Códigos consonánticos en 207 lenguas indoeuropeas

Un catálogo construido a ciegas

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

Resumen

Presentamos un catálogo de 10.116 códigos consonánticos que agrupan 94.152 formas léxicas de 207 lenguas indoeuropeas. Un código es un esqueleto consonántico compartido por palabras del mismo concepto en lenguas distintas.

Es la quinta familia de la serie y la primera con una bibliografía comparatista densa. El catálogo se ha construido sin consultarla: sin reconstrucción, sin listas de cognados, sin etimología, y sin comparar ningún resultado con dos siglos de método comparativo. Ésa es la decisión de diseño principal de este trabajo, y §1.1 explica por qué se toma y qué se gana con ella.

Las cifras: precisión 8,73 % frente a un azar del 0,084 %, es decir 104×, el segundo valor de la serie después del turco. La colisión es del 52,5 % y la aridad media de 3,35, la más alta empatada con el pama-ñungano. El catálogo cubre 1.946 conceptos, más del doble que cualquier familia anterior.

Cuatro resultados. `n·m|name`, con 103 lenguas, es el código de mayor alcance de todo el corpus trabajado. Los 61 pares de ramas dan las concentraciones de alternancia más marcadas de la serie. La geografía es monótona en los seis cubos pese a que la familia tiene lenguas trasplantadas a otros continentes. Y el descubrimiento automático de afijos vuelve a empeorar la ventaja sobre el azar — cuarta familia consecutiva.

1. Qué se busca, y por qué el indoeuropeo a ciegas

La lingüística histórica dispone de un instrumento poderoso y caro: el método comparativo. Se ha aplicado con una desigualdad enorme, y el indoeuropeo es el extremo de esa desigualdad — la familia con la reconstrucción más densa, la cognación experta más revisada y la bibliografía más profunda que existe. Cualquier estudio que se apoye en esa capa medirá, entre otras cosas, la densidad de la bibliografía.

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. Las cuatro familias anteriores —turco, urálico, austronesio, pama-ñungano— se eligieron en parte porque su bibliografía comparatista es menos densa, de modo que renunciar a ella costaba poco.

Aquí cuesta todo. Y por eso se hace.

1.1 El compromiso: no mirar la respuesta

Este catálogo se ha construido sin consultar la reconstrucción indoeuropea. En concreto:

- No se usa ninguna protoforma, ni para construir el catálogo ni para leerlo.

- No se usa ninguna lista de cognados. La fuente `iecor`, que aporta 25.731 formas a este corpus, trae juicios de cognación expertos y no se han leído: de ella se toman las formas y las glosas, como de cualquier otra fuente.
- No se cita ninguna ley fonética, ni por su nombre ni por su contenido.
- No se compara ningún resultado con la clasificación establecida de la familia más allá de las etiquetas de rama que el corpus ya trae como metadato, y que se usan igual que en las cuatro familias anteriores: para leer un catálogo ya cerrado, nunca para construirlo.
- Se excluye del catálogo el único lecto reconstruido que el corpus contiene (§3.1).

Por qué. Un instrumento que afirma poco es útil justamente donde no hay bibliografía. Pero para saber cuánto vale, hay que aplicarlo donde sí la hay — y la única forma de que la prueba signifique algo es no mirar la respuesta mientras se hace. En las otras cuatro familias no miramos porque no había; aquí no miramos habiendo, y eso es lo que convierte el ejercicio en una prueba.

La comparación queda pendiente, a propósito. Lo que este paper publica es el catálogo y sus medidas. Contrastarlo con la lingüística histórica indoeuropea es un trabajo distinto, posterior, y que puede hacer cualquiera con estos ficheros delante. No lo hacemos nosotros y no lo hacemos ahora.

