

Códigos consonánticos en 27 lenguas urálicas

La familia donde el instrumento por fin puede hacer lo que promete

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

Resumen

Presentamos un catálogo de 1.486 códigos consonánticos que agrupan 13.470 formas léxicas de 27 lenguas urálicas. Un código es un esqueleto consonántico compartido por palabras del mismo concepto en lenguas distintas. No se usa reconstrucción, ni cognación experta, ni ninguna hipótesis sobre el proto-urálico.

Es el tercer paper de la serie y el primero en el que la familia tiene la estructura que el método necesita: nueve ramas repartidas, de las cuales seis con datos suficientes, frente a las dos del turco y las seis muy desiguales del austronesio. Eso permite hacer por primera vez lo que el protocolo prometía y no había podido probar: calcular las sustituciones por par de ramas sobre treinta y seis pares en vez de uno.

Cuatro resultados. Los códigos de mayor alcance — $n \cdot m$ | name con 20 lenguas y $378 \times$, $v \cdot r$ | blood, $k \cdot l$ | fish, $m \cdot n$ | I, $t \cdot n$ | THOU— son los que cualquier uralista reconocerá de inmediato, obtenidos sin consultar ninguna reconstrucción. La búsqueda de códigos hermanos —pares del mismo concepto que difieren en una posición— reparte lenguas en grupos sin usar etiqueta genealógica alguna, y su filtro de disyunción resulta imprescindible: un cuarto de los pares comparte lenguas y por tanto no reparte nada. La geografía da aquí un gradiente monótono en los seis cubos, que el turco no pudo medir por falta de rango. Y el despiece automático de afijos vuelve a empeorar el resultado —sube diez puntos de agrupamiento y dieciocho de nulo— confirmando en una segunda familia aglutinante lo que el paper turco encontró.

Con tres familias medidas por el mismo código, la razón sobre el azar ordena por profundidad temporal —turco $216 \times$, urálico $78 \times$, austronesio $37 \times$ — pero la precisión observada no. Es un dato, no una tesis. (*Con las ocho familias que la serie tiene ya, la ordenación no se sostiene: ver la nota de §5.6.*)

1. Qué se busca, y por qué el urálico

La lingüística histórica dispone de un instrumento poderoso y caro: el método comparativo. Se ha aplicado con una desigualdad enorme —el indoeuropeo tiene reconstrucción densa y cognación experta, la mayoría de las familias no— de modo que cualquier estudio que se apoye en la capa histórica del léxico medirá, sin quererlo, la densidad de la bibliografía comparatista y no el pasado de las lenguas.

La serie a la que pertenece este trabajo pregunta si se puede hacer algo útil renunciando a la reconstrucción: no mejorarla ni sustituirla, prescindir de ella y ver qué queda. El paper austronesio mostró que queda un catálogo de códigos consonánticos; el turco preguntó qué significan sus cifras.

Este paper pregunta hasta dónde llega el instrumento cuando nada se lo impide. El paper turco declaró tres límites que resultaron ser de la familia y no del método: su corpus solo distingue dos ramas —con una sola lengua en una de ellas—, así que la regla de calcular las sustituciones por par de ramas se probó sobre un único par; el análisis geográfico salió degenerado, con 467 de 496 pares en un solo cubo; y la búsqueda de códigos hermanos llegó al final, sin nada con que contrastarla.

El urálico levanta esas tres restricciones. Nueve ramas repartidas, dispersión geográfica de miles de kilómetros sin ser un archipiélago, y una división primaria establecida.

Y aquí hay que ser muy preciso, porque se malinterpreta con facilidad. Esas nueve ramas no son una entrada del método: el catálogo se construye igual que en las otras dos familias, sin saber que existen. Lo que las ramas permiten es comprobar lo que el catálogo produjo — y por eso el urálico es una buena familia para este paper: no porque el instrumento las necesite, sino porque nosotros necesitamos algo contra lo que contrastarlo.

1.1 El método, y dónde está descrito

El procedimiento —qué es un código, en qué orden se construye, qué nulos usa y qué se puede afirmar al batirlos— está descrito una vez para toda la serie en «El catálogo de códigos consonánticos», y no se repite aquí. Lo esencial, para poder leer lo que sigue:

El catálogo se construye sin genealogía. Los códigos salen del esqueleto consonántico de una forma, del concepto que expresa y de un umbral de tres lenguas distintas. Ninguna clasificación en ramas, ninguna protoforma, ninguna lista de cognados. El procedimiento no sabe que esta familia tiene ramas, y la genealogía entra al final solo para leer un catálogo ya cerrado.

No se reconstruye, no se afirma parentesco, no se distingue herencia de préstamo, y no se atribuye significado a un código.

El encaje de este procedimiento en la literatura de comparación automática —Dolgopolsky, Turchin/Peiros/Gell-Mann, ASJP— está en el paper del método.

2. Qué es un código, y qué cuesta en urálico

2.1 El esqueleto consonántico

El esqueleto es la secuencia de consonantes de una palabra, con las vocales retiradas —no perdidas: quedan en columna aparte— y cada consonante escrita con un representante canónico. El urálico usa 25 símbolos canónicos, de los cuales 23 funden algo. La tabla completa está en `runs/normalizacion.md` y no está escrita a mano: se deriva alineando cada forma con su esqueleto.

La tabla completa está en `runs/normalizacion.md`. Éste es el inventario canónico entero, que hace falta para leer §5.1 y §5.2 — sin él no se sabe si *u*, *c* o *z* son símbolos del esqueleto o erratas:

s r n t l k m j v p h d z g w x q c b f u ð θ β g

Y lo que conviene tener presente, porque decide qué códigos coinciden:

Se escriben	como	Consecuencia en los datos
ʃ tʃ ʒ ɕ tɕ ts dʒ ... (sibilantes y africadas)	s	‘s·l·m
eye‘ reúne <i>silmä</i> , <i>šalmä</i> , <i>šel‘me</i>		
ŋ n̥ n̩	n	‘n·m
name‘ reúne <i>nimi</i> , <i>nim</i> , <i>nam</i>		
ʔ h h̥	h	
ʏ ɤ	x	
z NO se funde con s	z	y por eso §5.2 puede ver ‘k·s
FIR frente a k·z‘		
k , q y x se mantienen separados		y por eso puede ver k·l frente a q·l y x·l

Se escriben	como	Consecuencia en los datos
v y w se mantienen separados	y por eso puede ver	v · r frente a w · r

Que **k**, **q**, **x** y **z** no se fundan es la decisión más productiva de la tabla en esta familia, y conviene declararlo antes de que los resultados la aprovechen: es lo que permite que §5.2 vea las alternancias velares y sibilantes como códigos distintos. En turco, la decisión contraria sobre las sibilantes destruyó una correspondencia diagnóstica.

Y una precisión que evita malentendidos al llegar a §5. Todo el paper —catálogo, sustituciones, códigos hermanos— opera sobre este mismo alfabeto de 25 símbolos, el de después de la normalización. Cuando §5.1 habla de un par **z** **s**, son los símbolos canónicos **z** y **s**, no dos sonidos anteriores a la tabla.

2.2 Un código es un esqueleto que se repite entre lenguas

Un código es un esqueleto que aparece, para el mismo concepto, en tres lenguas distintas o más. Se escribe **n·m**|name, nunca **n·m** = name: el código no significa «nombre», es la forma que «nombre» toma repetidamente en la familia.

2.3 El coste del instrumento: la colisión

El 59 % de las formas urálicas comparten esqueleto, dentro de su propia lengua, con una forma de otro concepto. Es la medida limpia de lo que el esqueleto no distingue, sin genealogía de por medio.

