       Khakas kyn       Khalaj kyn, kin    Ki-Akmuz 1 kyn
Nogai kyn        North Altay kyn  NorthAltai kyn     OldTurkic kün
Sakha kyn        Salar kun        Shor kyn           Uyghur kyn
```

j·r|earth · binary · 16 languages · enrichment 113×

```
Azerbaijani jer  BarabaTatar yer  Bashkir jer             CrimeanTatar yer
Gagauz jer       Karaim jer       Khalaj jer 'Ort; Erde'  Ki-Akmuz 1 jɛr
Kumyk jer        Nogai jer        OldTurkic yer           Turkish yer
Turkmen jer      Uyghur jɛr       Uzbek yer               WestYugur ? jer
```

Animals

Three codes, all with the same shape: `b·t|louse` (*bit*), `q·s|bird` (*quš*), `b·l·q|fish` (*balïq*). The block is small and that too is data: animal names are among the areas of the lexicon where Turkic yields fewest codes, probably because the fauna changes with the geography and the lexicon changes with it.

b·t|louse · binary · 22 languages · enrichment 188×

```
Azerbaijani bit  Bashkir bet      CrimeanTatar bit    KarachayBalkar bit
Karaim bit       Kazakh bit       Khalaj bit          Ki-Akmuz 1 bit
Nogai biyt       NorthAltai biyt  Northern Uzbek bit  Sakha bït
Tatar bet        Tofa bïʿt        Turkmen bit         Tuvan bït
```

q·s|bird · binary · 20 languages · enrichment 140×

```
BarabaTatar ḳuš  Bashkir qoʃ         CrimeanTatar ḳuš  KK-Halqabad 1 qus
KazanTatar ḳoš   Ki-Akmuz 1 ḳuš      Kumyk quš         Nogai qus
NorthAltai quʃ   Northern Uzbek quʃ  OldTurkic ḳuš     Salar quʃ
Tatar qoʃ        Tuvan quʃ           Uyghur quʃ        Uzbek ḳuš
```

b·l·q|fish · ternary · 19 languages · enrichment 374×

BarabaTatar balïḳ	Bashkir balïḳ	CrimeanTatar balïḳ	KK-Halqabad 1 balïḳ
Kazakh balïḳ	KazanTatar balïḳ	Kumyk baliq	Nogai balïḳ
North Altay balɯq	NorthAltai balïq	Northern Uzbek baliq	Tofa balïḳ
Turkish balïḳ	Turkmen balïḳ	Tuvan balïḳ	Uyghur beliq

q·r·t|worm · ternary · 19 languages · enrichment 434×

BarabaTatar ḳOrt	CrimeanTatar ḳurt	KK-Halqabad 1 qurt	KarachayBalkar qurt
Kazakh qurt	KazanTatar ḳort	Ki-Akmuz 1 qurt	Kumyk qurt
Nogai qurt (jer qurtɯ)	North Altay qurt	NorthAltai qurt 'worm'	OldTurkic ḳurt
Shor qurt	Tatar qort	Tofa ḳuˁrt	Tuvan qurt

Plants and food

t·m·r|root (*tamïr*) is a clean ternary of good reach. t·z|salt (*tuz*) is the opposite case: two very frequent consonants, high reach and little specificity. The section shows the trade-off well — more consonants is harder to hit by chance, but also harder to preserve — which §2.4 measures.

t·z|salt · binary · 18 languages · enrichment 133×

CrimeanTatar tuz	Gagauz tuz	KK-Halqabad 1 tuz̦	KarachayBalkar tuz
Karaim tuz	Kazakh tuz	KazanTatar toz	Khalaj tu(:)z
Kumyk tuz	Nogai tuz	Northern Uzbek tuz	OldTurkic tuz
Tatar toz	Turkish tuz	Uyghur tuz	Uzbek tuz

s·j·n|CHEW · ternary · 17 languages · enrichment 823×

Azerbaijani čeynemag	Bashkir säynä	Gagauz ʧijne	KK-Halqabad 1 ʃajna
KarachayBalkar ʧajna	Karaim ʧajna	Kazakh ʃajna	KazanTatar čäynä
Khalaj ʧɛjnɛ, ʧɛjlɛ	Ki-Akmuz 1 ʧajna	Kumyk čaynamaq	Nogai ʃajna
North Altay ʃajna	NorthAltai ʧajna	Salar čäynä	Turkmen ʧejne

Verbs of action and state

This is the section where the phase-0 stripping shows, and it is worth saying so here. The sources nel and ids cite verbs in the infinitive (*jaʃamɒq*, *ашамакъ*) and lexibank in the root (*yaša*). Without removing that −mAK, the same verb came out with two skeletons and formed no code: 4,974 forms, 9% of the family. A good share of this block's codes would not exist without that decision, which is declared in §3.2 and measured.

