Consonant Codes in 111 Afro-Asiatic Languages

The family where the consonantal root was already a theory

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

Abstract

We present a catalogue of 2,121 consonant codes grouping 11,082 lexical forms from 111 Afro-Asiatic languages. A code is a consonant skeleton shared by words for the same concept across distinct languages. Its construction uses no etymology, no proto-forms and no cognacy judgements: only the transcribed form, the concept it expresses, and the requirement that the coincidence occur in at least three languages.

The procedure groups 17.4% of the available forms, with a precision 44 times above chance (5.26% against 0.119%). On grouping without the three-language threshold, the advantage is 25.0 points — 29.7% observed against 4.7% for the null, 6.3 times chance.

This family was chosen for a reason no other in the series offers: it is the only one where the consonantal root was already a theory before we arrived. The Semitic triliteral root — *k-t-b* 'write' — is exactly our object, described by another route and more than a thousand years ago. For the first time the instrument has something to be tested against other than itself.

The contrast was pre-registered before anything was run. Its principal result is confirmed: Semitic concentrates 50.6% of its forms in skeletons of exactly three consonants against 31.3% for Chadic (+19.3 points, $p = 0.0001$ with a null permuting branch over languages). And it is not that its words are shorter: mean consonants per form are 2.99 · 3.13 · 3.02 · 3.02 across the four branches. It is concentration, not length. A procedure blind to grammar recovers the shape of the triliteral root without having used it.

Two hypotheses from the same pre-registration are not confirmed, and they diagnose the instrument rather than the family. They are published in the same detail.

The codes of widest reach reproduce the vocabulary Afro-Asiatic comparison had already established — `m·t` for *die* in 38 languages across two branches, `d·m` for *blood* in 29, `s·m` for *name* in 25 — without having reconstructed anything.

1. What is sought, and why Afro-Asiatic

This work continues a series applying the same instrument to different families. 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:

A code is a consonant skeleton that appears for the same concept in three or more distinct languages. It is not claimed that its members are cognate, that they descend from a proto-form, or that the code means anything. A group is a group: two observable facts — form and concept — and nothing else.

And this family was chosen for a reason no other in the series offers. In Austronesian, Turkic, Uralic, Pama-Nyungan and Indo-European the consonant skeleton was an object of ours: we defined it, mea-

sured it, and tested it against its own null. Not here. The Semitic triliteral root is a consonant skeleton, described by Arabic and Hebrew philology more than a millennium ago, discovered by a route with nothing to do with ours.

That makes Afro-Asiatic the only place in the corpus where the instrument can converge with, or depart from, an external description. It is the most informative test in the series, and also the riskiest: if the skeleton does not recover the root where the root exists and is described, the problem is the skeleton's.

1.1 What this paper does NOT do

It does not reconstruct Proto-Afro-Asiatic and does not consult it. It does not use the cognacy judgements that exist for this family. It proposes no etymologies. And it does not claim the codes of widest reach are cognates, though many are: the catalogue cannot distinguish inheritance from contact, and will not pretend otherwise.

What it does do: build the catalogue as in the five previous families, and then measure the contrast with the triliteral root as a pre-registered hypothesis, with its null, its declared direction and its falsifiers written before looking.

2. The population, and the decisions that fix it

2.1 What enters

111 languages · 63,768 forms with a skeleton · 2,031 concepts · 4 branches.

Branch	Languages	Forms with skeleton
Chadic	80	47,202
Semitic	27	12,162
Berber	2	2,262
Cushitic	2	2,420

72% of the languages are Chadic, and two branches have two languages each. This governs what the paper can sustain and is said here, before the results, not after.

Omotic and Egyptian are missing. The corpus does not carry them for this family. The catalogue does not cover Afro-Asiatic: it covers the four branches the corpus contains, and should be cited that way.

2.2 One language excluded, and twenty-nine that are not

Phase 0 excludes **`Bara`**, declared Chadic. The corpus holds two lects of that name, both with 414 forms and both with coordinates in southern Madagascar: one Austronesian (`bara1369`) and one Afro-Asiatic (`bara1395`). Over the 207 shared concepts, 66% of the forms are identical, and those that differ do so as two surveys of the same language. The sample leaves no doubt: `biby` 'animal', `zaza` 'child', `velo` 'alive', `satria` 'because', `manapaky` 'cut'. It is Malagasy. An Austronesian wordlist attached to a Chadic glottocode.

Two clarifications about this exclusion, because the 66% invites a misreading. First: it is *not* a duplicate under the series' mechanical rule — same glottocode, or more than 80% identical skeletons — which `Bara` does not meet: the glottocodes differ and 66% is below the threshold. It is excluded as `mal_`

asignado, a different class with different evidence: the vocabulary and the coordinates. The 66% corroborates; it does not decide.

Second: excluding it does not change the results. With Bara in, the family's precision is 43× instead of 44×, and nothing else moves. The exclusion is correct and it is immaterial, and both are worth saying: a catalogue that documents only the decisions that change something hides the ones that do not.

And here we must report a check we deliberately failed. The same phase-0 procedure proposed excluding 30 of 112 languages by distance from the family's geographic centre. Among them: Standard Arabic, Modern Hebrew, Amharic, Ge'ez, Maltese, Tigrinya, Tarifit. That is, almost all of Semitic and all of Berber.

The check is not broken: it measures what it was asked to, and what it was asked to makes no sense here. The family's centre falls at 10.8°N · 13.8°E — the Chadic heartland — because 72% of its corpus languages are Chadic. Measured against that centre, a family stretching from Morocco to Oman declares its own northern half an intruder. In Uralic, a compact family, that same check caught a Tai-Kadai language 6,000 km out; here 6,000 km is Bukharan Arabic, which is legitimate.

The 29 are rejected. And it is written down, because it holds for the whole series: distance from the centre is informative in compact families and says nothing in extensive ones. The check should declare the family's dispersion alongside the list, and abstain from proposing when the median is high.

Applying those 29 exclusions would have deleted Semitic, which is the reason this family was chosen.

2.3 The forms that do not enter, and cost no coverage

Of 68,823 forms, 4,777 obtain no skeleton (7%), and they are exactly the 4,777 from the IDS source: verified source by source, lexibank 62,239 of 62,239 with a skeleton, nel 2,221 of 2,221, ids 0 of 4,777. This is not a silent drop in the pipeline: it is the scope decision already standing in the series — IDS supplies orthography and not IPA — making itself visible.

Zero concepts are lost entirely. Three languages depend on IDS in part — Ancient Aramaic 50%, Polci 37%, Hausa 28% — and none falls to zero.

One practical consequence: the citation-form check, which in Turkic uncovered the costliest phase-0 finding — the -mAK infinitive that some sources supply and others do not — here flags endings (-tā, -rā, -nā, -lā, between 2% and 6%) that are all from IDS. They do not enter the catalogue. It is noise from a source that does not take part.

2.4 Two declared transcription artefacts

Ethiosemitic labialisation written as a digraph. In Mesqan, Ge'ez and Kistane, gwəbbe = /gʷəbbe/, kwəll = /kʷəll/ 'all', čəkkwərə: the labialised consonant is written Cw and the skeleton counts the w as a separate consonant. Some twenty forms. It is declared and not corrected, because distinguishing it from a genuine /kw/ cluster — Maltese 'gwɛrre, an Italian loan; Arabic lwara:ʔ, article plus root — would require a per-language decision the corpus does not support.

The glottal modifier counted as a consonant. ʾ and ʿ are modifier letters — glottalisation, pharyngealisation — and not consonants. The sources' segmenter turns them into their own ʔ/ʕ segment when they follow a vowel and absorbs them when they follow a consonant. 248 Afro-Asiatic forms in 17 languages fell into that case. It was corrected in the corpus before this catalogue was built, leaving untouched the word-initial glottal onsets, which in Chadic may be full consonants.

3. What the instrument costs here

Collision: 38.1%. The percentage of forms whose skeleton, *within its own language*, covers more than one concept. It is the pure cost of the instrument, with no genealogy involved: how much information is lost simply by removing the vowels.

Mean arity: 3.16 consonants per form. The family sits in the cheap half of the series: above Sino-Tibetan (2.22) and Austronesian (2.79), below Pama-Nyungan and Indo-European (3.35).

This is coherent with what the family is. A morphology whose canonical root has three consonants produces words with three consonants, and three consonants are enough to distinguish: 85% of the resulting codes have three or fewer.

Nine strings cover five concepts or more within one language, and not all are defects of the data. `kwam` in Polci covers eight concepts that are all 'father-in-law / mother-in-law / daughter-in-law' — real colexification of a kinship system; `diɓɨkidì` in Miya covers seven decades; `dáukàa` in Hausa covers seven ways of carrying something. They are reported so that collision can be read knowing how much comes from there.

4. The catalogue

2,121 codes over 2,031 concepts, grouping 11,082 of 63,768 forms (17.4%). Of these, 1,732 are binary or ternary (81.7%), with median enrichment of 100× and a median of 3 languages per code.

All those figures depend on a threshold that is convention and not finding, and it is worth showing how much:

Language threshold	Codes	Forms grouped	Coverage
2	5,651	18,960	29.7%
3	2,121	11,082	17.4%
4	1,032	7,439	11.7%
5	620	5,606	8.8%
6	424	4,509	7.1%

Each step roughly halves the catalogue, and there is no elbow. No threshold is discovered in the data: three is the series' convention — two languages do not distinguish coincidence from repetition — and it is kept for comparability across families, not because these data point to it.

What is worth retaining is what depends on it and what does not. Coverage, the number of codes and the whole of §6 depend. The paper's principal result, H1, does not: it is measured over forms and their skeletons, without passing through the catalogue, so it is indifferent to the threshold.

Arity	Codes	%
1	157	7.4%
2	830	39.1%
3	902	42.5%
4	191	9.0%
5+	41	1.9%

Precision: 5.26% against 0.119% chance = 44× [44–44× across seeds]. The ratio is given decomposed on purpose: observed precision is low in absolute terms, and what makes it informative is what it is compared to.