El caso que mejor lo enseña es **k·l**, que aparece con «pez» y con «lengua». Y basta el finés para verlo: dice *kala* y *kieli*. Son dos palabras que ninguna descripción confundiría, y el esqueleto las escribe igual porque la diferencia está entera en las vocales, que es lo que descarta. Lo mismo **t·l** con «fuego» (*tuli*) e «invierno» (*talvi*), y **n·l** con «cuatro» (*neljä*) y «flecha» (*nuoli*).

Cuatro pares de palabras distintas bajo un mismo código entre las que el anexo enseña. No es un defecto que se arregle con más datos: es la aritmética de reducir una palabra a dos o tres consonantes, y en urálico pesa a pesar de que sus palabras son las más largas de la serie.

2.4 Aridad

Aridad	Códigos	
1 (unario)	166	11,3 %
2 (binario)	652	44,2 %
3 (ternario)	523	35,5 %
4 o más	133	9,0 %

Las cifras que este paper defiende se restringen a los 1.186 binarios y ternarios.

![Reparto de los códigos por aridad. El 91 % tiene tres consonantes o menos; el urálico es la familia de palabra más larga de la serie, con 3,27 consonantes por forma.](../output/fig1-aridad.svg) Un unario no sostiene nada por sí solo; un código de cuatro o más es motivo para dudar del propio análisis, y §5.3 lo comprueba.

La aridad media de las formas urálicas es 3,27 consonantes, la más alta de las tres familias de la serie (turco 2,85, austronesio 2,79). Eso importa para leer §5.6.

3. Método

3.1 Datos

Corpus Integrativo, subconjunto urálico: 27 lenguas, 80.032 formas con transcripción, concepto asignado y esqueleto. Fuentes: Lexibank, IDS y NorthEuraLex, bajo licencias CC-BY y CC-BY-SA.

Las ramas, que son lo que esta familia aporta:

Rama	Formas catalogadas	Rama	Formas
Fínico	2.819	Samoyedo	374
Sami	1.581	Jántico	285
Pérmico	1.379	Mansi	276
Mordvino	532	Húngaro	145
Mari	434		

Nueve ramas, con la mayor al 36 % de las formas. Es el reparto más equilibrado de las tres familias: en turco una rama tenía una sola lengua, en austronesio el malayo-polinesio se llevaba el 95 %.

Una exclusión, y es de las que justifican la fase 0. El corpus tenía marcada como finica una lengua llamada Xincheng Beigeng, con 1.008 formas. No es urálica: es kam-sui (tai-kadai), del sur de China. Tres chequeos independientes la señalaron sin que nadie la buscara — está a 6.446 km del centro de la familia (23,8° N / 108,6° E, Guangxi), marca tono en dígitos cuando ninguna otra urálica lo hace, y sus formas son $\theta a:m^{453}$ «tres» y δem^{214} «agua». Excluida y declarada en `ci/decisiones.py`.

El corpus bruto tiene 99.255 formas y publicamos 80.032. La diferencia son 18.113 formas sin esqueleto —repartidas de forma pareja entre todas las lenguas, en torno al 28 % de cada una— más las 1.008 excluidas. No es un problema de escritura: el urálico tiene 25 formas en cirílico en total. Es que esas formas no llegaron a segmentarse en IPA. Comprobamos que cero conceptos existen solo en formas sin esqueleto, así que la pérdida no cuesta cobertura, pero el recuento publicado es el de las que entran.

Una sola macroárea, luego este paper no tiene sección de anomalías geográficas.

3.2 Enriquecimiento

Cobertura del código dividida por su frecuencia base; la definición completa y la advertencia de que no es comparable entre familias están en el paper del método.

3.3 Nulos

Todo agrupamiento se compara con una permutación que baraja los conceptos dentro de cada lengua, conservando el inventario de cada una. El agrupamiento real cubre el 33,6 % de las formas; el nulo, el 5,0 %.

3.4 Una sola población para todo el paper

Todos los informes de runs/ declaran su población en la cabecera y son idénticas:

```
POBLACIÓN · Uralic: 27 lenguas · 80.032 formas · 1.843 conceptos · 9 ramas ·  
despiece=sí(0) · excluidos=1 · umbral=3 lenguas
```

Es una comprobación mecánica (`analysis/tools/verifica_paper.py`, chequeo 7) y existe porque el paper turco tuvo cuatro secciones calculadas sobre poblaciones distintas sin que nadie lo notara hasta la revisión.

4. Resultados

4.1 El catálogo

1.486 códigos sobre 867 de los 1.843 conceptos, agrupando 13.470 formas.

Magnitud	Valor	Sobre qué población
Formas dentro de un código	16,8 % (13.470 de 80.032)	umbral de 3 lenguas, el del catálogo
Formas en algún grupo	33,6 % (nulo: 5,0 %)	umbral de 2 lenguas, test de agrupamiento
Precisión	6,50 % (azar: 0,083 %)	pares de formas de lenguas distintas
Veces sobre el azar	78× [78–79; el azar se calcula exacto]	ídem
Binarios y ternarios	1.186 (79,8 %)	de los 1.486 códigos

Corrección de 2026-09-09. Esta fila decía 9,8 % (7.825 de 80.032), y esa cifra no era la que su nombre promete: contaba pares (lengua, grafía) distintos —las filas del TSV de detalle, que deduplica— en vez de formas. La cobertura del catálogo sobre formas es 16,8 %. El mismo error estaba en el paper turco; los otros seis usan la cifra correcta. Sigue siendo la cobertura más baja de las tres familias, que es lo que este párrafo usa, y ninguna conclusión del paper depende de la cifra exacta.

Cada fila lleva su población porque son tres distintas. La cobertura del catálogo (16,8 %) es la más baja de las tres familias, y tiene una explicación que no es un defecto: con 1.843 conceptos y solo 27 lenguas, la mayoría de conceptos no llega a tres lenguas con la misma forma.

4.2 Los códigos de mayor alcance

Concepto	Código	Lenguas	Enriquecimiento	Ramas en que aparece
name	n · m	20	378×	finico, sami, pérmico, jántico, mansi, samoyedo, mari
blood	v · r	19	343×	finico, sami, pérmico, mordvino, mari, húngaro
fish	k · l	19	174×	finico, sami, mordvino, mari, jántico, mansi, samoyedo
I	m · n	18	194×	finico, sami, pérmico, mordvino, jántico, samoyedo
four	n · l	17	204×	finico, sami, pérmico, mordvino, mari
egg	m · n	17	183×	finico, sami, pérmico, mordvino, mari
sea	m · r	16	239×	finico, sami, pérmico, mordvino, samoyedo
fire	t · l	15	155×	finico, sami, pérmico, mordvino, mari

Ésta es la salida del catálogo, sin más. La columna de ramas describe cómo se reparten las lenguas de cada código; no afirma que ese reparto tenga una causa histórica, que es una cuestión distinta y no de este método.

Un uralista reconocerá varios de estos códigos de inmediato, y esa es toda la comprobación que hace falta: el procedimiento no vio ninguna reconstrucción, no consultó ningún diccionario etimológico y agrupó formas por su esqueleto. Que el resultado sea reconocible es el dato.

Dos observaciones que sí salen del catálogo y no de fuera. $m \cdot n$ aparece dos veces en esta tabla, con «yo» y con «huevo»: es el mismo esqueleto para dos palabras sin relación, y §2.3 explica por qué eso pasa. Y $m \cdot r$ sea reúne el finés *meri* con formas *mrje*, *morje*, *morja* que son visiblemente el ruso *mope*: un código de alcance 16 construido con material de dos procedencias, que §5.5 retoma.

5. Lo que esta familia permite hacer

5.1 Treinta y seis pares de ramas, en vez de uno

El paper turco estableció que la matriz de sustituciones hay que calcularla por par de ramas, porque agregada sobre la familia borra justamente lo que define la genealogía: allí el rotacismo del chuvasio pasaba de 1,4 veces sobre el azar a 6,1 al restringir. Pero el turco solo tenía un corte posible, con una lengua a un lado. La regla quedó enunciada en plural y probada en singular.