k·l with "come" and with "laugh" is the pair the reverse index announces: *kel-* and *kül-*, two verbs separated only by the vowel. Here the catalogue over-groups, and has no way of noticing.

k·s|CUT · binary · 29 languages · enrichment 132×

Azerbaijani kesmag	BarabaTatar kës-	CrimeanTatar kes	KK-Halqabad 1 kes
Karaim kes	KazanTatar kis	Khakas kis	Ki-Akmuz 1 kes
Nogai kes	NorthAltai kes	OldTurkic kes	Salar kes
Tatar kisʉ	Turkish kes	Tuvan kes	Uzbek kes

t·t|GRASP · binary · 29 languages · enrichment 287×

Azerbaijani tut	BarabaTatar tOt−	Chuvash tĭt	Dolgan tut
KK−Halqabad 1 tut	Karaim tut	KazanTatar tot	Khalaj tut
Kumyk tut	North Altay tut	OldTurkic tut	Sakha tut
Shor tut	Turkish tut	Tuvan tut	Uzbek tut

k·r|see · binary · 27 languages · enrichment 168×

BarabaTatar kör−	Bashkir kyr	CrimeanTatar kör	KK−Halqabad 1 kör
KarachayBalkar kör	Kazakh kör	Khakas kör	Khalaj ker
Nogai kör	NorthAltai kör	Northern Uzbek kormɒq	Sakha kör
Shor kör	Tatar kʉrʉ	Tuvan kör	Uzbek kọr

t·r|stand · binary · 27 languages · enrichment 86×

BarabaTatar tOr	Bashkir tor	CrimeanTatar tur	KK−Halqabad 1 trʊː
KarachayBalkar tur	Kazakh tru	Khakas tur	Khalaj tur
Kumyk tur	North Altay tur	NorthAltai tur	OldTurkic tur
Shor tur	Tatar toru	Tuvan tur	Uzbek tur

k·l|come · binary · 26 languages · enrichment 104×

BarabaTatar kĕl	Bashkir kil	CrimeanTatar kel	Dolgan kel
KarachayBalkar kel	Kazakh kel	KazanTatar kil	Khalaj kɛl
Nogai kel	North Altay kel	Northern Uzbek kɛlmɒq	OldTurkic kel
Shor kel	Tofa kel	Tuvan kel	Uzbek kel

b·l|know · binary · 26 languages · enrichment 111×

Azerbaijani bil	Bashkir bel	Dolgan bil	Gagauz bil
KarachayBalkar bil	Kazakh blʊ	KazanTatar bel	Ki−Akmuz 1 bil
Nogai bil	NorthAltai bil	OldTurkic bil	Sakha bil
Tatar belʉ	Turkish bil	Turkmen bil	Uyghur bil

t·r·t|pull · ternary · 26 languages · enrichment 196×

BarabaTatar tart−	Bashkir tart	CrimeanTatar tart	Dolgan tart
KarachayBalkar tart	Kazakh tart	KazanTatar tart	Ki−Akmuz 1 itɛrɛt
Nogai tart	North Altay tart	Northern Uzbek tɒrtmɒq	OldTurkic tart
Shor tart	Tofa tiˁrt	Tuvan tĭrt	Uzbek tɔrt

t·p|FIND · binary · 24 languages · enrichment 294×

Azerbaijani tap	BarabaTatar tap−	Chuvash tup	CrimeanTatar tap
Karaim tap	Kazakh tap	Khakas tap	Ki−Akmuz 1 tap
Nogai tap	North Altay tap	Northern Uzbek tɒpmɒq	OldTurkic tap
Tofa tiˁp_	Turkmen tap	Uyghur tap	Uzbek tɔp