And it has to be decomposed once more, because the figure is not homogeneous. The series plan warns that a high × `chance` may measure the corpus's redundancy — many close varieties of the same material — before it measures the instrument's accuracy. That happens here:

	Precision	× chance
Whole family	5.26%	44×
Chadic only (80 lgs)	5.08%	34×
Semitic only (27 lgs)	18.68%	175×
Family without the 13 Arabic lects	4.40%	36×

Semitic alone yields 175×, four times the family figure, and removing the thirteen Arabic lects drops the headline from 44× to 36×: 18% of the headline comes from that block. Eight Arabic dialects sharing a common concept list share skeletons because they are varieties of the same thing, not because the method is more accurate on them.

The 44× is correct and is the family's figure as the corpus supplies it. But it is not read alone: it is read beside the 36× without the Arabics, which is the part of the figure that does not depend on how many close varieties the source happened to collect.

And on grouping without the three-language threshold, which is the measure that compares to the null directly: 29.7% observed against 4.7% null — an advantage of 25.0 points, 6.3× chance.

4.1 The codes of widest reach reproduce the classical comparanda

These are the fifteen widest-reaching codes in the catalogue, unfiltered, because the unfiltered list says more than a selection:

Code	Concept	Languages	Arity	Dominant branch
`m·t`	*die*	38	2	Chadic
`m`	*water*	33	1	Chadic
`l`	*cow*	31	1	Chadic
`d·m`	*blood*	28	2	Semitic
`m·b·j`	*cassava*	27	3	Chadic
`b·b`	*father*	26	2	Chadic
`s·m`	*name*	25	2	Semitic
`s`	*drink*	25	1	Chadic
`s·n`	*tooth*	22	2	Semitic
`l·m`	*name*	22	2	Chadic
`t·r`	*moon*	22	2	Chadic
`l·m`	*ear*	21	2	Chadic
`j·m`	*water*	20	2	Chadic
`f·d`	*four*	20	2	Chadic
`v·j`	*rainy season*	20	2	Chadic

There are three distinct classes in there, and conflating them would be the easy error.

Established comparanda. `m·t` is pan-Afro-Asiatic *mwt* 'die', and it is the widest-reaching code in the whole catalogue; `d·m` is Semitic *dam* 'blood'; `s·m` is *šm* 'name' — Arabic *ism*, Hebrew *šem*; `s·n` is *sinn* 'tooth'. The procedure did not look for them: they emerged from requiring a skeleton to repeat for the same concept across three languages.

Unary codes, and here precision matters because the easy answer is false. `m|water` in 33 languages, `l|cow` in 31, `s|drink` in 25. The comfortable thing to say is that a single consonant coincides by

availability rather than by history. But `l|cow` is *sla*, *łâ*, *łà*, *slà*: the lateral fricative **ɬ** that the normalisation writes **l**, and that *ła* 'cow' is a Chadic etymon in its own right. The code is real; it just has one consonant.

The problem, then, is not that unary codes are false: it is that the catalogue cannot distinguish a genuine monoconsonantal etymon from a coincidence of availability, because both give it the same evidence. That is why the paper works with binaries and ternaries — 81.7% of the catalogue — and counts the unaries without reading them. They are not discarded as false; they are set aside as indistinguishable.

And a modern loan, which the catalogue does not distinguish from the rest. `m·b·j|cassava` reaches 27 languages. Cassava is an American crop introduced into Africa in colonial times: that code is neither inheritance nor ancient contact, it is a recent loan travelling with its word. The catalogue groups it exactly as it groups **m·t**, and has nothing with which to separate them. It is the best example of the limit §1.1 declares, and that is why it stays in the table instead of being removed.

What the section does sustain, then, is narrower: where an established, wide-reaching correspondence exists, the catalogue contains it. What it does not sustain is the converse — that everything of wide reach is a correspondence — and the cassava is there to remember it.

5. The spine: the contrast with the triliteral root

Everything that follows was pre-registered before anything was run — the document is listed under Data and reproducibility below — with its declared direction and its falsifiers written down. It was done that way because the failure mode diagnosed in earlier work was "headline first, null afterwards", and here the temptation is at its greatest: Semitic philology has spent a thousand years saying what ought to come out.

And the pre-registration has two weaknesses better said here than left for the reader to find.

First: its falsifiers fix no magnitude. The one for H1 says "if Semitic does not concentrate on 3 more than Chadic, the contrast does not exist". Written that way, any positive difference confirms: 33% against 31% would have passed. A pre-registration with no declared minimum effect cannot fail against a small but correctly-signed result, and that is half of what a pre-registration is for. It did not matter here because the difference was 19.3 points with complete separation between languages, but it did not matter by luck, not by design.

Second: H1 was the least risky of the three. That the Semitic root has three consonants is known to anyone who has opened an Arabic grammar. What was in doubt was not Semitic but whether our skeleton would recover it, and that is a test of the instrument, not of the language — the paper says so in §8, but the pre-registration apparatus lends H1 an air of risk it did not carry. H2 and H3 were genuinely uncertain, and both failed. Registering three hypotheses and confirming one sounds like discipline; it is worth adding that the one that survived was the one with least at stake.

5.1 The prediction, which inverts the obvious one

The naive reading would be "the instrument will work better in Semitic". The registered prediction is the opposite, for a concrete reason: Semitic morphology is templatic. Its affixes are interdigitated vowels and patterns, not consonants attached at the edge. If the native theory is right, then the consonant skeleton of a Semitic word already is close to its root, and should concentrate on three.

5.2 H1 · confirmed, and it is the paper's result

Measure	Semitic	Chadic	diff.	p
Forms with a skeleton of exactly 3 consonants	50.6%	31.3%	+19.3	0.0001

The null permutes the branch label over doculects — 27 languages against 80, not 12,017 forms against 47,162 — with 10,000 permutations. This is the lesson of pseudoreplication: the unit of exchangeability is the language, not the form.

And that null still assumes too much, so a figure that assumes nothing is worth having. Permuting over languages treats the 27 Semitic ones as exchangeable, and they are not: eleven are Ethiosemitic with an identical concept list and eight are Arabic dialects. The effective n is smaller than 27, and with a much smaller n the p-value would soften.

The answer is to look at the language level directly. All 27 Semitic languages sit above the Chadic median. All twenty-seven. The Semitic range runs from 40.3% (Harari) to 62.4% (Egyptian Arabic); the Chadic range from 23.1% to 40.9%. Over the 2,160 language-to-language pairs, Semitic concentrates more in 2,158, an AUC of 0.999. The two pairs that go the other way are the same case: Harari, the lowest Semitic, against Mbazla, the highest Chadic.

That number assumes no independence between languages, no normality, nothing: it says that if one Semitic and one Chadic language are picked at random, the Semitic one concentrates more on three consonants 99.9% of the time. Even if the effective n were three or four independent groups, the separation would still be complete.

5.2-bis · And it is not an effect of the project that collected the data

One more objection remains, graver than the previous ones, and it must be answered with data. The corpus's Chadic and Semitic come from disjoint fieldwork projects: `kraftchadic` (61 languages) and `gravinachadic` (30) on one side; `ratcliffearabic` (14) and `felekesemitic` (11) on the other. Each project transcribes by its own conventions. The contrast between branches could be a contrast between teams.

The answer is that three projects cover both branches, and serve as an internal control:

Project	Chadic	Semitic	Difference
`gravinachadic`	31.1% (17,231 forms)	45.8% (753)	+14.8
`idssegmented`	31.7% (1,983)	55.1% (1,924)	+23.5
`polyglottaafricana`	29.6% (270)	47.1% (626)	+17.5

Within the same project and the same transcription practice, Semitic concentrates between 15 and 24 points more. And Chadic is remarkably stable across five independent projects — 29.6 · 31.1 · 31.2 · 31.7 · 32.3%: the project effect is around 3 points and the between-branch gap around 19.

The objection was correct and the answer is that it does not explain the result.

And the check that makes the number readable: they are not shorter words.

consonants	Semitic	Chadic	Berber	Cushitic
1	2.8%	5.6%	7.1%	4.0%
2	25.9%	30.7%	27.9%	32.5%
3	50.6%	31.3%	35.6%	36.1%
4	14.0%	17.4%	18.0%	16.1%
5	3.6%	7.4%	8.0%	7.2%
6+	3.0%	7.6%	3.3%	4.1%
mean	2.99	3.13	3.02	3.02

The means are practically identical across the four branches. What distinguishes Semitic is concentration: it puts half its mass on a single value while the other three spread theirs.

Berber (35.6%) and Cushitic (36.1%) fall in between, in the direction of the templaticity gradient the pre-registration anticipated. Nothing rests on them: they are two languages each, and that was declared before looking.

And there is a second route, not pre-registered and therefore offered as corroboration rather than proof. H1 measures forms; the same can be measured over the codes already built, asking which branch dominates each one by its arity:

Code arity	Chadic	Semitic	% Semitic
1	190	14	6.8%
2	868	150	14.7%
3	476	237	33.2%
4	116	32	21.6%
5	21	3	12.5%

The maximum is exactly at three, and ternary codes are 2.3 times more Semitic than binary ones. The curve rises to three and falls after it, which is the shape H1 predicts carried from the level of the form to that of the code.

With two cautions. First: two thirds of ternary codes are still Chadic (476 against 237). The claim is about relative proportions, not a partition. Second: this is not independent of H1 — codes are built from forms — so it corroborates without adding a new contrast.

Before reading this figure it was checked, also as pre-registered, that the citation form does not manufacture it: no ending exceeds 2.9% in either branch, and there is nothing like the infinitive convention that proved decisive in Turkic. One asymmetry is declared: Semitic draws on two sources and Chadic on one.

5.3 H2 · falsified

Measure	Semitic	Chadic	diff.	p
Forms that automatic segmentation modifies	65.6%	59.6%	+6.0	0.049

The prediction was Semitic lower than Chadic. It comes out the other way, and weakly: with two contrasts and a multiplicity correction it lands at 0.10. Nothing rests on the sign; what is claimed is that the prediction fails.

And the reading is not about the branches but about the instrument: a segmentation that modifies three of every five forms in both branches is not detecting morphology.