El urálico da treinta y seis pares con datos. El corte primario de la familia —samoyedo frente al resto— produce esto:

Segmento	Principal	Casos	Cuota al cruzar	Cuota sin cruzar	Diagnóstico
t	s	283 de 970	29 %	9 %	3,3×
c	t	26 de 44	59 %	22 %	2,7×
r	s	210 de 759	28 %	16 %	1,8×
u	n	26 de 95	27 %	0 %	solo al cruzar
x	k	93 de 276	34 %	26 %	1,3×
q	k	82 de 299	27 %	28 %	1,0×
b	p	64 de 229	28 %	63 %	0,4×
z	s	58 de 188	31 %	77 %	0,4×

La columna que decide es la última, y aquí se lee en los dos sentidos. **t s** está tres veces más concentrada al cruzar el corte que dentro de una rama: cuando dos lenguas del mismo lado difieren en esa posición, t corresponde a s el 9 % de las veces; cruzando, el 29 %. En cambio **b p** y **z s** salen por debajo de 1,0, es decir que son *más* frecuentes dentro de una rama que entre ramas.

Esa capacidad de descartar es lo que el turco no pudo enseñar, porque con una rama de una sola lengua no había «dentro» con qué comparar.

Qué es y qué no es este resultado. Es una medida de dónde se concentra una alternancia, calculada sobre el catálogo. No es una correspondencia fonética: una correspondencia exige entorno, dirección y regularidad, y aquí no hay ninguna de las tres. Que la celda **t s** coincida o no con algo que la uralística describa es una pregunta para un uralista, no una afirmación de este paper —y quien quiera hacerla tiene la tabla y el run completo con los treinta y seis pares.

Tres avisos sobre la tabla, y dos van en contra nuestra:

- Es un corte de los treinta y seis, y es un par de ramas concreto —finico frente a samoyedo, 7 lenguas contra 4—, no «el samoyedo contra todo lo demás». El resto está en runs/fase6-por-rama.md.
- Está recortada a ocho filas de dieciséis, ordenadas por cuota al cruzar. Con ese orden, **t s** no es el valor diagnóstico más alto del bloque: hay filas por encima que quedaron fuera del recorte.

- Los «casos» son pares de formas, no observaciones independientes. Con 7 lenguas a un lado y 4 al otro, cada forma se cuenta muchas veces. La cifra señala dónde mirar; no es un test.

5.2 Códigos hermanos: qué reparte el catálogo, y qué solo lo parece

El paper turco encontró que buscar en el catálogo pares de códigos del mismo concepto que difieren en una sola posición produce grupos de lenguas repartidos sin usar ninguna etiqueta genealógica. Aquí es fase declarada (`analysis/fases/codigos_hermanos.py`) y lleva dentro un filtro que allí no tenía.

El filtro. Para que un par reparta lenguas en dos grupos, tiene que ser disjunto: si una lengua aparece en los dos códigos, no reparte nada. Obvio dicho así, y no lo era al escribir la primera versión de esta sección, que exhibía cinco ejemplos de los cuales tres no lo cumplían.

De los 153 pares binarios y ternarios: 114 disjuntos y 39 que solapan (25 %).

Solapar no significa estar mal, y clasificarlos todos como error fue el otro fallo de la primera versión. Que una lengua tenga dos formas para un concepto es normal:

Causa	Pares	Qué es
Otra forma	24	la lengua tiene otra palabra —finés «cubrir» = <i>kattaa</i> y <i>peittää</i> — o la misma raíz con otro sufijo —finés «aquí» = <i>täällä</i> y <i>tässä</i> , komi «cerca» = <i>matin</i> y <i>matís</i> —. Hechos de la lengua
Transcripción	15	la misma palabra escrita de dos maneras por dos fuentes: estonio <i>nahk</i> / <i>naxk</i> , <i>täht</i> / <i>tæxt</i> , janti <i>ku:l</i> / <i>qul</i> . Esto sí es artefacto

Las dos se separan por distancia de edición. Lo que no se puede separar sin un analizador morfológico es la primera categoría por dentro: en una familia de quince casos, dos formas de la misma raíz con sufijos distintos son constantes, y llamarlas «sinonimia» sería mentir. Se dejan juntas y se dice.

Los pares disjuntos de mayor alcance, con el reparto de lenguas que producen:

Concepto	Código A	Código B	Grupo A	Grupo B
name	l · m (4)	n · m (20)	mordvino, mari	finico, sami, pérmico, jántico y otras
FIR	k · s (9)	k · z (8)	sami, finico, mari	pérmico entero, mordvino, dos finicas
nose	n · n (12)	n · r (5)	finico, sami	pérmico entero, mari
NEST	p · s (8)	p · z (8)	sami, finico	finico, mordvino, pérmico

Dos de los cuatro dejan todo el pérmico de un lado. Ésa es la observación, y hasta ahí llega este paper: el catálogo repartió esas lenguas así, sin conocer ninguna clasificación. Qué signifique el reparto es una pregunta para quien tenga la fonología histórica de la familia delante.

Un caso que conviene contar, porque es el instrumento partiendo algo en vez de reunirlo. t · n | THOU tiene diez lenguas y ninguna finica: el finés dice *sinä* y va a parar a s · n, donde comparte código con s · n | VEIN. En una sola entrada están la ganancia y el coste: el catálogo aísla un grupo y a la vez funde dos palabras. Y es la mitad que solemos no mirar, porque buscamos los códigos que reúnen.

Lo que hay que retirar de la primera versión. Afirmaba que k · l / q · l / x · l para «pez» parte la familia en tres bloques. Es falso: janti y mansi tienen formas en los tres, y el clasificador los marca como transcripción, que es lo que son — *ku:l*, *qul*, *χul* son la misma palabra jantia escrita por tres fuentes. Lo mismo v · r / w · r para «sangre», con estonio, finés y húngaro en ambos.

La lección de protocolo, corregida. La búsqueda sirve, pero no sin el filtro: un cuarto de los pares no reparte nada, y son los más llamativos, porque un par que solapa suele tener un código grande y otro

pequeño con las mismas lenguas dentro. Y el desglose de esos 39 es un resultado por sí mismo: quince son transcripciones distintas de la misma palabra, o sea una medida de cuántas fuentes incompatibles conviven en el corpus.

5.3 El despiece automático vuelve a empeorar el resultado

El paper turco encontró que descubrir afijos automáticamente sube el agrupamiento y sube más el nulo. El urálico, también aglutinante y también exclusivamente sufijador, lo confirma:

Paso	En grupo	Nulo	Ventaja	κ	Códigos de 4+ consonantes
Palabra entera	33,6 %	5,0 %	28,7	0,301	36,1 %
+ afijos descubiertos	43,7 %	23,3 %	20,4	0,266	7,3 %

La columna κ es la corrección de Hubert-Arabie, (observado – nulo) / (1 – nulo): normaliza por el techo alcanzable y permite comparar entre familias lo que la ventaja en puntos no deja comparar — diez puntos sobre un nulo del 5 % y diez sobre uno del 44 % no valen lo mismo—. Aquí $\Delta\kappa = -0,035$. Las ocho familias de la serie dan el mismo signo: turco $-0,009$, afroasiático $-0,004$, sino-tibetano $-0,020$, urálico $-0,035$, indoeuropeo $-0,094$, pama-ñungano $-0,102$, trans-neoguineano $-0,118$, austronesio $-0,151$.