t·k|sew · binary · 24 languages · enrichment 329×

Azerbaijani tik	BarabaTatar tik–	CrimeanTatar tik	Dolgan tik
KarachayBalkar tik	Karaim tik	KazanTatar tek	Khakas tɘk
Ki–Akmuz 1 tik	Kumyk tik	North Altay tik	NorthAltai tik–
Sakha tik	Shor tik	Uyghur tik	Uzbek tik

k·l|laugh · binary · 23 languages · enrichment 95×

BarabaTatar köl–	Bashkir köl	Chuvash kul	KK–Halqabad 1 kyl
KarachayBalkar kyl	Kazakh kyl	KazanTatar köl	Khalaj kil
Ki–Akmuz 1 kyl	Kumyk kulemaq	North Altay kyl	NorthAltai kül–
Sakha kyl	Salar kulli	Uyghur kyl	Uzbek kul

j·n|play · binary · 23 languages · enrichment 145×

Azerbaijani ojna	Bashkir ujna	Dolgan o:jn(')o:	KK–Halqabad 1 ojna
KarachayBalkar ojna	Kazakh ojna	Khakas ojna	Kumyk ojna
Nogai ojna	North Altay ojna	Northern Uzbek ojnamɒq	Salar ojna
Tatar ujnau	Turkish ojna	Tuvan ojna	Uyghur ojna

q·l|REMAIN · binary · 22 languages · enrichment 157×

BarabaTatar ḳal–	Bashkir qal	CrimeanTatar ḳal	KarachayBalkar qal
Kazakh qal	KazanTatar ḳal	Ki–Akmuz 1 qal	Kumyk qal
North Altay qal	NorthAltai ḳal–	Northern Uzbek qɒlmɒq	Salar qal
Shor qal	Tatar qalu	Uyghur qal	Uzbek ḳɔl

b·r|give · binary · 22 languages · enrichment 86×

BarabaTatar ber–	Bashkir bir	CrimeanTatar ber	KK–Halqabad 1 ber
KarachayBalkar ber	Karaim ber	KazanTatar bir	Ki–Akmuz 1 ber
Nogai ber	NorthAltai ber	Northern Uzbek bɛrmɒq	Sakha bier
Tatar birᵾ	Tofa ber	Tuvan ber	Uyghur bär

b·r|go · binary · 22 languages · enrichment 86×

BarabaTatar bar	Bashkir bar	CrimeanTatar bar	KK–Halqabad 1 ? bar
KarachayBalkar bar	Karaim bar	KazanTatar bar	Ki–Akmuz 1 bar
Nogai bar	NorthAltai bar	Northern Uzbek bɒrmɒq	Sakha bar
Salar bar	Tofa bar	Tuvan bar	Uyghur bar

s·t|hear · binary · 21 languages · enrichment 85×

Azerbaijani ešit	BarabaTatar išit–	Bashkir išet	CrimeanTatar ešit
KK–Halqabad 1 esit	KarachayBalkar ešit	Karaim esit	KazanTatar išet
Khalaj iʃyt	Kumyk ešit	Northern Uzbek æʃitmɒq	OldTurkic ešit
Salar išti	Turkish iʃit	Turkmen ešit	Uyghur išit

s·n|count · binary · 20 languages · enrichment 66×

KK–Halqabad 1 sana	KarachayBalkar sana	Kazakh sana	KazanTatar sana
Khalaj sa:ana	Ki–Akmuz 1 sana	Kumyk sana	Nogai sana
NorthAltai sana–	OldTurkic sana	Salar sana'	Tatar sanau
Turkmen sana	Tuvan sana	Uyghur sana	Uzbek sana

q·r·q|fear · ternary · 20 languages · enrichment 373×