5.4 H3 · does not discriminate

The pre-registered list of pan-Semitic consonantal affixes was m-, t-, n-, ʔ-, with ʔ– declared untestable because the normalisation merges ʔ ʕ ħ h into one symbol.

The prefixes automatic discovery finds blind, by number of languages in which they appear: k– 63%, m– 63%, d– 56%, h– 56%, s– 48%, b– 44%, r– 37%, n– 33%, g– 33%, t– 30%.

m– is at the top. But k–, d–, b–, s–, r– and g– appear at the same level or above, and they are not Semitic affixes: they are root consonants. That m– comes first is compatible with its being real and with its being chance; with this table it cannot be decided, and it is not.

5.5 What the two fallen hypotheses teach

H2 and H3 say the same thing by two routes: automatic affix discovery over-marks. Measured, it marks as an affix 69% of each Afro-Asiatic language's consonant inventory.

And that finding generalises beyond this family: the same measurement in the previous five gives between 59% and 89%. In a family whose root is a trio of consonants, marking seven of every ten consonants as an affix means trimming the root.

Measured on grouping, which is the criterion the protocol declares:

Step	In a group	Null	Advantage	κ	Codes of 4+ consonants
Whole word	29.7%	4.7%	25.0	0.262	29.9%
+ discovered affixes	38.8%	17.5%	21.3	0.258	9.6%

Eight points of signal and thirteen of noise. The κ column is the Hubert-Arabie correction, `(observed - null) /(1 - null)`, which normalises by the attainable ceiling and allows comparison across families: here $\Delta\kappa = -0.004$. 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. *(List 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.)* The 4+-consonant column collapses from 29.9% to 9.6% and looks spectacular, but the criterion is not that arity falls — that is achieved by over-trimming — but that grouping rises above the null, and it does not.

That is why this paper's catalogue does not use automatic affix discovery, just as the rest of the series does not. The phase remains a diagnostic, and its result is published because it is informative, not because it is applied.

6. What this family allows, and what it does not

6.1 Genealogical candidates are few, and we say why

With four branches and 72% of the languages in a single one, a code crossing three branches is in practice a Chadic-Semitic code that also surfaces in one of the four remaining languages. The 349 candidates Pama-Nyungan yielded with fifteen balanced branches do not recur here, and not through any fault of the method.

What does work is the principal division, Chadic against Semitic, which splits 80 and 27 languages: 14,328 substitutions cross that cut, 15.7% of the total, and yield correspondence candidates with their concentration share.

6.2 The anomaly list does exist: two macroareas

This is one of the five families in the corpus with more than one macroarea — the others are Indo-European, Austronesian, Arawakan and Chibchan. 393 codes (7.0%) cross macroareas and are the ones deserving a second look: coincidence of form and sense between languages that share no neighbourhood.

6.3 Substitution pairs: 513, of which 450 usable

Pairs of codes for the same concept differing in exactly one position. Of the 513, 450 are disjoint in languages — no language is in both — and only those can be proposed as an isogloss. The 63 that overlap are not erroneous: a language may have two words, the same root with two suffixes, or two transcriptions.

Those of widest reach: `l s` in *name* (Chadic `l·m` in 22 languages against `s·m` in 25, Semitic and Chadic), `c t`, `s t` and `r t` in *die* against the majority `m·t`, `l r` in *oil*, `b h` in *father*.

None was sought. They emerged from counting differences in pairs already aligned by concept and code.

6.4 Emergent classes, with validation on unseen languages

The 111 languages are split in two halves: 55 to discover and 56 to check. The pairs surviving in both, with their ratio over chance:

Pair	Times (discover)	Ratio	Times (check)	Ratio
`f ↔ v`	71	6.9×	266	18.3×
`c ↔ t`	158	6.4×	281	8.0×
`l ↔ ð`	114	8.7×	143	7.7×
`c ↔ s`	287	8.0×	386	7.6×
`b ↔ v`	95	4.7×	202	7.1×
`f ↔ p`	118	7.0×	165	6.9×
`h ↔ x`	184	4.8×	290	5.4×
`b ↔ p`	201	6.0×	244	5.2×

These are alternations of voicing and manner within the same place of articulation. They are not regular correspondences — environment and direction are missing — but candidates, with the place to look for them.

6.5 Geography works

Languages sharing more codes are closer together, and the relation is not flat:

Codes shared	Pairs	Mean distance	vs chance
1	1,369	1,827 km	1.16
2	701	1,139 km	0.72
3–5	1,018	608 km	0.39
6–10	463	469 km	0.30
11–20	176	766 km	0.49
21+	216	907 km	0.58

The minimum sits at 6–10 shared codes, at under a third of the distance expected by chance.

That the curve rises again from 11 upward has a concrete explanation, and it is not the obvious one. It is not pan-family codes appearing everywhere: it is which branch fills each band. Counting the composition of the pairs:

Codes shared	% Semitic-Semitic pairs	% Chadic-Chadic
1	1.9%	36.4%
2	4.5%	54.8%
3–5	6.5%	77.2%
6–10	6.7%	89.5%
11–20	17.8%	77.7%
21+	41.3%	56.5%

The distance minimum coincides with the most Chadic band (89.5%), and Chadic occupies a compact area between Nigeria and Chad. The band of highest overlap fills with Semitic pairs (41.3%, twenty times the share in the first band), and the corpus's Semitic runs from Morocco to Oman and Ethiopia.

Put another way: the curve does not measure a relation between similarity and distance, but the uneven geographic spread of the two large branches. Chadic being compact and Semitic dispersed is enough to produce this shape with nothing to interpret about the codes. The section stays because its first half — more shared codes, less distance, up to 6–10 — is informative within a branch; the second half is an artefact of how the family is sampled.

7. Limits

The catalogue does not cover Afro-Asiatic. It covers Chadic, Semitic, Berber and Cushitic. Omotic and Egyptian are missing, and Berber and Cushitic enter with two languages each, which is presence and not a sample.

The spine runs on 27 languages and 12,162 forms, a fifth of the family. Every claim about the contrast with the triliteral root carries that denominator.

The corpus's Semitic has a median of 139 concepts per language, the lowest of the six families in the series — Uralic has 970. Eleven Ethiosemitic languages share an identical 139-concept list and eight Arabic dialects bring about 100. This is not a filter of ours: it is the length of the source wordlists, and there is nothing to recover — IDS adds zero new concepts and kaikki's two Semitic entries are empty.

The normalisation merges consonants that are distinct in Semitic, and there are two mergers, not one.

The first, ʔ ʕ ħ h into a single symbol h. It weighs 5.5% of Semitic segments against 1.9% of Chadic ones: it is asymmetric between the two branches being contrasted and favours Semitic in any grouping measure.

The second weighs more, and it is the one that was missing from this list: **s** gathers **ʃ**, **ʒ**, **tʃ**, **dʒ**, **ts** and the other sibilants, 6.9% of Semitic segments against 3.9% of Chadic ones. In Semitic the *s* / *š* opposition is not a phonetic detail: it is the shibboleth's own distinction, and it separates roots. Measured over the catalogue, 63 of the 418 Semitic codes — 15.1% — gather forms with **s** and forms with **ʃ** inside the same code. One in seven exists because the table writes them alike.

Neither affects H1, which counts positions and not identities — and that is why the paper's principal result is the one that depends least on the normalisation. Both affect the catalogue, the enrichment and the substitution pairs, and are declared here rather than in a note.

The skeleton does not capture tone. Over the forms entering the catalogue no mark appears, in digits or diacritics, even though the global profile gave 47% diacritic tone for this family: that figure comes from the forms that do not enter. It is declared and not pretended to be measured.

8. Conclusion

Afro-Asiatic was, in the design of the series, the family where the instrument could break against an external description. It did not break: the consonant skeleton recovers the shape of the triliteral root without using it, with a concentration on three consonants that doubles its sister branch's and that is not an effect of word length.

It is a convergence, not a discovery. The triliteral root was already discovered. What this work adds is that a procedure knowing nothing of Semitic grammar, running over transcriptions of vocabulary lists, arrives at the same shape. For an instrument whose declared problem is that outside Indo-European it rests only on itself, that is exactly the external support it lacked.

And the two hypotheses that fell are worth as much as the one that held. Automatic affix discovery marks 69% of the consonant inventory as affixal, here and in the five previous families. In a family whose root is a trio of consonants, that means trimming the root — and without this family, without an external theory to measure it against, the failure would have gone on passing for a gain.

References

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

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

Frisch, S. A., Pierrehumbert, J. B. and Broe, M. B. (2004). Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22(1). — [the gradient quantification of the constraint on Semitic roots]

Greenberg, J. H. (1950). The patterning of root morphemes in Semitic. *WORD* 6(2). — [the classical constraint on the composition of the triliteral root]

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

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

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

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

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

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

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

Toledo Martínez, A. (2026). *The catalogue of consonantal codes: a comparative method without genealogy, cognacy or reconstruction*. — [the method of the series, described once]

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	`familias/afro-asiatic/output/catalogo-afro-asiatic.tsv` — 2,121 rows
Per-form detail	`familias/afro-asiatic/output/catalogo-afro-asiatic-lenguas.tsv`
Hygiene (§2.2)	`pipeline/fases/fase0_higiene.py` → `runs/fase0-higiene.md`
Data decisions (§2.2)	`runs/fase0-decisiones.md` · `ci/decisiones.py`
False segments (§2.4)	`pipeline/fases/fase0_segmentos_falsos.py` → `runs/fase0-segmentos-falsos.md`
Normalisation table (§3)	`pipeline/fases/tabla_normalizacion.py` → `runs/normalizacion.md`
Loading, population, warnings	`cod/familia.py`
Catalogue (§4)	`pipeline/fases/catalogo_familia.py`
Grouping and nulls (§4)	`pipeline/fases/agrupamiento_familia.py` → `runs/agrupamiento.md`
Spine pre-registration (§5)	`runs/preregistro-espina-semitica.md` — written before anything was run
Branch contrast (§5.2–5.3)	`pipeline/fases/contraste_por_rama.py` → `runs/contraste-semitico-chadico.md`
Reference truth (§5.4)	`runs/verdad-afijos-semiticos.md` — written before the sweep
Spine results (§5)	`runs/resultados-espina.md`
Automatic segmentation (§5.5)	`pipeline/fases/fase2_despiece.py` → `runs/fase2-despiece.md`
Sibling codes (§6.3)	`pipeline/fases/codigos_hermanos.py` → `runs/codigos-hermanos.md`
Emergent classes (§6.4)	`pipeline/fases/fase6_clases_emergentes.py` → `runs/fase6-clases.md`
Substitutions by branch (§6.1)	`pipeline/fases/fase6_por_rama.py` → `runs/fase6-por-rama.md`
Geography (§6.5)	`pipeline/fases/codigos_vs_parentesco.py` → `runs/codigos-parentesco.md`

Every figure in this paper comes from those scripts over the same population declared in §2.1, with the same exclusions. None was written by hand.