Un aviso al lector que conoce la familia. Varias tablas de §5 —en particular §5.1— muestran alternancias entre segmentos concentradas en cortes concretos. Quien conozca la fonología histórica indoeuropea les pondrá nombre de inmediato. Nosotros no se lo ponemos, y eso es el diseño y no un descuido. Lo que la tabla dice es dónde se concentra una alternancia en este catálogo; lo que esa concentración signifique es exactamente la pregunta que dejamos abierta.

1.2 Lo que este trabajo no intenta

No reconstruimos protoformas. No proponemos ninguna forma ancestral, ni siquiera implícitamente.

No afirmamos parentesco. Que dos palabras compartan código no dice que desciendan de una común.

No distinguimos herencia de préstamo. Sin protoformas es imposible. Un grupo de palabras con el mismo código puede ser herencia, préstamo, convergencia o coincidencia. El catálogo no elige — y en esta familia, con dos milenios de contacto documentado entre sus ramas, ese límite pesa más que en ninguna otra (§6).

No atribuimos significado a un código. Escribimos `n·m|name` y nunca `n·m = name`.

Y no comparamos el indoeuropeo con otras familias: las comparaciones de §5.6 son sobre el comportamiento del instrumento, no sobre las lenguas.

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 aquí

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 indoeuropeo usa 35 símbolos canónicos, de los cuales 31 funden algo: es la familia con el inventario canónico más grande de la serie, y también la que más fusiona. La tabla completa está en `runs/normalizacion.md` y se deriva alineando cada forma con su esqueleto, no se escribe a mano.

Los 35 símbolos, ordenados por frecuencia:

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

Dos fusiones hay que tener presentes al leer §5, porque son las que más borran:

- **s** reúne todas las sibilantes y africadas — ʃ, ʒ, ts, tʃ, dʒ, ʂ, ʐ y una docena más, sobre 149.209 posiciones. Es el símbolo más frecuente del inventario y el que más funde.
- **n** reúne todas las nasales salvo la labial — ŋ, ɲ, nʲ, ɳ y variantes.

Cualquier alternancia que dependa de distinguir dentro de esos dos grupos es invisible para este instrumento, y conviene decirlo antes de que el lector se pregunte por qué no aparece.

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.

2.3 El coste del instrumento: la colisión

El 52,5 % de las formas indoeuropeas comparten esqueleto, dentro de su propia lengua, con una forma de otro concepto. Es la medida limpia de lo que el esqueleto no distingue. En la serie queda en medio: por debajo del turco y el urálico (59 %), por encima del austronesio (46 %) y bastante por encima del pama-ñungano (34 %).

Que sea del 52,5 % con una aridad de 3,35 —la más alta de la serie, empatada— dice algo del inventario: la fusión sibilante de §2.1 se cobra buena parte de lo que la longitud de palabra debería haber ganado.

El índice inverso del anexo lo enseña. `m·r` aparece con «mar», «morir», «serpiente» y «hormiga»; `m·s` con «carne», «mosca», «hombre» y «ratón»; `s·t·r` con «estrella», «cuatro» y «viejo». Cuatro conceptos sin relación aparente bajo dos consonantes es lo que el instrumento cuesta, y hay que llevarlo puesto.

2.4 Aridad

Aridad	Códigos	
1 (unario)	705	7,0 %
2 (binario)	3.645	36,1 %
3 (ternario)	3.800	37,6 %
4 o más	1.947	19,3 %

Las cifras que este paper defiende se restringen a los 7.454 binarios y ternarios.

Esto importa aquí más que en ningún paper anterior, porque cuatro de los quince códigos de mayor alcance son unarios: `d|two` con 79 lenguas, `n|one` con 71, `t|THOU` con 53, `n|NOT` con 43. Una sola consonante no es evidencia de nada — es la clase de coincidencia que un inventario pequeño produce sola — y por eso quedan fuera. Se listan en §4.2 marcados, no escondidos.