BarabaTatar ḳorḳ–	Bashkir qurq	CrimeanTatar ḳorḳ	KK–Halqabad 1 qorq
Karaim qorq	Kazakh qorq	KazanTatar ḳurḳu	Khalaj qorq
Kumyk qorq	Nogai qorq	North Altay qoruq	NorthAltai ḳoruḳ–
Shor qoruq	Tatar qurqu	Uyghur qorq	Uzbek ḳorḳ

t·k·r|SPIT · ternary · 19 languages · enrichment 390×

BarabaTatar tökUr–	Bashkir tökör	CrimeanTatar tükür	Gagauz tykyr
KK–Halqabad 1 tkʊrʊ	KarachayBalkar tykyr	Kazakh tkəru	KazanTatar töker
Khakas tykyr	Ki–Akmuz 1 tykyr	Kumyk tukurmaq	North Altay tykyr
NorthAltai tykyr	Shor tykkyr	Tofa tükkür	Turkish tykyr

j·s|BE ALIVE · binary · 18 languages · enrichment 174×

Azerbaijani jaʃa	BarabaTatar yešä–	Bashkir jæʃæy	CrimeanTatar yaša
Gagauz jaʃa	KazanTatar yäšä	Ki–Akmuz 1 jaʃɔː	Kumyk jaʃa
Northern Uzbek jaʃamɒq	OldTurkic yaša	Salar jaʃa	Tatar jaʃæʉ
Turkish jaʃamak	Turkmen jaʃa	Uyghur jaʃa	Uzbek yaša

s·r|SUCK · binary · 18 languages · enrichment 60×

Azerbaijani sor	KK–Halqabad 1 sor	Kazakh sor	KazanTatar ?_suïr
Khakas sor	Khalaj suor	Ki–Akmuz 1 sor	Kumyk sor
North Altay sor	NorthAltai soor	OldTurkic sor	Shor sor
Tofa sor	Turkmen ? sor	Tuvan sor	Uyghur šora

q·z|dig · binary · 18 languages · enrichment 251×

CrimeanTatar ḳaz	KK–Halqabad 1 qaz	KarachayBalkar qaz	Kazakh qaz
KazanTatar ḳaz	Khalaj qaz	Ki–Akmuz 1 qaz	Kumyk qaz
Northern Uzbek qazimɒq	OldTurkic ḳaz	Salar qazɯ	Tatar qɑzu
Tofa ḳaz	Tuvan qaz	Uyghur qaz	Uzbek ḳazi

j·t|say · binary · 18 languages · enrichment 143×

BarabaTatar äyt	Bashkir äyt	KK–Halqabad 1 ajt	KarachayBalkar ajt
Karaim ajt	Kazakh ajt	KazanTatar äyt	Ki–Akmuz 1 ajt
Nogai ajt	North Altay ajt	NorthAltai ajt	Northern Uzbek ajtmɒq
Shor ajt	Tatar æjtʉ	Turkmen ajt	Uyghur ejt

j·t|LIE DOWN · binary · 17 languages · enrichment 144×

Azerbaijani yatmag	BarabaTatar yat–	Bashkir jat	CrimeanTatar yat
Gagauz jat	Karaim jat	KazanTatar yat	Khalaj jaːt
Kumyk jat	Nogai jat	OldTurkic yat	Salar jat
Turkish yat	Turkmen jat	Uyghur jat	Uzbek yɔt

s·s|swell · binary · 17 languages · enrichment 94×

Azerbaijani šiš	Bashkir šeš	Chuvash šĭś	CrimeanTatar šiš
Gagauz šiš	Karaim sis	KazanTatar šeš	Khakas sɔs
Khalaj ʃɯʃ	Ki–Akmuz 1 šiš	Kumyk šiš	Nogai sis
OldTurkic sïš	Shor šiš	Turkish šiš	Turkmen čiš

q·s|vomit · binary · 17 languages · enrichment 142×

CrimeanTatar ḳus	KK–Halqabad 1 qus	KarachayBalkar qus	Kazakh qŭs
KazanTatar ḳos	Khalaj qus	Ki–Akmuz 1 qus	Kumyk qus