Appendix A · One hundred codes, with commentary

The catalogue holds 2,121 codes and the complete data are in `familias/afro-asiatic/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 things about the presentation. Languages are spread along the alphabetical list, not taken from its head. Long names are trimmed; the forms are not altered. And the thematic grouping is a reading order declared by hand, not a semantic classification.

This appendix carries no emergent-code column. The phase-6 classes come out here as pairs — `f↔v`, `c↔t`, `l↔ð`, `b↔p` — and not as a closed partition of the inventory, so writing them beside each code would imply a precision they do not have. §6.4 publishes them as candidates, which is what they are.

And there is something this appendix shows, with a warning that must be read first. Going through the blocks, many ternary codes are recognisable triliteral roots — `l·s·n` for *tongue*, `b·t·n` for *belly*, `r·k·b` for *knee*, `k·l·b` for *dog*, `s·m·s` for *sun*, `k·b·r` for *big*. They were neither sought nor annotated: they are codes the procedure produced.

The warning: this appendix is not a representative sample of the catalogue, and it is biased in the direction that most favours that reading. Its hundred codes are chosen by reach, and reach favours Semitic because the corpus's Arabic dialects share a common concept list and therefore reach more languages per code. Measured:

	full catalogue	this appendix
codes with a Semitic dominant branch	21%	41%
among the ternaries	33.2%	62.9%

That is: the appendix doubles the Semitic share. The claim the catalogue does support is narrower and is measured in §5: ternary codes are more Semitic than binary ones — 33.2% against 14.7%, a factor of 2.3, with the maximum exactly at three consonants — and two thirds of the catalogue's ternaries are still Chadic. A reader who sees only the appendix will take away a stronger impression than the data license, and that is why it is said here and not in a footnote.

Auditing the appendix: which codes depend on the table, and how many are loans

Before the blocks it is worth giving the result of auditing the hundred codes one by one, because it changes how they read.

Nineteen of the ninety-one gather within themselves symbols the normalisation merges — `s` with `ʃ`, or `h` with `ʔ ʕ ħ`. But two cases must be separated:

In ten, the merger only relabels. All or nearly all their forms carry the same merged symbol: `s·m·s|sun` has a sibilant in 19 of 21 forms, `l·h·m|meat` a guttural in 13 of 14. Undoing the merger would leave them equally grouped, just written differently.

In nine, the merger is load-bearing: the code exists because the table writes alike two consonants the languages distinguish.

Code	Concept	Merger	Split	Example
`s·n`	tooth	sibilant	9 / 20	`ʃɛn` against `siŋ`
`p·s`	sun	sibilant	7 / 11	`pats` against `pɨs`
`w·h·d`	one	guttural	13 / 6	`weeħ_di` against `wə́hīdu`
`h·s·r`	stone	guttural	10 / 5	`ħad͡ʒar` against `háṣar`

Code	Concept	Merger	Split	Example
`s·n`	dream	sibilant	6 / 7	`šinè` against `sunà`
`s·t·n`	demon	sibilant	6 / 7	`shetan` against `setan`
`s·t`	pepper	sibilant	12 / 4	`šìta` against `dṣ́ita`
`p·l·s`	horse	sibilant	5 / 13	`pɨlɪšà` against `plís`
`h·m`	water	guttural	8 / 4	`a'am` against `həmə`

The case of **`s·n|tooth`** is worth reading slowly, because it is not a failure: the *s š* alternation in 'tooth' is a real Afro-Asiatic correspondence, and a comparatist would group those forms just as we do. The difference is how: he would do it by demonstrating the correspondence; we do it because the table writes them alike before looking. The result agrees and the grounds do not, and that distinction is the whole paper.

And a second finding from the audit that we were not looking for: the appendix holds more loans than §4.1 flags. To the cassava must be added, among these nine, `s·t·n|demon` — which is *šayṭān*, of Islamic diffusion — plus `s·t|pepper` and `p·l·s|horse`, words for a crop and for material culture that travel with the thing. The catalogue groups them exactly as it groups **`m·t|die`**, and has nothing with which to separate them. Read the appendix knowing that at least four of its hundred codes are things that arrived, not things that were inherited.

Reverse index: codes recurring across concepts

Only three codes in the appendix appear with more than one concept, and that is a low figure for the series — Turkic had ten concepts under a single code. It is coherent with the family's arity: with three consonants on average, the skeleton discriminates more than in a family of short words.

The three read differently, and the difference is worth separating because it runs through the whole appendix.

`m·n` covers 'who', 'what' and 'how many'. That is not instrument collision: it is the Semitic interrogative stem, the same *man* readable in Amharic, Ancient Aramaic, Argobba and the Arabic dialects of the block. A language using the same material for all three questions is colexifying, and the code reflects it rightly.

`l·m` covers 'name', 'ear' and 'market', and here suspicion is due. Three unrelated fields, and the code is binary over an inventory where `l` and `m` are among the most frequent symbols. It is the typical case of the wide-reaching binary: it groups by availability as much as by history.

`k·m` covers 'face', 'ten' and 'ear', with low enrichments — 42×, 37×, 30× — the lowest in the appendix. A low enrichment on a wide-reaching code is exactly the signal that the code is picking up what is frequent, not what is shared.

That is the distinction to hold while reading everything that follows: one code may gather the same word used for several things, or several words the instrument cannot tell apart, and the code alone does not say which. The 38% collision declared in §3 is this, counted.

Code	Concepts it appears with, and their enrichment
`l·m`	name (22, 90×) · ear (21, 71×) · market (12, 75×)
`m·n`	who (18, 66×) · what (12, 45×) · how many (10, 52×)
`k·m`	face (11, 42×) · ten (10, 37×) · ear (10, 30×)

Numerals

The numerals split the family in two, and they do it visibly.

The Chadic ones go binary and the Semitic ones ternary. 'Four' appears twice: `f·d` in twenty languages — *fúdu*, *fidi*, *faɗ*, *fódu*, *fuɗu* — and `f·w·d` in thirteen — *fwad*, *fwəɗà*, *fwòɗu*, *fwaɗ* — and both lists are entirely Chadic. The difference between the two codes is the `w`, that is the labialisation of the first consonant, written sometimes as a segment and sometimes not. It is the same numeral split by a transcription convention, the same class of artefact §2.4 declares for Ethiosemitic.

Against that, `w·h·d` for 'one' and `h·s·r` for 'ten' are textbook Semitic triliterals — *wāḥid*, *ʕašar* — with twelve languages each.

`k·m|ten` is worth reading beside **`h·s·r|ten`**. The same concept comes out with two codes, one binary at 37× enrichment and one ternary at 164×, and they do not compete: they are the two halves of the family counting in different ways.

`f·d|four` · binary · 20 languages · enrichment 173×

```
Bade fúdu   Banana fiɗi    Banana fidi  Cuvok faɗ
D□ ai fúdū   Hausa fódu     Hwona faɗà   Karekare fə̰́du
Lame fúdī   Mandara úfade  MargiS foɗu  Matakam faɗ
Misme fidi  Miya fɨɗì      Ngizim fuɗu  Tangale faɗàu
```

`f·w·d|four` · ternary · 13 languages · enrichment 569×

```
Bata fwad          Boka fwəɗà     Cibak fwòɗu   Fali Gili fwaɗ
Fali Mucella fwəɗ  Ga?anda fwəɗà  Gudu fwád     Higi Ghye fwaɗɛ
Kamwe-Futu fwaɗu   Margi fwoɗù    MargiS fwɔ̀dù  Psikye fwaɗɛ
West Margi fwoɗo
```

`h·s·r|TEN` · ternary · 12 languages · enrichment 164×

```
Amharic ʔassɨr  Argobba ʔassɨr  Ge'ez ʕəṣru    Harari ʔassir
Kistane ʔasɨr   Mesqan ʔasɨr    Silt'e ʔassɨr  Standard Arabic ʕaʃara
Tarifit ʕəšr'a  Tigre ʔasɨr     Wolane ʔassɨr  Zway ʔassɨr
```

`t·f|five` · binary · 12 languages · enrichment 207×

```
Bacama tuf        Bata túf     Cibak tufù  Fali Mucella tuf
Fali Ɓwagira tuf  Gudu tùf     Hwona tufù  Kilba tûfù
MargiS tufu       Ngwaxi tufù  Njanye tuf  West Margi tufù
```

`w·h·d|one` · ternary · 12 languages · enrichment 297×

```
(Arabic) Ádirar w□ hid    (Arabic) B□ ṛān wə́hid       BuxariArabic waaħid
DamasceneArabic wa:ħ_ed  Egyptian Spoken A wa:ħid   IraqiArabic wa:ħi_d
Maltese wi:ħ_ed          MoroccoArabic wa_ħed       Shoa Arabic wa:hid
Standard Arabic wa:ħid   United Arab Emira waaħ_id  YemeniticArabic wa:ħi_d
```

`m·k·r|three` · ternary · 12 languages · enrichment 258×

```
Bura makùř     Cibak makùř     Kilba mákrù        Margi makùř
MargiS maakər  Matakam màkâar  Mofu-Gudur maakar  Moloko makar
Muyang makar   Ngwaxi makùř    Ouldeme màkàr      West Margi makùr
```

k·m|TEN · binary · 10 languages · enrichment 37×

Boka kum	Bura kuma	Cibak kuma	Ga?anda kum
Hwona kûm	Kilba kúmà	Margi kumu	MargiS kumou
Ngwaxi kuma	West Margi kuma		

Body parts

The densest block in the appendix, and the one that best shows the paper's thesis.