3. Método

3.1 Datos

Corpus Integrativo, subconjunto indoeuropeo: 207 lenguas, 319.149 formas con transcripción, concepto asignado y esqueleto. Es la familia más grande de la serie por conceptos (2.778) y la segunda por formas.

Seis lectos excluidos, todos declarados en `ci/decisiones.py` y por dos criterios:

1. Un lecto reconstruido, `Proto-Italic`. Un estudio que se hace sin reconstrucción no puede tener una reconstrucción como dato. La regla se declara aunque esa entrada llegara vacía de todos modos: se excluye por construcción, no por suerte.
2. Cinco registros duplicados del mismo doculecto — mismo glottocode y solapamiento alto de formas escritas, es decir, el corpus tiene el mismo lecto dos veces desde fuentes distintas. Se conserva el registro con más formas: `Old Norse` sobre `Old Icelandic` (96 %), `Lower Sorbian` sobre `Sorbian: Lower` (87 %), `Ossetic` sobre `Ossetic: Iron` (70 %), `Breton` sobre `Breton_ST` (61 %), `Bihari` sobre `Maithili` (61 %).

Y tres exclusiones que la fase 0 propuso y NO se aplican, porque conviene decir también lo que se descarta descartar:

- Nadie se excluye por proximidad. El umbral del 80 % fijado en el paper turco no lo cruza ningún par de glottocodes distintos: el más alto es `Greek: Ancient` / `Greek: New Testament` al 79 %, seguido de las dos normas noruegas al 73 %. Bajar el umbral aquí sería usar dos criterios entre familias.
- Nadie se excluye por marcas de tono. La fase 0 señaló esloveno (90 % de sus formas), lituano (14 %), sueco (3 %) e inglés (0 %). Son marcas de acento tonal, no de tono léxico, y sobre todo no llegan al esqueleto: se comprobó que en toda la familia hay *cero* esqueletos con dígito o marca suprasegmental — `gla˩:d` da `g·l·d`. El chequeo estaba calibrado para familias tonales y aquí sobreavisa.
- Nadie se excluye por estar lejos. Cinco lenguas están a más de 8.000 km del centro de la familia (bislama, tok pisin, criollo jamaicano, saramacano, negerhollands). Excluirlas exigiría un juicio sobre su modo de formación, que es exactamente el tipo de juicio que este trabajo no hace. Se conservan, y §5.5 muestra el análisis geográfico con y sin ellas.

Una cuarta parte de las formas no llega al catálogo, y hay que decirlo fuerte. De 428.150 formas con concepto, 322.827 obtienen esqueleto y 105.323 no (24,6 %). La causa no es la calidad de las palabras sino un pase de segmentación incompleto: son formas perfectamente legibles —`topur`, `caill`, `bēl` en irlandés antiguo— cuyo campo de segmentos está vacío.

Y no es una pérdida repartida: es una fuente. Contando por la fuente real de cada forma —el prefijo de su identificador, no la etiqueta de la lengua, que es de dónde salió una versión anterior de esta tabla y era la columna equivocada—:

Fuente	Formas con concepto	Sin esqueleto	
`ids`	85.559	85.559	100 %
`kaikki`	147.921	18.353	12,4 %
`iec`	25.731	1.036	4,0 %
`lexibank`	128.289	310	0,2 %

Fuente	Formas con concepto	Sin esqueleto	
nel	40.650	65	0,2 %

IDS no pierde mucho: no produce nada. Ninguna de sus 85.559 formas con glosa inglesa tiene fila de esqueleto — no vacía, inexistente—, y esas 85.559 son el 81 % de toda la pérdida de la familia. Las otras cuatro fuentes juntas pierden el 5,7 % de lo suyo.

Eso reescribe el efecto sobre la cobertura, y a mejor: las lenguas que desaparecen enteras son las que solo existen en IDS. Sánscrito (2.684 formas) e inglés medio (1.995) están únicamente ahí, y por eso se pierden del todo. El alto alemán antiguo pierde sus 4.945 formas de IDS pero conserva 1.423 de 1.426 de `lexibank`; el irlandés antiguo pierde 1.353 y conserva sus 172 de `iec`. El hitita queda en 19 de 1.400 y el griego micénico en 4 de 136, éste por `kaikki` y no por IDS.