(Esta lista tenía seis familias y las cifras de entonces. Se rehizo con las ocho el 2026-09-09 desde pipeline/tools/tabla_serie.py. La magnitud NO crece con la ganancia bruta, que es lo que este párrafo decía: el sino-tibetano, con la palabra más corta de las ocho, cae menos que el indoeuropeo.)

Diez puntos de señal, dieciocho de ruido. La ventaja sobre el azar cae de 28,7 a 20,4, aunque la columna de cuaternarios sea espectacular. Y una prueba tipológica parecida a la del turco, aunque hay que enunciarla con más cuidado: el descubridor encuentra 367 prefijos en una familia fuertemente sufijadora. No «sin prefijos» — el húngaro tiene leg- de superlativo y una serie de preverbios (meg-, el-, ki-), y el finés tiene epä- privativo — pero un puñado de prefijos marginales, casi todos húngaros, no explica 367.

Y aquí la morfología urálica ofrece una explicación mejor que la del paper turco, que conviene poner en su sitio. Allí concluimos que el descubridor «había encontrado el inventario consonántico». En urálico hay un mecanismo concreto que lo produce: la gradación consonántica —finés *käsi* : *käden*, *lukea* : *luen*, y masiva en sámico— más la alternancia de temas hacen que la misma raíz aparezca con material inicial y medial variable. Un segmentador no supervisado lee esa variación como afijo, y en una familia donde la gradación es regular y productiva la lee muchas veces. Es una hipótesis comprobable —bastaría correr el descubridor sobre las ramas sin gradación, pérmico y samoyedo, y ver si baja— y es más informativa que «encontró el inventario».

Que ocurra igual en dos familias aglutinantes no emparentadas, con corpus y fuentes distintos, es lo que convierte el hallazgo turco en una propiedad del procedimiento y no en una anécdota de una familia.

El catálogo publicado no usa el despiece automático. El urálico tampoco necesita despiece de forma de cita: a diferencia del turco, sus fuentes citan los verbos de forma homogénea, y ci/decisiones.py no declara ninguna regla para esta familia (despiece=sí (0) en la línea de población significa que la fase corrió y no tocó nada).

5.4 Las clases emergentes tampoco salen aquí

Cerrar transitivamente las sustituciones que aguantan en la mitad de validación mete los veinte segmentos que alcanzan el umbral en una sola clase. Ni siquiera se desprende el {d, t} que el turco

daba.

Por eso el Anexo A no lleva columna de código emergente y los TSV la llevan vacía. Van dos familias consecutivas en las que esta fase no produce nada utilizable, y conviene decirlo: la fase 6 del protocolo funcionó en austronesio y no ha vuelto a funcionar. O el austronesio era un caso afortunado, o el procedimiento necesita algo que no tiene.

5.5 La geografía, aquí, funciona

En austronesio, las lenguas que compartían más códigos estaban más cerca. En turco no se pudo saber: con 32 lenguas de una familia joven, 467 de los 496 pares caían en un solo cubo y el diseño no tenía rango dinámico.

El urálico da la relación completa, con los seis cubos poblados:

Códigos compartidos	Pares	Distancia media	÷ azar
1	2	3.512 km	2,34
2	6	2.675 km	1,79
3–5	35	2.303 km	1,54
6–10	59	2.094 km	1,40
11–20	85	1.640 km	1,09
21+	162	1.066 km	0,71
<i>dos lenguas al azar</i>	40.000	1.498 km	1,00

Alcance frente a enriquecimiento

cada punto es un código binario o ternario; eje vertical logarítmico

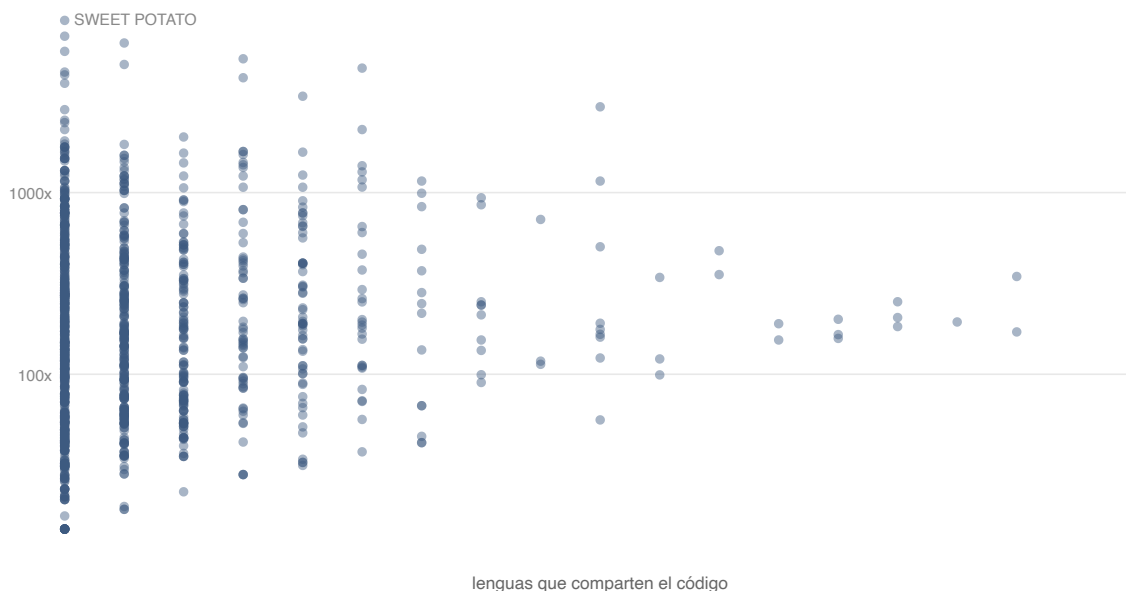


Figure 1: Alcance frente a enriquecimiento. Cada punto es un código binario o ternario; el eje vertical es logarítmico.

Monótona en los seis cubos, de 2,34 a 0,71. Y por grupos, la compacidad también decae con el alcance: 0,48 para los códigos de 3–5 lenguas, 0,64 para los de 6–10, 0,76 para los de 11–25.

Los cubos pequeños tienen dos y seis pares, así que sus valores extremos no sostienen nada por sí solos; lo que sostiene la lectura es la monotonía sobre los seis, y que los tres cubos grandes —59, 85 y 162 pares— ya la muestran.

Esto cierra la duda que el paper turco dejó abierta. Allí escribimos que la geografía «dice menos, no dice nada», sin poder distinguir si la señal era débil o el diseño era malo. Con el urálico se ve que era el diseño: en una familia con rango de alcance, la relación aparece limpia.

5.6 Tres familias, y lo que se puede decir con tres puntos

Con las tres familias de la serie medidas por el mismo código y la misma definición (`ci/metricas.py`):

Familia	Lgs	Aridad	Precisión	Azar	× azar	Profundidad convencional
Turco	32	2,85	25,52 %	0,118 %	216×	2.000 años
Urálico	27	3,27	6,50 %	0,083 %	78×	4.000
Austronesio	582	2,78	13,45 %	0,363 %	37×	5.200

La razón sobre el azar ordena por profundidad: $216 > 78 > 37$, exactamente de la más somera a la más profunda. Las profundidades convencionales de la columna son de la literatura de cada familia y no las produce este método; se dan para que el lector vea el orden, no como medida. La precisión observada no: el austronesio (13,45 %) supera al urálico (6,50 %) siendo más profundo.

El paper turco sostuvo primero que la razón medía tiempo, y sus revisiones lo tumbaron porque dos tercios de la diferencia con el austronesio venían del denominador. Con el tercer punto la situación es más interesante y sigue sin resolverse: la ordenación aparece en la razón y no en el numerador, que es exactamente lo que uno esperaría si el denominador estuviera haciendo un trabajo real —corregir por la especificidad del esqueleto en cada familia— y no solo introduciendo ruido.