Seven of its eleven codes are ternary, and all are recognisable triliterals: l·s·n *lisān* 'tongue', b·t·n *baṭn* 'belly', r·k·b *rukba* 'knee', s·h·r *šaʕr* 'hair'. The procedure did not know they existed; they emerged from requiring a skeleton to repeat for the same concept across three languages.

And **l·s·n|tongue** has a Chadic language inside it: Karekare *lusùŋ*, among Arabics and Aramaics. The catalogue groups it because it shares form and concept, and cannot say whether that is inheritance, loan or coincidence. It is exactly the kind of case §1.1 declares out of scope, and it is left in view rather than hidden.

b·t·n|belly has the block's highest enrichment, 666×, on a ternary of eleven languages: the combination b, t, n in that order and for that concept is extremely rare by chance. Compare l·m|ear, binary over twenty-one languages and only 71×. Reach and enrichment do not measure the same thing, and the appendix gives them separately on purpose.

d·m|blood · binary · 28 languages · enrichment 100×

(Arabic) Ádirar dámu	(Arabic) B□ ṛān dam	Ancient Aramaic dmā	Bole(wa) domì
CypriotArabic d-m-_y	DamasceneArabic dam	Ge'ez dəm	IraqiArabic Dam
Kistane dəm	Maltese 'dɛmm	Modern Hebrew dam	Ngamo dɔ̀m
Silt'e dəm	Standard Arabic da_m	Tigrigna dəm	YemeniticArabic Dam

s·n|tooth · binary · 22 languages · enrichment 47×

(Arabic) Ádirar sin	(Arabic) B□ ṛān sin	Ancient Aramaic šennā	Banana siino
DamasceneArabic sinn	Dira šinè	Egyptian Spoken A sinn	Geji sìŋ
Maltese 'sɪnnɐ	Mazera sine	Mbara siŋ	Mpade shán
Polci shen	Seya šɨn	Standard Arabic sinː	United Arab Emira sinn

l·m|ear · binary · 21 languages · enrichment 71×

Afade slɨm	Bura łîm	Buwal zlɑm	Cibak łɨmà
Dugwor zlam	Fali Kiria łɨmu	Gava łimà	Gudu łĭm
Higi Ghye łɨme	Kamwe-Futu łɨmu	Malgbe slɨm	Margi łɪmì
Mbara slumo	Merey zləm	Moloko zlam	Ngwaxi łɪm

h·n|eye · binary · 18 languages · enrichment 83×

Amharic ʔayn	Argobba ʔeen	BuxariArabic ʔayn	DamasceneArabic ʕeen
Egyptian Spoken A ʕeːn	Ge'ez ʕayn	Harari ʔin	IraqiArabic ʕeen
Kistane ʔin	Mesqan ʔen	Silt'e ʔiin	Standard Arabic ʕai_n
Tigre ʔɨn	Tigrigna ʕayni	United Arab Emira ʕeen	Wolane ʔin

r·s|head · binary · 14 languages · enrichment 104×

(Arabic) Ádirar rā́su	(Arabic) B□ ṛān rās	Amharic ras
Ancient Aramaic rēšā	BuxariArabic ṛaaṣ	DamasceneArabic ra:s
Egyptian Spoken A rˁaa_s	IraqiArabic ra:s	Ki-nubi ras
Maltese ras	Modern Hebrew ʁoʃ	MoroccoArabic rˁas_ˁ
Shoa Arabic ra:s	United Arab Emira raas	

l·s·n|tongue · ternary · 14 languages · enrichment 349×

(Arabic) B□ ṛān lísān	CypriotArabic l-s-n	DamasceneArabic lsa:n
Egyptian Spoken A lisaan	IraqiArabic lisa:n	Karekare lusùŋ
Ki-nubi ∗lisaan	Maltese lsi:n	Modern Hebrew laʃon
MoroccoArabic lsa_n	Shoa Arabic lisa:n	Standard Arabic lisa_:n
United Arab Emira lisaa_n	YemeniticArabic lisaa_n	

b·t·n|belly · ternary · 11 languages · enrichment 666×

(Arabic) Ádirar bátnu	BuxariArabic batin	DamasceneArabic batˁn
Egyptian Spoken A batˁ_n∗	IraqiArabic batˁi_n	Ki-nubi batna
Modern Hebrew bɛtɛn	Shoa Arabic bótana	Standard Arabic batˁ_n
United Arab Emira batˁn	YemeniticArabic batˁn	

r·k·b|knee · ternary · 11 languages · enrichment 308×

(Arabic) Ádirar ríkib	(Arabic) B□ ṛān ríkibẹ	DamasceneArabic rəkb_e
Egyptian Spoken A ruk_ba∗	IraqiArabic rukb_a	Ki-nubi ∗rukuba
Maltese rko_bba	MoroccoArabic rok_ba	Shoa Arabic rukuba
Standard Arabic rukba	United Arab Emira rikba	

m·j|mouth · binary · 11 languages · enrichment 71×

Boka miya	Dugwor mey	Kamwe-Futu myæ	Margi mìyà
MargiS miya	Mbara may	Mofu-Gudur mey	Moloko mey
Ngizim mìya	Ouldeme mey	West Margi myà	

l·n|tooth · binary · 11 languages · enrichment 72×

Cuvok zleŋ	Gaʔanda ɬêna	GizigaMarva slin	Gudu lın
Higi Ghye ɬɨnɛ	Hwona ɬànà	Kamwe-Futu ɬìnɛ	Lagwan sliɨni
Mbuko slan	Njanye line	Psikye sləna	

s·h·r|HAIR · ternary · 10 languages · enrichment 165×

DamasceneArabic ʃaʕra	Egyptian Spoken A ʃaʕɾ	IraqiArabic ʃaʕra	Maltese ʃaʕa_r	
Modern Hebrew ʃaʔaʁa	MoroccoArabic ʃʕer	Shoa Arabic ʃiʔir	Standard Arabic ʃaʕr	
United Arab Emira ʃaʕar	YemeniticArabic ʃaʕra			

k·m|ear · binary · 10 languages · enrichment 30×

Bole(wa) kúmo	Burma kɨm	Cip kɔ̀m	Dera kumo
Dira kumɛ̀	Dwot kum	Gera kumi	Ngamo kùmà
Polci kum	Sura kom		

Kinship and person

b·b|father with twenty-six languages and **m·m|mother** with thirteen are the pair that shows up in every family of the series, and here too: labial reduplication for the father, nasal for the mother. §6.1 of the Austronesian paper already warned that these are the least informative codes in a catalogue, because their shape is explained by child-rearing before history.

h·n·t|thou has the highest enrichment in the whole appendix, 1,063×, and must be read carefully because part of that figure is manufactured by our own table. The forms are *ʔant*, *ʔinta*, *ʔənte*, *hint*: the initial glottal stop is written h by the guttural merger §7 declares. A code beginning in h, in a family where h gathers four distinct consonants, is more frequent than its enrichment suggests. The figure is correct within our normalisation, and the normalisation is a decision.

n·g·s|king is the most cleanly branch-internal code in the appendix: ten languages, all Ethiosemitic — Amharic *nigus*, Ge'ez *niguś*, Harari *nʔgaaši*, Tigrinya, Tigre, Argobba, Silt'e, Kistane. It does not cross into Chadic or Berber, and should not: it is a word of political culture from a specific area. A catalogue without genealogy that produces a branch code is a catalogue measuring something.

b·b|FATHER · binary · 26 languages · enrichment 244×

Bacama bàba	Banana bàba	Dera bàba	Dira bàba
Dugwor baba	Gaʔanda bàba	Gavar bɑbɑ	Hwona bùba
Kamwe-Futu bàba	Karekare bə́ba	Matakam bəbu	Mbudum baba
Misme bouba	Muyang bɑbɑ	Ngizim bə́ba	Ouldeme bàbá

s·m|name · binary · 25 languages · enrichment 81×

Ancient Aramaic šmā	Angas sɨm	Banana šèma	Banana sema
DamasceneArabic ism	Egyptian Spoken A ism	IraqiArabic əsəm	Ki-nubi ⋆asma
Mazera sime	Misme sem	Modern Hebrew ʃɛm	Mser sɨm
Polci sɨm	Standard Arabic ism	Sura sum	Tarifit isəm

l·m|name · binary · 22 languages · enrichment 90×

Afade slɨm	Boka ɬɪmà	Bura ɬɪm	Cibak ɬɨmà
Dugwor zlam	Fali Kiria ɬɨmu	Gaʔanda ɬìma	Gavar zləm
Hwona ɬɨ̀mə̀	Kilba lém	Lagwan slɨmi	Margi ɬɪm
MargiS ɬɨmɔ̀	Mbara slim	Ngwaxi ɬɪm	Piðimdi lɪm

m·n|WHO · binary · 18 languages · enrichment 66×

Amharic man	Ancient Aramaic man	Argobba mani
BuxariArabic miin	CypriotArabic m-n	DamasceneArabic ma:n
Egyptian Spoken A mi:n	Ge'ez mənnu	IraqiArabic minu
Ki-nubi munu	Maltese min	Silt'e maani
Standard Arabic ma_n	Tigre mən	Tigrigna mən
United Arab Emira min		

m·m|MOTHER · binary · 13 languages · enrichment 106×

Bura mumà	Dugwor mama	Fali Kiria màma	Fali ƁwagiraÌ màma
Gavar mɑmɑ	Kamwe-Futu mama	Matakam mɔ̀m	Mbudum mama
Moloko mama	Ngwaxi mamà	Psikye màma	Shoa Arabic ómma
Vame māmá			

m·n|WHAT · binary · 12 languages · enrichment 45×

```
Amharic mɨn  Argobba mɨn  Bacama munò  Bata mɨno
Buduma məni  Harari min   Hwona mani   Kistane mɨn
Silt'e mɨn   Tarifit min  Wolane mɨn   Zway mɨn
```

n·g·s|KING · ternary · 10 languages · enrichment 240×

```
Amharic nɨgus  Argobba nɨgus  Ge'ez nɨguś  Harari nəgaaši
Kistane nɨgus  Silt'e nɨgus   Tigre nəgus  Tigrigna nɨgus
Wolane nɨgus   Zway nɨgus
```

n·d|PERSON · binary · 10 languages · enrichment 78×

```
Cibak ńdà          Cuvok nda    Fali Gili ǹdu  Fali Mucella ǹda
Kilba ǹdú          Lamang ùndù  MargiS ndū     Ngwaxi ndà
Shoa Arabic na:di  Sukur nda
```

h·n·t|THOU · ternary · 10 languages · enrichment 1063×

```
Amharic ʔantə          BuxariArabic hint        CypriotArabic ʔ–n–t
DamasceneArabic ʔənte  Egyptian Spoken A ʔinta  Ge'ez ʔantə
IraqiArabic ʔinta      Standard Arabic ʔan_ta   Tigre ʔɨnta
YemeniticArabic ʔant
```

Nature and sky

The block where the same concept comes out most often under different codes, and that is informative.