Es decir: el sesgo hacia lo antiguo es real, tiene una causa localizada y no es una propiedad del método ni del léxico.

Y es una decisión declarada, no una avería pendiente. IDS trae sus formas en ortografía y su distribución no incluye perfiles de conversión a IPA —el estándar CLDF para eso, `etc/orthography/`, uno por lengua—, así que segmentarlas exigiría escribir 274 perfiles. Hacerlo mal no pierde la forma: fabrica un esqueleto falso que entra al catálogo con aspecto de dato. Se decidió (2026-09-04) no trabajar esas lenguas, y la decisión está escrita con sus cifras en `docs/DECISION-IDS.md`, declarada en `ci/decisiones.py` y vigilada en cada corrida de la suite de QA. Este catálogo describe peor las lenguas cuyo único testimonio es IDS, y §6 lo retoma.

Siete macroáreas, por trasplante colonial (§5.5).

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. El agrupamiento real cubre el 43,9 % de las formas; el nulo, el 11,8 %.

El azar de la precisión (§4.1) no se muestrea: se calcula exacto, en forma cerrada sobre el reparto de conceptos y lenguas. Los papers anteriores lo estimaban por sorteo y publicaban intervalos que medían el muestreo propio; la corrección está documentada en `docs/ARQUITECTURA.md`.

3.4 Una sola población para todo el paper

Los siete informes de `runs/` que la declaran imprimen la misma línea:

```
POBLACIÓN · Indo-European: 207 lenguas · 319.149 formas ·
            2.778 conceptos · 12 ramas · despiece=sí(0) ·
            excluidos=6 · imposibles=34 · umbral=3 lenguas
```

Es una comprobación mecánica (`analysis/tools/verifica_paper.py`, chequeo 7).

4. Resultados

4.1 El catálogo

10.116 códigos sobre 1.946 de los 2.778 conceptos, agrupando 94.152 formas.

Magnitud	Valor	Sobre qué población
Formas dentro de un código	29,5 % (94.152 de 319.149)	umbral de 3 lenguas, el del catálogo
Formas en algún grupo	43,9 % (nulo: 11,8 %)	umbral de 2 lenguas, test de agrupamiento
Precisión	8,73 % (azar: 0,084 %)	pares de formas de lenguas distintas
Veces sobre el azar	104× [104–105; el azar se calcula exacto]	ídem
Binarios y ternarios	7.454 (73,7 %)	de los 10.116 códigos

Cada fila lleva su población porque son tres distintas.

El 104× es el segundo de la serie, y §5.6 insiste en lo que el paper turco estableció: es un cociente, y su denominador depende de la familia. El azar indoeuropeo (0,084 %) es el más bajo de las cinco, y lo es por una razón de corpus y no de lenguas — 2.778 conceptos repartidos entre las formas hacen que dos formas al azar coincidan en concepto muy rara vez. Con más conceptos, el denominador baja y el cociente sube sin que nada haya mejorado.