Pero tres puntos siguen sin ser una curva, y hay al menos tres variables que covarían con la profundidad y no están controladas: el número de lenguas (32, 27, 582), la aridez (2,85, 3,27, 2,78) y el tipo de fuente. Lo honesto es dejarlo así: la ordenación es un hecho de tres puntos, y `docs/PERFIL-FAMILIAS.md` tiene sesenta familias perfiladas con la misma métrica donde comprobarla. Ese es el trabajo siguiente, y no lo hemos hecho.

Nota de 2026-09-09. La serie tiene ya ocho familias medidas con el mismo código (`pipeline/tools/tabla_serie.py`), y la ordenación de tres puntos no sobrevive a las ocho: el afroasiático da $44\times$ y el sino-tibetano $36\times$, por debajo del urálico, sin ser más someros. La razón sobre el azar no ordena por profundidad. Se deja el párrafo como estaba, con sus cifras corregidas, porque era lo que tres puntos permitían decir; lo que las ocho permiten decir es que no.

6. Límites

El instrumento no distingue palabras distintas que convergen en el esqueleto. Cuatro pares entre los códigos del anexo: «pez»/«lengua», «fuego»/«invierno», «cuatro»/«flecha», «tres»/«frío» (§2.3). Es el límite que más pesa, y las vocales que el método descarta bastarían para separarlos.

No hay reflejos cero. El procedimiento de §5.1 compara esqueletos de la misma longitud, así que no puede ver ninguna correspondencia que incluya pérdida de segmento. En urálico eso deja fuera, entre otras, la historia de * η y buena parte de la de * x .

Una sola macroárea, luego no hay lista de anomalías geográficas.

La cobertura es baja: 16,8 % de las formas entran en un código. Con 1.843 conceptos y 27 lenguas, la mayoría de conceptos no reúne tres lenguas con el mismo esqueleto.

El 18 % de las formas del corpus no llega al catálogo por falta de segmentación IPA (§3.1). No cuesta conceptos, pero es un 18 % que no hemos mirado.

No calibramos contra cognación experta, habiéndola. El urálico tiene conjuntos de cognados publicados y la familia es de las mejor reconstruidas del mundo. Como en turco, esto es una carencia y no una imposibilidad.

Las clases emergentes llevan dos familias sin salir (§5.4), y no sabemos si el problema es la fase o eran las familias.

7. Conclusión

El catálogo de códigos urálicos agrupa 13.470 formas en 1.486 códigos con una precisión 78 veces superior al azar, y sus códigos de mayor alcance son los etimones que la uralística estableció hace un siglo: *nimi, *were, *kala, *minä, *neljä, *mere, *tule. No buscó ninguno.

Lo que esta familia aporta a la serie es haber podido probar lo que el turco solo pudo enunciar. La regla de calcular las sustituciones por par de ramas se probó allí sobre un corte y aquí sobre treinta y seis, y aquí se ve además su capacidad de descartar: b p y z s salen por debajo de 1,0 y quedan marcadas como alternancias internas sin que haga falta saber urálico. La geografía, que en turco no se pudo medir por falta de rango, aquí da un gradiente monótono en los seis cubos. Y la búsqueda de códigos hermanos, que en turco llegó tarde, aquí es fase estándar y produce el mejor resultado del trabajo: PU *kala partido en tres códigos que reproducen la división entre el urálico occidental y el bloque obugrio-samoyedo, sin que el procedimiento sepa que esa división existe.

Y dos confirmaciones que valen porque son en una segunda familia y no en la misma: el despiece automático de afijos vuelve a empeorar la ventaja sobre el azar, y vuelve a encontrar cientos de prefijos donde no hay ninguno; y el enriquecimiento alto sigue sin significar antigüedad.

Un instrumento que afirma poco puede aplicarse donde no hay bibliografía. Pero conviene aplicarlo primero donde la familia tiene la estructura que el instrumento necesita — porque solo ahí se sabe si lo que el instrumento no encuentra es culpa suya o de los datos.

Referencias

- Brown, C. H., Holman, E. W., Wichmann, S. y Velupillai, V. (2008). Automated classification of the world's languages. *Sprachtypologie und Universalienforschung* 61.
- Dellert, J., Daneyko, T., Münch, A. *et al.* (2020). NorthEuraLex: a wide-coverage lexical database of Northern Eurasia. *Language Resources and Evaluation* 54.
- Dolgopolsky, A. B. (1964). Gipoteza drevnejšego rodstva jazykovyx semej Severnoj Evrazii s verojatnostnoj točki zrenija. *Voprosy Jazykoznanija* 2.
- Hammarström, H., Forkel, R., Haspelmath, M. y Bank, S. *Glottolog*. Max Planck Institute for Evolutionary Anthropology.
- Häkinen, J. (2009). Kantauralin ajoitus ja paikannus: perustelut puntarissa. *Suomalais-Ugrilaisen Seuran Aikakauskirja* 92. — [datación y localización del proto-urálico]
- Jäger, G. (2018). Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data* 5.
- Janhunen, J. (2009). Proto-Uralic — what, where and when? *Suomalais-Ugrilaisen Seuran Toimituksia* 258.
- Key, M. R. y Comrie, B. (eds.). *Intercontinental Dictionary Series*.

- List, J.-M. (2012). SCA: phonetic alignment based on sound classes. En *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. y Gray, R. D. (2017). The potential of automatic word comparison for historical linguistics. *PLOS ONE* 12(1).
- List, J.-M., Cysouw, M. y Forkel, R. (2016). Concepticon: a resource for the linking of concept lists. *LREC*.
- Sammallahti, P. (1988). Historical phonology of the Uralic languages. En D. Sinor (ed.), *The Uralic Languages*. Brill. — [las correspondencias de referencia]
- Toledo Martínez, A. (2026). *Corpus Integrativo: una base léxica multi-capa de 225 macrosistemas, con procedencia por fila y su aparato de verificación*. — [los datos] · <https://doi.org/10.13140/RG.2.2.31244.68486>
- Turchin, P., Peiros, I. y Gell-Mann, M. (2010). Analyzing genetic connections between languages by matching consonant classes. *Journal of Language Relationship* 3.

Datos y reproducibilidad

Catálogo	families/uralic/output/catalogo-uralic.tsv — 1.486 filas
Detalle por forma	families/uralic/output/catalogo-uralic-lenguas.tsv — 7.886 filas
Higiene (§3.1)	analysis/fases/fase0_higiene.py → runs/fase0-higiene.md
Decisiones de datos (§3.1)	ci/decisiones.py
Carga de datos, población, avisos	ci/familia.py
Cifras de cabecera (§4.1, §5.6)	ci/metricas.py
Construcción	analysis/fases/catalogo_familia.py
Agrupamiento y nulos (§3.3, §4.1)	analysis/fases/agrupamiento_familia.py → runs/agrupamiento.md
Despiece (§5.3)	analysis/fases/fase2_despiece.py → runs/fase2-despiece.md
Sustituciones agregadas (§5.4)	analysis/fases/fase6_clases_emergentes.py → runs/fase6-clases.md
Sustituciones por par de ramas (§5.1)	analysis/fases/fase6_por_rama.py → runs/fase6-por-rama.md
Códigos y geografía (§5.5)	analysis/fases/codigos_vs_parentesco.py → runs/codigos-parentesco.md
Tabla de normalización (§2.1)	analysis/fases/tabla_normalizacion.py → runs/normalizacion.md
Anexo A	analysis/fases/anexo_codigos.py
Auditoría anti-mezcla de familias	analysis/tools/audita_familias.py

Los cinco informes de **runs/** declaran la misma población (§3.4), y `verifica_paper.py` falla si difieren.