'Sun' appears three times: `p·s` (18 languages), `s·m·s` (15) and `f·t` (13). `s·m·s` is *šams*, Semitic beyond dispute. The other two are Chadic. Three codes for one concept in a four-branch family is not noise: it is the division.

'Water' comes out twice, `j·m` in twenty languages and `h·m` in eleven, and 'night' twice, `l·l` — *layl* — and `v·d`.

`n·f·s|wind` (358×) is *nafas* 'breath', and its presence in the nature block rather than the body block is a reminder that the thematic grouping of this appendix is a reading order and not a classification: the concept is fixed by Concepticon, not by us.

t·r|moon · binary · 22 languages · enrichment 88×

```
Angas tar       Ankwe tàr      Bole(wa) téręˋ  Daba tìrà
Dera tere       Fali Gili tɪf̃  Gera tèra      Higi Ghye tɨra
Karekare tárei  Lamang tré     Mandara tɨre   Miya tɪr
Ngamo tɛ̀rè      Ngizim téra    Psikye trə     Sura tar
```

j·m|water · binary · 20 languages · enrichment 143×

```
Buwal yɑm        Cibak yìmi       Cuvok yam        Daba yim
Fali Gili yam    Fali Jilbu yɪmû  Fali Kiria yàmu  Ga?anda yèma
GizigaMarva yam  Kamwe–Futu yɛ̀mi  Mbudum jim       Merey yàm
Moloko yàm       Ngwaxi yìmì      Ouldeme yàm      Psikye yemú
```

p·s|sun · binary · 18 languages · enrichment 127×

```
Angas pus    Ankwe pɨs     Buwal pes   Cip pùs
Cuvok pas    Dugwor pats   Gavar pish  Gera pìši
Kilba pɨ́čì   Matakam pats  Mbazla pǎs  Mbudum pis
Moloko pats  Ouldeme pats  Sukur pis   Sura puus
```

s·m·s|sun · ternary · 15 languages · enrichment 268×

```
(Arabic) B□ ṛān sḗms   Ancient Aramaic šemšā    BuxariArabic ʃams
DamasceneArabic ʃams  Egyptian Spoken A ʃam_s  IraqiArabic ʃamis
Ki-nubi ∗semsi        Maltese ʃem_ʃ            Modern Hebrew ʃɛmɛʃ
MoroccoArabic ʃe_mʃ   Shoa Arabic ṣ́āmiṣ        Standard Arabic ʃa_ms
Tigre šɨmɨš           United Arab Emira ʃams   YemeniticArabic ʃams
```

f·t|sun · binary · 13 languages · enrichment 96×

```
Banana fata        Bata fótó      Bole(wa) fùtî  Fali Jilbu fitî
Fali Mucella fìtu  Karekare fətî  Lamang fítí    Lame fəta
Mandara fúti       Misme fata     Ngamo fɔtî     Njanye fɨ̀te
Ouldeme fāt
```

n·r|fire · binary · 12 languages · enrichment 177×

```
(Arabic) Ádirar nā́ru  (Arabic) B□ ṛān nār     Ancient Aramaic nūrā
BuxariArabic naar     DamasceneArabic na:rˁ  Egyptian Spoken A na:ɾ
IraqiArabic na:r      Ki-nubi nari           Maltese nar
Shoa Arabic na:r      Standard Arabic na:_r  YemeniticArabic naar
```

l·l|night · binary · 11 languages · enrichment 75×

```
(Arabic) Ádirar láilu  (Arabic) B□ ṛān leil     BuxariArabic layl
CypriotArabic l-y-l    DamasceneArabic leele   Egyptian Spoken A le:l
IraqiArabic leela      MoroccoArabic lila      Shoa Arabic le:le
Standard Arabic laila  United Arab Emira leel
```

v·d|night · binary · 11 languages · enrichment 145×

```
Bata vùɗó        Cuvok vaɗ          Daba vùɗu        Dghwede vuɗì
Fali Jilbu vɨɗi  Fali Mucella vùɗu  Kamwe-Futu viɗi  Matakam vaɗ
Njanye vɨ̀ɗe      Psikye víɗi        Sukur vəɗ
```

h·m|water · binary · 11 languages · enrichment 58×

```
Gera həmə    Karekare ʔèmù  Margi ʔimi  MargiS ʔɪmì
Mazera aˈɨm  Mbara ’am      Mbuko a’am  Ngamo hâm
Sura ʔâm     Tangale hâm    Vame āhwám
```

b·r·d|cold · ternary · 10 languages · enrichment 237×

```
(Arabic) B□ ṛān bā́rid     BuxariArabic bard        DamasceneArabic barəd
Egyptian Spoken A bar_d  IraqiArabic beerə_d      Maltese bi:r_ed
MoroccoArabic bar_ed     Standard Arabic ba:_rid  United Arab Emira baari_d
YemeniticArabic baari_d
```

h·s·r|stone · ternary · 10 languages · enrichment 119×

```
(Arabic) Ádirar háṣar    (Arabic) B□ ṛān háṣar    CypriotArabic ħ–_ʤ–r
DamasceneArabic ħaʤa_ra  IraqiArabic ħaʤa_r      MoroccoArabic ħeʒ_ra
Shoa Arabic hiʤer        Standard Arabic ħad͡ʒar  United Arab Emira ħaʤa_r
YemeniticArabic ħaʤa_r
```

n·f·s|wind · ternary · 10 languages · enrichment 358×

```
Amharic nɨfas  Argobba nɨfas  Ge'ez nəfas  Kistane nəfas
Mesqan nɨfas   Silt'e nɨfas   Tigre nɨfas  Tigrigna nɨfas
Wolane nɨfas   Zway nɨfas
```

Animals

k·l·b|dog (*kalb*, 16 languages, 258×) and **k·r|dog** (15 languages, 36×) are the appendix's best pair for understanding enrichment. Same concept, almost identical reach, and a factor of seven between them: the ternary is rare by chance and the binary is not. Reading reach without enrichment would have rated the two equally good.

p·l·s|horse (344×) crosses branches, and is among the codes §6.2 flags as candidates: the horse reaches the continent by a known route and its name travels with it. The catalogue cannot say it is a loan — it does not distinguish inheritance from contact — but it can point at where to look.

k·l·f|fish · ternary · 20 languages · enrichment 474×

```
Bura kɨlfà    Cuvok kəlef        Daba kìlif     Dghwede klfe
Gava kilifà   GizigaMarva kilef  Gələvda kɪlfà  Kilba kálfì
MargiS kalfi  Matakam klef       Mazera kɨlfa   Mbudum kəlːif
Merey kəlef   Moloko kəlef       Muyang kilif   Ngwaxi kulfu
```

p·l·s|HORSE · ternary · 16 languages · enrichment 344×

```
Daba pìliš         Dghwede pɨlišì  Dugwor pəles   Gava pɨ̀lisà
GizigaMarva piles  Gələvda pɨlšà   Lamang plís    Matakam pliš
Mbara pilis        Mbuko pəles     Merey pəles    Moloko pəles
Muyang pilis       Nakatsa pə̀ləse  Ouldeme pəlìs  Vame pālésh
```

k·l·b|dog · ternary · 16 languages · enrichment 258×

```
(Arabic) Ádirar kálbu  (Arabic) B□ ṛān kḗlb   Ancient Aramaic kalbā  BuxariArabic kalb
CypriotArabic k–l–b    DamasceneArabic kalb  Egyptian Spoken A klb  Ge'ez kəlb
IraqiArabic kalb       Maltese kelb          MoroccoArabic kel_b    Shoa Arabic kalb
Standard Arabic kal_b  Tigre kəlɨb           Tigrigna kəlbi         YemeniticArabic kalb
```

k·r|dog · binary · 15 languages · enrichment 36×

```
Dugwor kəra   Fali Gili kɨri  Fali Kiria kuřì  GizigaMarva kre
Hausa kare    Higi Ghye kuře  Kamwe–Futu kɨřɛ  Lamang kəré
Mandara kɨ̀rè  Merey kərɑ      Moloko kəra      Muyang kərɑ
Ouldeme kərà  Psikye kuřɛ     Sukur kəra
```

t·m|SHEEP · binary · 12 languages · enrichment 107×

```
Ankwe tɨm  Bura tɨmà  Cibak tɨmà       Cip tìm
Dira tɨmè  Dwot tum   Kamwe-Futu tɨmɛ̀  Ngwaxi tɪmà
Polci tɨm  Seya tîm   Sura tum         West Margi tɨmà
```

t·r|bird · binary · 11 languages · enrichment 44×

```
(Arabic) Ǻdirar tǻer   (Arabic) B□ ṛān tǻer        BuxariArabic tayra
DamasceneArabic tˁeer  Egyptian Spoken A tˁee_r∗  IraqiArabic tˁeer
Ki-nubi ter            MoroccoArabic tˁir         Shoa Arabic tˁe:ra
Standard Arabic tˁɑir  United Arab Emira tˁeer
```