Nogai qus	North Altay qus	NorthAltai qus	OldTurkic ḳus
Tofa ḳuˁs	Tuvan qus	Uyghur qus	Uzbek ḳus

t·r|sit · binary · 16 languages · enrichment 51×

Azerbaijani otur	BarabaTatar otïr–	CrimeanTatar otur–	Gagauz otur
KK–Halqabad 1 otïr–	Kazakh otïr–	KazanTatar utïr–	Ki–Akmuz 1 otur–
North Altay otur	NorthAltai otur–	Northern Uzbek otirmɒq	Salar otïr–
Tatar utɤru	Turkish otur	Turkmen otur	Uzbek o̡tir–

s·r|ASK (REQUEST) · binary · 15 languages · enrichment 47×

BarabaTatar sora–	CrimeanTatar sora	KK–Halqabad 1 sora	Kazakh sŭra
KazanTatar sora	Khakas sur	Ki–Akmuz 1 sura	Nogai sora
North Altay sura	NorthAltai sura	Shor sura	Turkish sor
Turkmen sora	Uyghur sora	Uzbek so̡ra	

j·t|SWALLOW · binary · 15 languages · enrichment 123×

BarabaTatar yut–	Bashkir jot	CrimeanTatar yut	Gagauz jut
Karaim jut	KazanTatar yot	Khalaj ju:ut	Kumyk jut
Nogai jut	North Altay jyt	NorthAltai yüt–	Northern Uzbek jutmɒq
Tatar jotu	Turkish jutmak	Uzbek yut	

j·r|WALK · binary · 15 languages · enrichment 106×

Azerbaijani jeri	BarabaTatar yör–	Bashkir jør(ø)	CrimeanTatar yür
Karaim jyry	KazanTatar yör	Khalaj jorɯ	Kumyk jyry
Nogai jyr	Northern Uzbek jurmɒq	OldTurkic yorï	Salar jyr, jer
Turkish jyɾymɛk	Turkmen jøre	Uzbek yur	

Objects and culture

A single code, t·j·q|stick. It is the appendix's poorest section, for a reason that matters: material culture is the most borrowable lexicon, and in a family that crossed Persian, Arabic, Russian and Chinese, it is where the inherited form is least preserved. The scarcity is informative.

t·j·q|stick · ternary · 18 languages · enrichment 514×

BarabaTatar tayaḳ	Bashkir tajaq	CrimeanTatar tayaḳ	KK–Halqabad 1 tajaq
KarachayBalkar tajaq	Kazakh tajaq	KazanTatar tayaḳ	Ki–Akmuz 1 tajaq
Nogai tajaq	North Altay tajaq	NorthAltai tajaq	Northern Uzbek tajɒq
Shor ? tajaq	Tofa tayaḳ	Turkmen tajaq	Uyghur tajaq

Qualities

`q·r|black` (*qara*), `s·r|yellow` (*sarï*), `q·z·l|red` (*qïzïl*), `j·s·l|green` (*jašïl*). The basic colours appear almost complete, and the last two share the same `–l` suffix visible in the skeleton: the catalogue places it adjacent without knowing what it is, just as with the `–z` of the pronouns.

`t·l|full` (*tolï*) shares its code with `t·l|tongue` (*til*). The same warning again, and that makes three: in this family, two consonants do not suffice to separate two words.