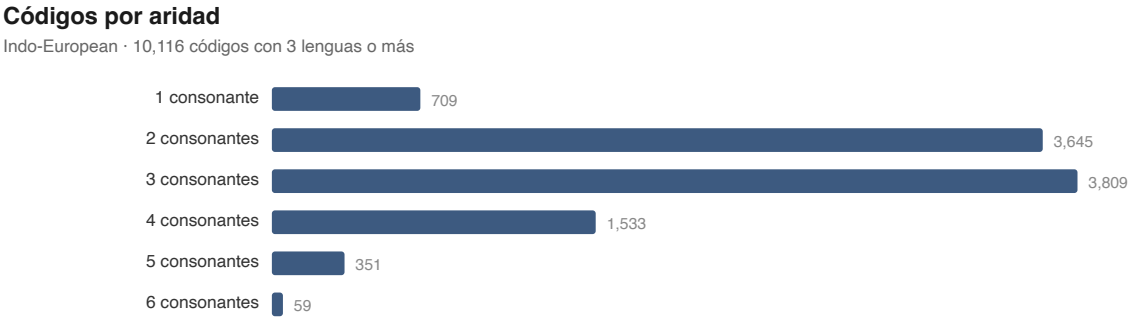

Figure 1: Reparto de los códigos por aridad.

4.2 Los códigos de mayor alcance

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

Las filas en cursiva son unarias y no cuentan para ninguna cifra de este paper (§2.4): un código de una sola consonante reúne lenguas por coincidencia con demasiada facilidad. Se muestran porque esconderlas sería peor, y porque su presencia en la cabecera de la lista es en sí un dato sobre el instrumento.

`n·m|name`, con 103 lenguas, es el código de mayor alcance de todo el corpus trabajado — por delante del `m·r|hand` pama-ñungano (69) y de cualquier código austronesio.

La columna de ramas cuenta en cuántas ramas aparece cada código; no afirma que ese reparto tenga una causa histórica, que es una cuestión distinta y no de este método.