Plants and food

l·h·m|meat (*laḥm*, ternary, 289×) against **l·w|meat** (binary, 20 languages, 149×): again the same concept split between the Semitic ternary and the Chadic binary, and again the ternary pays more enrichment for less reach.

m·m|honey with fifteen languages is a reduplication case like those in the kinship block, and calls for the same caution.

l·w|meat · binary · 20 languages · enrichment 149×

```
Ankwe lowa         Boka ɬuwə     Cuvok slaw   Dghwede ɬuwè
Fali Mucella lùwa  Ga?anda lîwa  Gava ɬuwà    Gələvda ɬʊwà
Lamang slùw        Matakam ɬuwe  Merey zlaw   Miya ɬiwi
Nakatsa ɬuwa       Ngizim ɬùwəi  Njanye lìwo  Sura luwa
```

m·m|HONEY · binary · 15 languages · enrichment 129×

```
Cibak mɨmà    Dugwor amam  Gava mamà        GizigaMarva amam
Gələvda màmà  Malgbe mam   Margi mùmù       MargiS mumɔ
Mbuko umam    Moloko omom  Mpade màm        Mser mam
Nakatsa màma  Sukur mam    West Margi mɨmɛ̀
```

s·r·b|drink · ternary · 11 languages · enrichment 198×

```
BuxariArabic ʃarab       CypriotArabic ʃ-r-b    DamasceneArabic iʃrib
Egyptian Spoken A ʃɾb    IraqiArabic ʃirab      Maltese ʃoro_b
MoroccoArabic ʃre_b      Shoa Arabic ʃarab      Standard Arabic ʃari_ba
United Arab Emira ʃarab  YemeniticArabic ʃirib
```

h·k·l|eat · ternary · 10 languages · enrichment 365×

```
Ancient Aramaic ʔkal   BuxariArabic ʔukul       CypriotArabic ʔ-k-l
DamasceneArabic ʔakal  Egyptian Spoken A ʔkl    IraqiArabic ʔakal
Ki-nubi ‘akulu         Standard Arabic ʔak_ala  United Arab Emira ʔakal
YemeniticArabic ʔakal
```

l·h·m|meat · ternary · 10 languages · enrichment 289×

```
BuxariArabic læħəm        CypriotArabic l-ħ-_m   DamasceneArabic laħm
Egyptian Spoken A laħ_ma  IraqiArabic laħa_m     Maltese laħa_m
MoroccoArabic lħa_m       Standard Arabic laħ_m  United Arab Emira laħa_m
YemeniticArabic laħm
```

Verbs of action and state

m·t|die is the widest-reaching code in the whole catalogue: thirty-eight languages, twenty-eight Chadic and twenty-two Semitic. It is the only one in the appendix crossing the two large branches with that force. Afro-Asiatic comparison has been discussing that *mwt* for a century; the catalogue finds it having consulted nothing.

The rest of the block are triliterals: `s·m·h` *samiʕa* 'hear', `h·k·l` *ʔakala* 'eat', `h·r·f` *ʕarafa* 'know', `s·r·b` *šariba* 'drink'. All four have between ten and twelve languages and enrichments from 198× to 365×.

m·s|walk (87×) is out of place, for the same reason as `l·m` in the reverse index: binary, ten languages of reach, low enrichment. It is in the appendix by reach, not by quality.

m·t|die · binary · 38 languages · enrichment 164×

```
(Arabic) Ádirar mét  Angas mùt      Buduma matə             Egyptian Spoken A maa_t
Fali Ɓwagira mti     Higi Ghye mtɛ  Ki-nubi mutu            Lamang mta
Margi mtu            Misme mat      Modern Hebrew mɛt       Mser mato
Ngizim mùtu          Ouldeme mət    Standard Arabic ma:_ta  United Arab Emira maat
```

s·m·h|hear · ternary · 12 languages · enrichment 243×

```
BuxariArabic samaʕ        CypriotArabic s-m-_ʕ      DamasceneArabic ismiʕ
Egyptian Spoken A smʕ     IraqiArabic sima_ʕ        Maltese sam_a'
Modern Hebrew ʃamaʔ       MoroccoArabic sm_eʕ       Shoa Arabic simi'
Standard Arabic sa_miʕ_a  United Arab Emira sima_ʕ  YemeniticArabic simiʕ_∗
```

m·s|WALK · binary · 10 languages · enrichment 87×

```
CypriotArabic miʃi      DamasceneArabic maʃa    Egyptian Spoken A miʃi  IraqiArabic miʃa
Ki-nubi Mashi           Maltese meʃ_a           MoroccoArabic mʃ_a      Shoa Arabic maʃa
Standard Arabic ma_ʃa:  United Arab Emira miʃa
```

h·r·f|know · ternary · 10 languages · enrichment 341×

```
BuxariArabic ʔaraf       CypriotArabic ʕ-r-f  DamasceneArabic ʕarˁaf
Egyptian Spoken A ʕirif  IraqiArabic ʕiraf    Ki-nubi ‘arufu
MoroccoArabic ʕrˁe_f     Shoa Arabic ʕirif    Standard Arabic ʕar_afa
United Arab Emira ʕaraf
```

Objects and culture

The smallest block, and predictably so: objects of material culture are where borrowing weighs most and where concept lists agree least between sources.

l·b·r|needle reaches 730× with ten languages, the appendix's second-highest enrichment. A ternary for a concrete object with that concentration is exactly the profile of an isogloss candidate, and §6.3 would pick it up if its pair were in the catalogue.

l·b·r|NEEDLE (FOR SEWING) · ternary · 10 languages · enrichment 730×

```
Afade libɨra  Buduma líbūra   D□ ai lí́bīra     Karekare alí́bura
Lagwan libra  Malgbe libra    Maltese 'lebrɐ  Mpade líbrà
Mser libra    Ngizim alí́bera
```

Qualities

j·b·s|dry at 938× is the highest enrichment after **h·n·t**, and unlike that one it contains no merged guttural: it is *yābis*, with `j`, `b` and `s` cleanly separate in the normalisation table. Of the whole appendix, it is the code whose enrichment deserves most belief.

`k·b·r|big` (*kabīr*, 266×) and `t·w·l|long` (*ṭawīl*, 259×) complete a block that is entirely ternary and entirely Semitic. That the qualities all come from one branch while the numerals split in two says something about which fields this family shares and which it does not — but saying exactly what would require the genealogy this work does not use.

j·b·s|dry · ternary · 13 languages · enrichment 938×

```
(Arabic) B□ ṛān yábes     Ancient Aramaic yabbīš  BuxariArabic yabs
DamasceneArabic ji:bis   Ge'ez yɨbus             IraqiArabic jeebə_s
Kándin yābű́ṣi           MoroccoArabic jab_es    Shoa Arabic ja:bis
Standard Arabic ja:b_is  Tigre yɨbus             United Arab Emira jaabi_s
YemeniticArabic jaabi_s
```

k·b·r|big · ternary · 12 languages · enrichment 266×

```
(Arabic) Ádirar kab□ ru  (Arabic) B□ ṛān ikbér     BuxariArabic ka’biir
CypriotArabic k-b-_r     DamasceneArabic kbi:r    Egyptian Spoken A kibiːɾ
Ki-nubi kbir             Maltese kibi_r           MoroccoArabic kbi_r
Shoa Arabic kabi:r       Standard Arabic kab_i:r  YemeniticArabic Kabii_r
```

t·w·l|long · ternary · 11 languages · enrichment 259×

```
CypriotArabic tˁ-w-_l       DamasceneArabic tˁawi:_l  Egyptian Spoken A tˁaw_i:l
IraqiArabic tˁ-w-_l         Ki-nubi towil             Maltese twil
MoroccoArabic tˁwi_l        Shoa Arabic tˁawil        Standard Arabic tˁa_wi:l
United Arab Emira tˁawii_l  YemeniticArabic tˁawi:_l
```

m·l|full · binary · 10 languages · enrichment 129×

```
Amharic mulu  Ancient Aramaic mlē  Argobba mulu        IraqiArabic əmmal_a
Kistane mula  Mesqan mula          Modern Hebrew malɛ  Silt'e mulla
Wolane mulli  Zway muli
```

Other

What did not fit the previous blocks, left in so the appendix does not give the impression of being tidier than it is. The thematic order is for reading; the codes did not come out ordered.

m·b·j|CASSAVA · ternary · 27 languages · enrichment 581×

```
Boka mbəy         Cuvok mbay          Dghwede mbàyà  Fali Gili mbàyi
Fali Jilbu mbayà  Fali Mucella mbàya  Ga?anda mbə̀y   Gava mbàyà
Gudu m̀bə́y        Higi Ghye mbàyà     Hwona mbə̀y     Mandara mbə̀yà
MargiS mbàyu      Merey mbay          Muyang mbɑy    Ouldeme mbàyà
```

v·j|RAINY SEASON · binary · 20 languages · enrichment 309×

Bura viyà	Buwal vəyɑ	Cibak vɪyà	Cuvok viya
Fali Kiria vìya	Fali Mucella vìyà	Gava vìyà	Gavar viyɑ
Kamwe-Futu viya	Kilba vúyà	Mandara vìyà	Margi viya
Matakam viya	Mbudum vija	Mbuko viya	Piðimdi vìya

m·t|DEATH · binary · 19 languages · enrichment 109×

Angas mut	Ankwe mut	Banana mita	Bole(wa) mòto
Bura ḿtà	Cibak ḿtì	Fali Gili m̀ti	Kamwe-Futu mto
Lagwan mɨti	Lame mat	Margi ḿtù	Ngamo màtò
Ngizim mùtu	Psikye mtə	Standard Arabic maut	Tangale mutò

k·w·r|DONKEY · ternary · 18 languages · enrichment 341×

Bata kwà:ríò	Boka kwarè	Bura kwəra	Cibak kwara
Fali Kiria kwarā	Fali ƁBwagira kwəra	Ga?anda kwari	Gudu kwára
Hwona kwara	Kamwe-Futu kwara	Kilba kwárá	Margi kwara
MargiS kwara	Ngizim kwara	Ngwaxi kwəra	Psikye kwarà