`z·n|long` · binary · 28 languages · enrichment 471×

Azerbaijani uzun	BarabaTatar uzïn	Gagauz uzun	KarachayBalkar uzun
Kazakh ŭzïn	KazanTatar ozïn	Khalaj uzu:n	Kumyk uzun
North Altay uzun	NorthAltai uzun	OldTurkic uzun	Shor uzun
Tofa uzun	Turkish uzun	Tuvan uzun	Uzbek uzun

`q·r|black` · binary · 23 languages · enrichment 138×

BarabaTatar ḳara	Bashkir qara	CrimeanTatar ḳara	KarachayBalkar qara
Kazakh qara	Khalaj qara	Ki–Akmuz 1 qara	Nogai qara
North Altay qara	NorthAltai qara	OldTurkic ḳara	Salar qara
Tatar qɑra	Tofa ḳara	Uyghur qara	Uzbek ḳɔra

`s·k|old` · binary · 22 languages · enrichment 132×

Azerbaijani æskɪ	BarabaTatar ĕski	CrimeanTatar eski	KK–Halqabad 1 eski
KarachayBalkar eski	Karaim eski	KazanTatar iske	Ki–Akmuz 1 eski
Nogai eski	North Altay eski	NorthAltai eski	OldTurkic eski
Salar eski	Shor eski	Turkish eski	Uyghur äski

`s·l|left` · binary · 20 languages · enrichment 70×

Azerbaijani sol	BarabaTatar sol	CrimeanTatar sol	KK–Halqabad 1 sɔl
Kazakh sol	KazanTatar sul	Khakas sol	Ki–Akmuz 1 sol
Nogai sol	North Altay sol	NorthAltai sol	Northern Uzbek sol
Shor sol	Tatar sul	Turkish sol	Uyghur sol

`t·r|narrow` · binary · 20 languages · enrichment 68×

Bashkir tar	CrimeanTatar tar	KK–Halqabad 1 tar	KarachayBalkar tar
Kazakh tar	KazanTatar tar	Khakas tar	Khalaj ta:ar
Kumyk tar	Nogai tar	NorthAltai tar	OldTurkic tar
Tofa tar	Tuvan tar	Uyghur tar	Uzbek tɔr

`q·z·l|red` · ternary · 20 languages · enrichment 418×

BarabaTatar ḳïzïl	KK–Halqabad 1 qɨzɨl	KarachayBalkar qɯzɯl	Kazakh qəzəl
Khalaj qɯzɯl	Ki–Akmuz 1 qɨzɨl	Kumyk qizil	Nogai qɯzɯl
NorthAltai qɯzɯl	Northern Uzbek qizil	OldTurkic ḳïzïl	Shor qɯzɯl
Tofa ḳïzïl	Tuvan qɯzɯl	Uyghur qizil	Uzbek ḳizil

`k·n|wide` · binary · 20 languages · enrichment 124×

BarabaTatar käŋ	Bashkir kiŋ	CrimeanTatar keŋ	KK-Halqabad 1 keŋ
Karaim ken	Kazakh keŋ	KazanTatar kiŋ	Khalaj ki:eŋ
Nogai keŋ	North Altay keŋ	NorthAltai keŋ	Northern Uzbek kɛŋ
Sakha kieŋ	Tatar kiŋ	Uyghur käŋ	Uzbek keŋ

t·r·n|DEEP · ternary · 19 languages · enrichment 477×

BarabaTatar täräŋ	Bashkir tärän	Chuvash tarən	CrimeanTatar teren
KK-Halqabad 1 tereŋ	KarachayBalkar teren	Kazakh tereŋ	KazanTatar tirän
Khakas tireŋ	Ki-Akmuz 1 tereŋ	Kumyk teran	North Altay tereŋ
NorthAltai tereŋ	OldTurkic teriŋ	Salar tiroŋ	Shor tereŋ

k·p|MANY · binary · 19 languages · enrichment 438×

BarabaTatar köp	Bashkir kyp	KK-Halqabad 1 köp	Karaim köp
Kazakh köp	KazanTatar küp	Ki-Akmuz 1 köp	Kumyk kop
Nogai köp	North Altay køp	NorthAltai köp	Shor köp
Tofa köˁp	Turkmen köp	Uyghur köp	Uzbek koָp

t·l|full · binary · 19 languages · enrichment 114×

BarabaTatar tol-	Bashkir tulï	Chuvash tuli	CrimeanTatar tolu
KK-Halqabad 1 tɔlɜ	KarachayBalkar tolu	Kazakh tolï	KazanTatar tulï