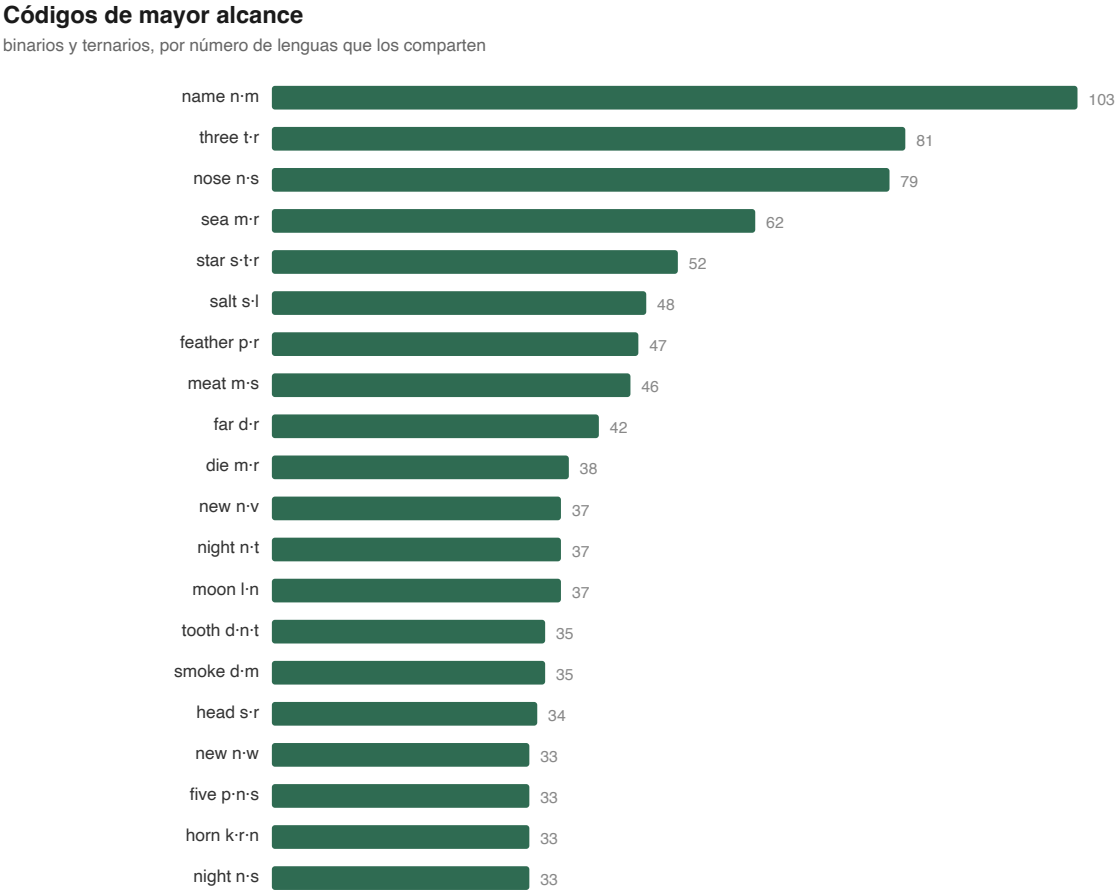

Figure 2: Los veinte códigos binarios y ternarios de mayor alcance.

5. Lo que esta familia enseña

5.1 Sesenta y un pares de ramas

El paper turco estableció que la matriz de sustituciones hay que calcularla por par de ramas, porque agregada sobre la familia diluye lo que distingue a los grupos. El turco pudo probarlo sobre un corte, el urálico sobre treinta y seis, el pama-ñungano sobre noventa y seis. Aquí hay sesenta y uno, y son los de muestras más grandes de toda la serie.

Un corte, el germánico ↔ románico:

Segmento	Principal	Casos	Cuota al cruzar	Cuota sin cruzar	Diagnóstico
h	k	5.799 de 12.343	47 %	8 %	6,2×
k	h	5.799 de 22.418	26 %	3 %	7,8×
θ	t	1.050 de 2.597	40 %	20 %	2,0×
f	p	4.104 de 12.001	34 %	15 %	2,3×
p	f	4.104 de 13.160	31 %	13 %	2,5×
m	l	5.627 de 15.580	36 %	6 %	5,9×

Y otro, el indoiraniano ↔ románico:

Segmento	Principal	Casos	Cuota al cruzar	Cuota sin cruzar	Diagnóstico
θ	s	586 de 877	67 %	29 %	2,3×
l	r	8.747 de 21.622	40 %	17 %	2,4×
t̪	s	921 de 2.733	34 %	8 %	4,3×
q	k	194 de 484	40 %	20 %	2,0×

La última columna compara cuánto más concentrada está la alternancia al cruzar el corte que dentro de una rama.

Qué es y qué no es. 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. Y —§1.1— no le ponemos nombre. Quien conozca la fonología histórica de esta familia verá en estas dos tablas cosas que nosotros no decimos; decirlas sería consultar la respuesta, que es justamente lo que este trabajo se ha comprometido a no hacer. Los ficheros están publicados.

Tres avisos. Las tablas son dos cortes de sesenta y uno y están recortadas a las filas de mayor cuota; el resto está en `runs/fase6-por-rama.md`. Los «casos» son pares de formas, no observaciones independientes: la cifra señala dónde mirar, no es un test. Y §2.1 recuerda que toda alternancia interna a las sibilantes o a las nasales es invisible aquí por construcción.

5.2 Códigos hermanos

La búsqueda de pares de códigos del mismo concepto que difieren en una sola posición (`analysis/fases/codigos_hermanos.py`) da aquí 3.647 pares binarios y ternarios, con diferencia el número más alto de la serie.

	Indoeuropeo	Pama-ñungano	Urálico	Turco
Pares	3.647	653	153	357
Solapan en lengua	827 (23 %)	35 (5 %)	39 (25 %)	140 (39 %)

El 77 % son disjuntos: ninguna lengua aparece en los dos códigos, así que cada par reparte las lenguas en dos grupos sin excepción. De los 827 que solapan, el clasificador separa 425 como «otra forma» —la lengua tiene dos palabras para el concepto, o la misma raíz con distinta flexión, y ninguna de las dos cosas es un error— y 402 como artefacto de transcripción, dos fuentes escribiendo lo mismo de dos maneras.