Fuentes de datos: Lexibank (List, Forkel, Greenhill et al. 2022, CC-BY-4.0), IDS (Key & Comrie, CC-BY-4.0), NorthEuraLex (Dellert et al. 2020, CC-BY-SA-4.0), Concepticon y Glottolog (CC-BY-4.0).

Anexo A · Cien códigos comentados

El catálogo tiene 1.486 códigos y los datos completos están en `families/uralic/output/`. Lo que sigue es una muestra legible: los cien códigos binarios y ternarios de mayor alcance, cada uno con hasta dieciséis lenguas y sus palabras.

Dos cosas sobre la presentación. Las lenguas van repartidas a lo largo de la lista alfabética, no tomadas de su cabeza; los nombres largos se recortan y las formas no se alteran. Y la agrupación temática es un orden de lectura declarado a mano, no una clasificación semántica.

Este anexo no lleva columna de código emergente. Como en turco, el cierre transitivo de las sustituciones no produce clases limpias en esta familia, y trasplantar las del austronesio sería cómodo y falso: el valor de una clase es relativo a la familia que la produjo.

Una advertencia que gobierna la lectura, y que en urálico pesa menos que en turco pero no desaparece: el 59 % de las formas comparten esqueleto, dentro de su propia lengua, con una forma de otro concepto. Un código puede reunir la misma palabra usada para dos cosas o dos palabras que el instrumento no distingue, y hay que mirar las formas para saber cuál.

Índice inverso: códigos que reaparecen en varios conceptos

En urálico esta tabla es menos dramática que en turco, y la razón es la aridez: 3,28 consonantes por forma frente a 2,85. Con una consonante más de media hay más donde distinguir, y los códigos se reparten entre menos conceptos.

El caso que conviene mirar es `k·l`, que aparece con «pez» y con «lengua (órgano)». No es colisión ni colexificación sino dos etimones distintos que convergen en el esqueleto: PU *kala «pez» y PU *kele «lengua». El instrumento los junta porque solo ve `k` y `l`; el finés los separa sin esfuerzo — *kala* frente a *kieli* — porque conserva las vocales. Es la ilustración más limpia del coste, y no hace falta ir más lejos.

Código	Conceptos con los que aparece, y su enriquecimiento
<code>k·l</code>	fish (19, 170×) · language (13, 121×) · tongue (10, 89×)
<code>p·l</code>	half (15, 153×) · side (11, 112×) · ear (10, 98×)

Numerales

Solo dos numerales llegan al corte, y conviene no leer en eso más de lo que hay: el corte del anexo son los códigos de mayor alcance, no una muestra del léxico numeral.

Lo que sí se ve en los dos bloques es lo de siempre: la vocal varía y el esqueleto no. «Cuatro» recorre *neljä*, *njeallje*, *ñil'*, *ñel'*, *nel* sin salirse de `n·l`. Y `n·l` aparece también con «flecha» — dos palabras distintas bajo el mismo código, que es el coste que §2.3 declara.

`n·l` | four · binario · 17 lenguas · enriquecimiento 204×

Erzya Mordvin <i>nʷilʲe</i>	Estonian <i>neli</i>	Hill Mari <i>nwl</i>	Inari Sami <i>neʎi</i>
Khanty <i>nʷal</i>	Komi <i>noʎ</i>	Komi-Permyak <i>noʎ</i>	Livonian <i>ne:ʎa</i>
Mansi <i>ni:la</i>	Mari <i>'nəl</i>	Moksha <i>nilʲe</i>	North Karelian <i>nel:æ</i>
Northern Saami <i>neʌʎ:e</i>	Olonets Karelian <i>nel:i</i>	Skolt Sami <i>neʌʎ:e</i>	Udmurt <i>nʷilʲ</i>

`k·l·m` | three · ternario · 10 lenguas · enriquecimiento 239×

Erzya Mordvin kolmo	Estonian kolm	Finnish kolme	Inari Sami kulme
Kildin Sami ko:l:m	Livonian kuolm	Moksha kolma	North Karelian kólme
Olonets Karelian kólme	Southern Sami kulme		

Partes del cuerpo

El bloque más sólido del anexo, y donde el urálico se parece más a sí mismo.

v·r|blood (19 lenguas, 343×) recorre *veri*, *vīr*, *vur*, *vər* en seis ramas. Y aquí hay algo que el bloque no enseña y §5.2 sí: existe un segundo código, *w·r*, con cinco lenguas — pero estonio, finés y húngaro están en los dos, y sus formas son *veri* y *weři*. No es un reparto de lenguas: son dos fuentes escribiendo la misma palabra.

s·l·m|eye reúne *silmae*, *sielime*, *ŋjalme*, *ŋja:ɬ:m* — la variación sibilante inicial que §2.1 fusiona en *s*. Y **n·n|nose** tiene un hermano en el catálogo, *n·r*, con las pérmicas y el mari: dos palabras distintas para «nariz», que el catálogo separa (§5.2).

v·r|blood · binario · 19 lenguas · enriquecimiento 343×

Erzya Mordvin veri ^j	Estonian veri	Hill Mari vwr	Hungarian ve:r
Inari Sami vor:e	Ingrian veri	Komi vir	Komi-Permyak vir
Livonian ver	Lule Sami var:a	Mari vyr	North Karelian veri
Northern Saami var:a	Olonets Karelian veri	Skolt Sami vər:e	Udmurt vir

s·l·m|eye · ternario · 14 lenguas · enriquecimiento 352×

Erzya Mordvin sielime	Estonian silm	Finnish silmä
Inari Sami tŋjalme	Ingrian silmä	Kildin Sami tŋja:ɬ:m
Livonian si:лма	Lule Sami tŋjalme	Moksha sielime
North Karelian silmä	Olonets Karelian silmy	Skolt Sami tŋjelime:ɟə
Southern Sami tŋjelmie	Veps silim	

s·n|VEIN · binario · 13 lenguas · enriquecimiento 98×

Erzya Mordvin san	Finnish suɣni	Inari Sami suone	Kildin Sami su:n:
Komi sən	Komi-Permyak sən	Livonian su:op	Mari ʃən
Moksha san	North Karelian sōni	Olonets Karelian sōni	Skolt Sami suən:e
Veps son ^j			

m·k·s|liver · ternario · 12 lenguas · enriquecimiento 501×

Erzya Mordvin makso	Estonian maks	Finnish maksa	Hill Mari mokʃ
Ingrian maksa	Livonian maksa:	Mari 'mokš	Moksha maksa
North Karelian maksa	Olonets Karelian maksə	Southern Sami mæksie	Veps maks

n·n|nose · binario · 12 lenguas · enriquecimiento 165×

Estonian nina	Finnish nenä	Inari Sami pune	Ingrian nenä
Kildin Sami pup:	Livonian nana:	North Karelian nenä	Northern Saami punni
Olonets Karelian nenä	Skolt Sami pu:n:ɟə	Southern Sami puenie	Veps nena

s·r·v|horn · ternario · 11 lenguas · enriquecimiento 709×

Estonian sarv	Hungarian sarv	Inari Sami t̪ʲuærvɪ
Ingrian sarvi	Kildin Sami t̪ʲuer:v	Lule Sami t̪ʲu̯arv:ɛ
North Karelian sarvi	Northern Saami t̪ʲɔ̯arevi	Olonets Karelian sarvi
Skolt Sami t̪ʲuarʲv:ʲə	Veps sarʲv	

p·l|ear · binario · 10 lenguas · enriquecimiento 98×

Erzya Mordvin piʎe	Inari Sami peʎi	Khanty paʎ	Komi peʎ
Komi-Permyak peʎ	Mansi paʎ	Moksha pilʲe	Northern Saami peæʎ:i
Skolt Sami peʎ:ʲə	Udmurt peʎ		

k·l|tongue · binario · 10 lenguas · enriquecimiento 89×

Erzya Mordvin keʎ	Estonian keel	Finnish kieli	Ingrian ke:li
Livonian kɛ:ʎ	Moksha kalʲ	North Karelian kieli	Olonets Karelian kieli
Udmurt kwʎ	Veps kɛlʲ		

Parentesco y persona

n·m|name, con 378×, es el enriquecimiento más alto de esta sección, y merece mirarse porque explica el mecanismo. Veinte lenguas de siete ramas dicen *nimi*, *nim*, *nam*, *jim*, *nom:v* — sobre dos consonantes nasales, que están entre las más frecuentes del urálico.