Khalaj tu(:)ola	Kumyk tolu	Nogai tolï	Northern Uzbek tola
OldTurkic tolu	Salar tolu	Tatar tulɤ	Uzbek toָla

s·r|yellow · binary · 19 languages · enrichment 55×

Azerbaijani saru	BarabaTatar sarï	Chuvash 'sarə	CrimeanTatar sarï
Gagauz sarï	KK-Halqabad 1 sarï	Karaim sarï	Kazakh sarï
KazanTatar sarï	Ki-Akmuz 1 sarï	Kumyk sari	North Altay sarɯ
NorthAltai sarï	Salar sarï	Tatar sarɤ	Turkish sarɯ

s·n|TRUE · binary · 18 languages · enrichment 57×

BarabaTatar čïn	Chuvash t͡ɕən	KK-Halqabad 1 ʔ ʃɯn	Kazakh šïn
KazanTatar čïn	Khakas sïn	Ki-Akmuz 1 čïn	Nogai ušïn
NorthAltai čïn	OldTurkic čïn	Salar čeŋ	Shor šïn
Tofa šïn	Turkmen čïn	Tuvan šïn	Uyghur čin

j·s·l|green · ternary · 18 languages · enrichment 523×

Azerbaijani jaʃɯl	BarabaTatar yišil	Bashkir jæʃel	CrimeanTatar yešil
Gagauz jeʃil	Karaim jesil	KazanTatar yäšel	Kumyk jaʃɯl
Northern Uzbek jaʃil	OldTurkic yašïl	Salar ja:ʃɯl	Tatar jaʃel
Turkish jeʃil	Turkmen jaʃɯl	Uyghur jeʃil	Uzbek yašil

s·r|rotten · binary · 18 languages · enrichment 58×

Azerbaijani čürü	Bashkir ser	Chuvash śər	CrimeanTatar čürü
Gagauz čürü	KK-Halqabad 1 šir	KarachayBalkar čir	Karaim ciri
KazanTatar čer	Khalaj ʧirry	Ki-Akmuz 1 čir	Kumyk čiri
Nogai širi	NorthAltai čiri	Salar čür	Turkish čürü

q·l·n|thick · ternary · 18 languages · enrichment 844×

Bashkir qalɯn	CrimeanTatar ḳalïn	KK-Halqabad 1 qalən	KarachayBalkar qalɯn (2D)
Kazakh qalən	KazanTatar ḳalïn	Ki-Akmuz 1 qalɯn	Kumyk qalɯn
NorthAltai qalɯn	Northern Uzbek qalin	OldTurkic ḳalïn	Shor qalɯn
Tatar qɑlɤn	Tuvan qɯlɯn (2D)	Uyghur qelin	Uzbek ḳalin

q·s·q|short · ternary · 17 languages · enrichment 396×

Bashkir qɯsqa	CrimeanTatar ḳïsḳa	KK-Halqabad 1 qɨsqa	Kazakh qsqa
KazanTatar ḳïsḳa	Ki-Akmuz 1 qɨsqa	Kumyk qɨsqa	Nogai qɯsqa
North Altay qɯsqa	NorthAltai ḳïsḳa	Northern Uzbek qisqa	Shor qɯsqa
Tatar qɤsqɑ	Tofa ḳïˁsḳa	Tuvan qɯsqa	Uyghur qisqa

t·p|BOTTOM · binary · 16 languages · enrichment 216×

BarabaTatar tüp	Bashkir töp	Chuvash tӳp	CrimeanTatar tüp
KK-Halqabad 1 typ	Karaim typ, kip etc.	Kazakh typ	KazanTatar töp
Khakas typ	Ki-Akmuz 1 typ	Kumyk typ	Nogai typ
North Altay typ	NorthAltai typ	Uyghur typ	WestYugur tïp

Other

s·k·r|JUMP · ternary · 15 languages · enrichment 408×

CrimeanTatar sekir	KK-Halqabad 1 sekir	KarachayBalkar sekir	Karaim sekir
Kazakh sekĭr	KazanTatar siker	Ki-Akmuz 1 sekir	North Altay sekir
NorthAltai sekir	Northern Uzbek sakramɒq	OldTurkic sekri	Salar seker
Tatar sikerʉ	Uyghur ? sɛkrɛ	Uzbek sakra	

k·j|PUT ON · binary · 15 languages · enrichment 326×

Azerbaijani kej	BarabaTatar käy-	Bashkir kej	KK-Halqabad 1 kij
KarachayBalkar kij	Karaim kij	Kazakh kij	Ki-Akmuz 1 kij
Nogai kij	North Altay kij	NorthAltai kij	Northern Uzbek kijmɒq
Salar kij	Tatar kijʊ	Uyghur kij	