m·r|OIL (ORGANIC SUBSTANCE) · binary · 17 languages · enrichment 87×

Bata maare	Bole(wa) moř̀ì	Burma mar	Cip mòr
Dira miri	Fali Mucella mara	Gera mori	Gudu mář
Karekare mərù	Lame mar	Moloko amar	Ngamo mɔř
Njanye mare	Piðimdi məř	Polci mɪr	Seya mɪr

m·l|OIL (ORGANIC SUBSTANCE) · binary · 16 languages · enrichment 99×

Banana mula	Bura mə̂l	Buwal mel	Cibak màl
Cuvok mal	Fali Jilbu malî	Gavar mɑli	Geji mili
GizigaMarva mal	Kilba màl	Margi mâl	MargiS mal
Mbara mal	Ngwaxi məl	Ouldeme āmāl	West Margi màl

s·r|TRUTH · binary · 16 languages · enrichment 50×

Bade jiř̂ai	Bura jɪrì	Cibak jɪrì	Dera jirè
Gava jirà	Gələvda jirà	Kamwe-Futu jɪrì	Karekare jirè
Kilba ǰírí	Mandara jɪrè	Margi jìrì	MargiS jɪri
Nakatsa jira	Ngamo jìrê	Piðimdi jìrè	West Margi jɪrì

l·w|FLESH · binary · 16 languages · enrichment 155×

Ankwe lowa	Boka ɬuwà	Cuvok slaw	Dghwede sluwe
Dugwor slaw	Fali Gili ɬìwu	Fali Mucella lùwa	Ga?anda ɬìwa
Gava ɬuwà	Gera lèwi	Gələvda ɬʊwà	Hwona ɬuwa
Miya ɬiwì	Nakatsa ɬuwa	Njanye lìwe	Sura luwa

p·t|CAT · binary · 15 languages · enrichment 104×

Cibak pàtu	Fali Gili patù	Kamwe-Futu pàtu	Mandara pàtu
Margi pàtu	MargiS patu	Matakam patu	Muyang pɑtu
Ngamo pàtu	Ngizim pā́tu	Ngwaxi pàtu	Piðimdi pàtu
Sukur patu	Vame pātú	West Margi pàtu	

k·l|ALL · binary · 14 languages · enrichment 45×

Ancient Aramaic kull	BuxariArabic kullu	CypriotArabic kull	DamasceneArabic kəll
Egyptian Spoken A kull	Harari kullu	IraqiArabic kull	Ki-nubi kulu
Maltese Koll_u	MoroccoArabic koll	Standard Arabic kull	Tigrigna kullu
United Arab Emira kil	YemeniticArabic kull		

s·n|DREAM (SOMETHING) · binary · 13 languages · enrichment 41×

Angas sun	Bura sunì	Cibak šinì	Daba šìni
Fali Mucella sinî	Gera sunə	Karekare sunà	Mandara šinè
MargiS šuni	Ngamo sùnâ	Ngwaxi šinì	Piðimdi jina
West Margi sɨnì			

s·t|PEPPER · binary · 13 languages · enrichment 87×

Angas šɨta	Ankwe šità	Bole(wa) dṣ́ita	Cip šìta
Daba šitə	D ai dṣ́ita	Fali Kiria šatà	Karekare dṣ́ita
Kilba čìtá	Ngizim ṣ́ita	Shoa Arabic ṣ́ẹ̀ta	Sura žìta
Tangale sità			

g·m|RAM · binary · 13 languages · enrichment 93×

Ankwe gàm	Bade g ma	Bole(wa) gèmi	Bura gâm
Dera gəm	D ai gam	Kamwe-Futu game	Karekare gəmu
Kilba gâm	MargiS gāmo	Ngamo gèm	Ngwaxi gəm
West Margi agəmɛ̀			

k·w·l|PENIS · ternary · 13 languages · enrichment 228×

Bata kwáàló	Bura kwâl	Cibak kwalæ	Fali Gili kwala
Fali Kiria kwəla	Higi Ghye kwala	Kamwe-Futu kwala	Kilba kwăl
Margi kwal	MargiS kwal	Ngwaxi kwə̀l	Psikye kwala
West Margi kwal			

s·t·n|DEMON · ternary · 12 languages · enrichment 773×

(Arabic) Ǻdirar ṣéitān	(Arabic) B ṛān séitān	Afade setan	Karekare ṣ́itan
Malgbe shetan	Maltese ʃɪ'tɐːn	Mandara sátẹn	Merey setene
Mpade sìtán	Ngizim ṣ́itan	Polci sheetan	Shoa Arabic séitān

l·m|MARKET · binary · 12 languages · enrichment 75×

Bata lú:mù	Boka lumu	Daba lumû	Dghwede lumà
Fali Jilbu lumà	Fali Kiria lumà	Fali Mucella lumà	Ga?anda lìmo
Gudu lú:mú	Lame lumo	Njanye lumu	Tangale lomò

z·v·n|GUINEA FOWL · ternary · 12 languages · enrichment 625×

Buwal zɑvɑn	Daba zàvun	Fali Gili zùvuni	Fali Jilbu zàvɨ̀nî
Fali Kiria zɨvɨn	Fali Mucella zàvunu	Gavar zɑvən	Gudu zúvʌn
Kamwe-Futu zivəno	Lagwan zavan	Mbudum zavaŋ	Psikye zɨ̀vùnɛ

k·d·k|SWEET POTATO · ternary · 12 languages · enrichment 514×

```
Ankwe kudùkû    Buwal kwɑdɑkwɑ       Dira kudùku    Fali Gili kudàku
Gavar kwɑdɑkwɑ  GizigaMarva kudaaku  Kilba kúdàkú   MargiS kɨdakù
Mbuko kwadakwa  Merey kodoko         Ngwaxi kudàku  Njanye kudàku
```

r·z|RICE · binary · 12 languages · enrichment 1034×

```
Amharic ruz  Ancient Aramaic rōzā  Argobba ruz  Ge'ez ruz
Harari ruuz  Kistane ruz           Mesqan ruz   Silt'e ruz
Tigre ruz    Tigrigna ruz          Wolane ruz   Zway ruz
```

k·m|FACE · binary · 11 languages · enrichment 42×

```
Bura kɨmà          Cibak kumà      Fali Gili kuma   Fali Kiria kuma
Fali Mucella kuma  Higi Ghye kɨma  Kamwe-Futu kɨma  Ngwaxi kɨmà
Psikye kuma        Sukur kəma      West Margi kɨmà
```

b·c|MAT · binary · 11 languages · enrichment 450×

```
Buduma boci   GizigaMarva buce  Kamwe-Futu bùci  Karekare bucì
Mandara bùce  Margi buci        MargiS buci      Mbuko buce
Muyang bʉci   Ngizim bucì       West Margi buci
```

ɓ·l·m|YAM · ternary · 11 languages · enrichment 407×

```
Bura ɓulùm     Cibak ɓɨlma     Dugwor ɓəlam      Margi àɓılım
MargiS ɓulùm   Mbuko ɓəlam     Merey ɓəlem       Muyang ɑɓəlum
Ngwaxi ɓılıma  Ouldeme āɓālām  West Margi ɓɨlma
```

g·m·l|CAMEL · ternary · 11 languages · enrichment 998×

```
Amharic gɨməl   Ancient Aramaic gamlā  Argobba gamela   Ge'ez gəməl
Harari gaaməla  Kistane gimə̄l         Mesqan gamela    Silt'e gaameela
Tigre gemə̄l     Tigrigna gɨməl         Wolane gaameela
```

l·g·m|CAMEL · ternary · 11 languages · enrichment 862×

```
Buduma loguəme    Dghwede zlugwama   Dugwor zlegweme  Malgbe logome
MargiS dlugwam    Mbuko zlugweme     Merey zləgweme   Moloko ezlegweme
Muyang ezligwemi  Ouldeme ázləgwèmè  Vame ázləgwèmè
```

k·r|DONKEY · binary · 11 languages · enrichment 35×

```
Bade koro   Bole(wa) koro  Kamwe-Futu kwara  Karekare koro
Lame koro   MargiS kwara   Mbara kuro        Mpade kóró
Ngamo kɔrô  Piðimdi koro   Psikye kwará
```

b·t|CHEAP · binary · 11 languages · enrichment 157×

```
Bata bú:tù         Bura butù        Cibak butù       Fali Kiria butù
Fali Mucella butà  Kamwe-Futu butì  Mandara butù     Margi bùtù
MargiS butu        Ngwaxi bùtu      West Margi butu
```

x·l|STEAL · binary · 10 languages · enrichment 129×

```
Bura xula        Cibak xulæ       Daba àxɨl    Fali Gili ɣèli
Fali Kiria ɣùlù  Kamwe-Futu ɣùlì  MargiS xulà  Ngwaxi xɪlà
Psikye ɣùlì      West Margi xɨlɛ
```

b·n|COFFEE · binary · 10 languages · enrichment 434×

```
Amharic bunna  Argobba bunn  Ge'ez bunna  Harari bun
Kistane buna   Silt'e buno   Tigre bun    Tigrigna bun
Wolane buno    Zway buna
```

k·k|GRANDPARENTS · binary · 10 languages · enrichment 180×

```
Boka kàka     Bole(wa) kàka    Bura kɨka    Dira kàka
Ga?anda kàka  Hausa kȁkā       Ngwaxi kàka  Piðimdi kàka
Polci kaka    West Margi kàka
```

m·n|HOW MANY · binary · 10 languages · enrichment 52×

```
Bole(wa) munò      Burma mənà  Dira mùnè      Dwot mənî
Fali Mucella mànà  Gera munu   Karekare manà  Miya mɨnà
Ngamo mɔ̀nɔ̂        Polci mèni
```

d·f|MUSH · binary · 10 languages · enrichment 227×

```
Fali Gili ɗàfa  Fali Kiria ɗəfa  Fali Mucella ɗəfa  Gudu ɗà:f
Higi Ghye ɗàfa  Kamwe-Futu ɗàfa  Mandara ɗəfa       Margi ɗufù
Psikye ɗàfa     West Margi ɗɨfɛ
```