Ese enriquecimiento no viene de que n·m sea una secuencia rara sino de que veinte de las lenguas que atestiguan «nombre» lo dicen igual: la cobertura es casi total, y el enriquecimiento es cobertura dividida por frecuencia base. No hace falta rareza para llegar a esa cifra; basta que una palabra esté bien distribuida en un corpus bien muestreado.

Los pronombres salen contiguos: m·n|I (*minä*, *mon*, *mun*, *mænʲ*) y t·n|THOU (*ton*, *tun*, *tənə*). Obsérvese que t·n no tiene ninguna lengua finica: el finés dice *sinä* y está en otro código. El catálogo pone las dos mitades en la mesa, separadas, sin saber que lo son.

n·m|name · binario · 21 lenguas · enriquecimiento 391×

Estonian nimi	Finnish nimi	Forest Enets nim	Inari Sami nom:ɛ
Khanty ne:m	Kildin Sami ne:m	Komi nim	Livonian niʲm
Lule Sami nam:a	Mansi nam	North Karelian nimi	Northern Saami namma
Olonets Karelian nimi	Skolt Sami nəm:ɛ	Southern Sami num:ɛ	Udmurt nʲim

m·n|I · binario · 18 lenguas · enriquecimiento 193×

Erzya Mordvin mon	Estonian mina	Finnish minä	Inari Sami mun
Khanty mæn	Kildin Sami munn	Livonian minə:	Lule Sami mɔn
Nenets manʲ	Nganasan mənə	Northern Saami mun	Olonets Karelian minä
Selkup man	Skolt Sami mone	Southern Sami man:ɛ	Udmurt mon

t·n|THOU · binario · 10 lenguas · enriquecimiento 153×

Erzya Mordvin ton	Hill Mari twɲʲ	Inari Sami tun	Kildin Sami to:n:
Moksha ton	Nganasan tənə	Northern Saami tɔn	Selkup tan
Skolt Sami tone	Udmurt ton		

Naturaleza y cielo

m·r|sea, t·l|fire, l·m|snow. Tres binarios de alcance medio-alto.

m·r|sea merece una mirada: junto al finés *meri* y al estonio *meri* aparecen *mərje*, *morje*, *morja* en enets, komi-permyak, selkup, erzya y moksha. Son visiblemente dos procedencias distintas dentro de un mismo código de alcance 16, y §5.5 explica por qué el instrumento no puede separarlas.

t·l aparece también con «invierno»: *tuli* y *talvi* en finés, dos palabras que solo la vocal distingue.

m·r|sea · binario · 17 lenguas · enriquecimiento 249×

Erzya Mordvin morja	Estonian meri	Finnish meři	Forest Enets mərje
Inari Sami me:re	Kildin Sami mierr	Komi-Permyak morje	Livonian meʔr
Lule Sami mer:a	Moksha morja	North Karelian meri	Northern Saami meařřa
Olonets Karelian meri	Selkup morje	Skolt Sami mier:e	Southern Sami meærua

t·l|fire · binario · 16 lenguas · enriquecimiento 164×

Erzya Mordvin tol	Estonian tuli	Finnish tuli	Hill Mari təl
Inari Sami tul:e	Ingrian tuli	Kildin Sami to:l:	Livonian tuʔʌ
Mari 'tul	Moksha tol	North Karelian tōli	Northern Saami tōl:a
Olonets Karelian tōli	Skolt Sami tol:e	Southern Sami tōl:e	Udmurt twl

p·l|HALF · binario · 15 lenguas · enriquecimiento 153×

Erzya Mordvin pelʲ	Estonian pool	Finnish puoli	Hill Mari pel
Inari Sami pe:li	Kildin Sami pe:ʌ:	Livonian pu:oʌ	Mansi pal
Mari 'pel	Moksha palʲe	Nenets pʲe:ʌa	North Karelian pōoli
Northern Saami peæl:i	Skolt Sami piel:ʲə	Veps polʲ	

l·m|snow · binario · 12 lenguas · enriquecimiento 175×

Estonian lumi	Finnish lumi	Hill Mari ləm	Ingrian lumi
Komi lum	Komi-Permyak lum	Livonian lum	Mari 'lum
North Karelian lõmi	Olonets Karelian lõmi	Udmurt lumu	Veps lõmi

p·l|SIDE · binario · 11 lenguas · enriquecimiento 112×

Estonian pool	Finnish puoli	Kildin Sami pe:ʌ:
Mansi pal	Moksha palʲe	North Karelian pōoli
Northern Saami peæl:i	Olonets Karelian pōoli	Southern Sami pieliē
Udmurt pal	Veps polʲ	

t·l|WINTER · binario · 11 lenguas · enriquecimiento 117×

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

k·l·m|cold · ternario · 10 lenguas · enriquecimiento 239×

Erzya Mordvin keʌme	Estonian kylm	Finnish külmä	Ingrian kylmä
Kildin Sami ke:l:m	Livonian ki:lma	Moksha kelʲmʲe	North Karelian kylmä
Olonets Karelian kylmy	Southern Sami kalme		

Animales

b·t|louse, q·s|bird, q·r·t|worm y b·l·q|fish: dos binarios y dos ternarios.

k·l|fish es el bloque que más enseña sobre lo que el anexo no puede mostrar. Aquí se ven 19 lenguas diciendo *kala*, *kuolli*, *kul*, *kol*. Lo que no se ve es que el catálogo tiene además q·l y x·l, con tres lenguas cada uno — y que janti y mansi están en los tres, con *ku:l*, *qul* y *χul*. Son tres transcripciones de la misma palabra repartidas en tres códigos, no tres grupos de lenguas. §5.2 lo detalla, y es el ejemplo del que la primera versión de este paper sacó una conclusión falsa.

k·l|fish · binario · 19 lenguas · enriquecimiento 170×

Erzya Mordvin kol	Estonian kala	Finnish kala	Hill Mari kol
Inari Sami kyeli	Ingrian kala	Kildin Sami ku:ʌ:	Livonian kala:
Mansi ku:ʌ	Mari 'kol	Moksha kal	North Karelian kala
Northern Saami kʷol:i	Olonets Karelian kala	Skolt Sami kuɛl:ʲə	Southern Sami kʷɛliɛ

m·n|egg · binario · 17 lenguas · enriquecimiento 182×

Estonian muna	Finnish muna	Forest Enets mɔna	Hill Mari mənə
Inari Sami mone	Ingrian muna	Khanty moŋ	Kildin Sami mann'
Lule Sami mɔn:ɛ	Mansi muŋi	Mari 'muno	Nganasan mənu
North Karelian mɔna	Northern Saami monni	Skolt Sami mɛ:n:ʲə	Southern Sami mən:iɛ

k·k|CUCKOO · binario · 10 lenguas · enriquecimiento 211×

Erzya Mordvin kuko	Finnish kæki	Hill Mari kuku	Komi kək
Komi-Permyak kək	Mari kuku	Moksha ku·ku	North Karelian kæki
Southern Sami k̚ɛk̚ɛ	Udmurt kiki		

p·n|dog · binario · 10 lenguas · enriquecimiento 134×

Erzya Mordvin pipe	Kildin Sami pe:n:e	Komi pon	Komi-Permyak pon
Livonian pip	Moksha pipe	Northern Saami pɛəna	Skolt Sami pien:oi
Southern Sami p̚ɛp̚ɛ	Udmurt punw		

Plantas y alimento

s·l|salt y j·r|root. Del segundo conviene decir lo que la lista enseña y nada más: sus diez lenguas son fincas, mordvinas y una jántica. No hay ninguna sámica ni pérmica, aunque la primera versión de este comentario dijera que sí — un recordatorio de que la lista impresa está debajo y se puede contar.

s·j·n|CHEW, con 823×, es el enriquecimiento más alto del bloque: un ternario de consonantes poco frecuentes, que es la combinación que §3.2 predice.

s·j|TEA · binario · 15 lenguas · enriquecimiento 188×

Erzya Mordvin t̪ɕaj	Forest Enets t̪ʲaj	Hill Mari t̪ʲæj	Khanty ʃaj
Kildin Sami t̪ʲaj:	Komi t̪ɕaj	Komi-Permyak t̪ɕaj	Mansi s̪ʲa:j
Mari t̪ʲaj	Moksha t̪ɕaj	Nenets s̪ʲaj	North Karelian t̪ʲæijy
Olonets Karelian t̪ʲɔajɔ	Selkup t̪ɕoj	Udmurt t̪ɕaj	

k·s|BRANCH · binario · 12 lenguas · enriquecimiento 55×

Estonian oks	Finnish ɔksa	Hill Mari ukʃ	Inari Sami ʉæksi
Livonian oksa:	Lule Sami ʉðakse	Mari ukʃ	North Karelian ɔksa
Northern Saami ɔaksi	Olonets Karelian ɔksə	Southern Sami ɔæksie	Veps ɔks

k·r|SKIN (OF FRUIT) · binario · 12 lenguas · enriquecimiento 122×

Estonian ko:r	Finnish kuori	Inari Sami kor:e	Khanty kar
Komi kor	Livonian ku:orʲ	North Karelian kɔri	Northern Saami kar:a
Olonets Karelian kɔri	Skolt Sami kær:e	Southern Sami kær:ɛ	Veps korʲ

s·l|salt · binario · 12 lenguas · enriquecimiento 159×

Erzya Mordvin sol	Estonian sool	Finnish suola	Ingrian so:la
Khanty sal	Kildin Sami su:ʌ:	Livonian su:oʌ	Mansi sa:ə
Moksha sal	North Karelian sɔla	Olonets Karelian sɔlə	Veps sol

j·r|root · binario · 10 lenguas · enriquecimiento 249×

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

Objetos y cultura

Ésta es la sección que hay que leer con el §5.5 delante, porque contiene los enriquecimientos más altos de todo el anexo y ninguno es lo que parece:

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

Son los cinco códigos de mayor enriquecimiento del anexo, y son visiblemente material que estas lenguas comparten con el ruso. El instrumento no puede distinguirlos de una palabra heredada — ambos son «muchas lenguas diciendo lo mismo»— y de eso trata §5.5. Puestos en fila así, son la mejor ilustración disponible de que un enriquecimiento alto no dice nada sobre la edad de una palabra.

k·r|bark comparte esqueleto con k·r|SKIN (OF FRUIT), y son la misma palabra usada para dos cosas.

n·l|ARROW · binario · 16 lenguas · enriquecimiento 199×

Erzya Mordvin nal	Estonian nool	Finnish nuoli	Hungarian niil
Inari Sami puole	Khanty nʉul	Kildin Sami pu:l:	Livonian nu:oʌ
Mansi nʉol	Moksha nal	North Karelian nɔli	Northern Saami puolla
Skolt Sami puæl:e	Southern Sami puʌɛ	Udmurt nʉəl	Veps nɔl

k·r|bark · binario · 16 lenguas · enriquecimiento 157×

Erzya Mordvin kerʲ	Estonian ko:r	Finnish kuoři	Inari Sami kor:e
Ingrian ko:ri	Khanty ka:r	Kildin Sami ke:r:	Livonian ku:orʲ
Mansi kær:	Moksha ker	North Karelian kɔri	Olonets Karelian kɔri
Skolt Sami kær:e	Southern Sami kær:ɛ	Udmurt kur	Veps korʲ

l·n·t | RIBBON · ternario · 14 lenguas · enriquecimiento 477×

Erzya Mordvin l̥ienta	Estonian l̥iint	Forest Enets ʌent	Kildin Sami ʌa:ŋ:t
Komi ʌenta	Livonian li:nta	Mansi ʌe:nta	Mari ʌe:ntʲe
Moksha l̥ienta	Nenets ʌe:nta	Olonets Karelian lentʊ	Selkup ʌenta
Skolt Sami l̥ient:e	Udmurt l̥ienta		

s·l·m | KNOT · ternario · 13 lenguas · enriquecimiento 339×

Erzya Mordvin s̥ʲulmo	Estonian səlm	Finnish solmu	Inari Sami t̥ʲuolme
Kildin Sami t̥ʲu:l:m	Livonian suoʌm	Lule Sami t̥ʲʊu̯lm:a	Moksha s̥ʲulma
North Karelian sɔlmʊ	Olonets Karelian sɔlmi	Skolt Sami t̥ʲuəlm:e	Southern Sami t̥ʲuʌlme
Veps sɔl̥im			

k·l | LANGUAGE · binario · 13 lenguas · enriquecimiento 121×

Erzya Mordvin kelʲ	Estonian keel	Finnish kieli	Inari Sami kiele
Kildin Sami ki:l̥	Livonian kɛ:ʌ	Moksha kalʲ	North Karelian kieli
Northern Saami kiel:a	Olonets Karelian kieli	Southern Sami kiele	Udmurt k̥il
Veps kɛlʲ			

k·n·g | BOOK · ternario · 12 lenguas · enriquecimiento 1154×

Erzya Mordvin kinʲiga	Forest Enets kniga	Kildin Sami knīga	Komi kniga
Komi-Permyak kniga	Mari knʲigaʰ	Moksha kniga	Nenets kniga
Nganasan knige	Olonets Karelian kni:gʊ	Selkup kniga	Udmurt knʲiga

l·m·p | TORCH OR LAMP · ternario · 12 lenguas · enriquecimiento 2964×

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

s·m·k | BAG · ternario · 10 lenguas · enriquecimiento 933×

Hill Mari sumkə	Kildin Sami sum̥ʲ:k	Komi sumka	Komi-Permyak sumka
Mansi sumka	Mari sumkaʰ	Olonets Karelian sʊmkʊ	Selkup sumka
Udmurt sumka	Veps sʊmk		

r·s·t | CROSS · ternario · 10 lenguas · enriquecimiento 855×

Estonian rist	Finnish risti	Inari Sami riste	Kildin Sami r̥isst
Livonian rift	North Karelian risti	Northern Saami risto	Olonets Karelian ristu
Skolt Sami rist:e	Veps rist		

Cualidades

k·l·m | cold y n·r | YOUNG. El primero comparte esqueleto con k·l·m | three: dos palabras sin relación bajo un mismo código, y con eso van cuatro pares así en el anexo — k·l, t·l, n·l y k·l·m.

La insistencia es el punto. En una familia con el 59 % de colisión, dos o tres consonantes no bastan para separar dos palabras, y el lector tiene que llevarlo puesto en cada bloque.

n·r | YOUNG · binario · 12 lenguas · enriquecimiento 189×

Estonian noor	Finnish nuori	Inari Sami nuore	Kildin Sami nu:r:
Livonian nu:orʲ	Lule Sami nuõr:a	North Karelian nuõri	Northern Saami nuoõrã
Olonets Karelian nuõri	Skolt Sami nuær:e	Southern Sami nuλre	Veps nɔrʲ