El caso mayor es `name`: `n·m` con 103 lenguas frente a `n·n` con 10 — y las diez son en su mayoría de una rama distinta de las ciento tres. El catálogo pone las dos mitades en la mesa, separadas, sin saber que lo son.

5.3 El despiece automático empeora el resultado, por cuarta vez

Paso	En grupo	Nulo	Ventaja	κ	Códigos de 4+ consonantes
Palabra entera	43,9 %	11,8 %	32,1	0,364	38,8 %
+ reduplicación	44,4 %	—	—	—	38,0 %
+ afijos descubiertos	58,9 %	43,7 %	15,2	0,270	7,7 %

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 $\Delta\kappa = -0{,}094$. 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.)

Quince puntos de señal, treinta y dos de ruido. La ventaja sobre el azar cae a menos de la mitad.

Es la cuarta familia consecutiva en que ocurre —turco, urálico, pama-ñungano y ésta— y las cuatro son de morfología muy distinta. A estas alturas el hallazgo ya no es de ninguna familia: descubrir afijos por parecido de forma infla el nulo más que la señal, porque el mismo recorte que junta las palabras emparentadas junta también las que no lo están.

Aquí el descubridor encuentra 2.029 prefijos y 1.778 sufijos repartidos en 198 lenguas. Que la cifra de prefijos supere a la de sufijos en una familia de morfología mayoritariamente sufijadora vuelve a decir lo mismo que dijo en pama-ñungano: el descubridor está encontrando el inventario consonántico, no la morfología.

El catálogo publicado no usa despiece automático.

5.4 El cierre de sustituciones no produce nada, y ya van tres

Cerrar transitivamente las sustituciones que aguantan en la mitad de validación dio tres clases limpias en austronesio y tres en pama-ñungano. En turco y urálico no dio ninguna. Aquí tampoco: a umbral $\geq 2\times$ no queda ningún grupo publicable, y este anexo se publica sin columna de código emergente.

Con cinco familias medidas, el reparto es tres a dos y la explicación candidata es la de §5.4 del paper pama-ñungano: allí el cierre funcionó porque la normalización conservaba la dimensión sobre la que se organiza el inventario de la familia —el punto de articulación, con cinco series coronales distintas—. Aquí §2.1 hace lo contrario: funde todas las sibilantes en `s` y casi todas las nasales en `n`, que son precisamente las dimensiones más pobladas del inventario indoeuropeo. Donde la normalización borra la dimensión que la familia usa, el cierre no encuentra nada.

Eso no es una excusa: es una predicción comprobable. Una normalización que conservara las sibilantes debería hacer aparecer clases aquí, y podría empeorar otras cifras. No lo hemos probado.

5.5 La geografía, con lenguas trasplantadas dentro

Códigos compartidos	Pares	Distancia media	÷ azar
1	4.680	3.649 km	1,09

Códigos compartidos	Pares	Distancia media	÷ azar
2	2.823	3.433 km	1,02
3–5	3.024	3.384 km	1,01
6–10	1.293	3.157 km	0,94
11–20	1.041	2.497 km	0,74
21+	1.893	1.637 km	0,49
dos lenguas al azar	40.000	3.355 km	1,00

Monótona en los seis cubos, y ninguno baja de mil pares.

Y aquí hay una comprobación que las otras familias no necesitaban. Esta familia tiene lenguas a doce mil kilómetros de su centro por trasplante colonial —un hecho sobre barcos, no sobre diversificación— y esas lenguas comparten muchos códigos con las de origen mientras están lejísimos, lo que debería aplanar el gradiente. De las 207 lenguas del catálogo, 204 tienen coordenadas y 198 de ellas están en Eurasia; repitiendo el cálculo solo con esas 198:

Códigos compartidos	÷ azar (todas)	÷ azar (solo Eurasia)
1	1,09	1,12
2	1,02	1,05
3–5	1,01	1,00
6–10	0,94	0,86
11–20	0,74	0,64
21+	0,49	0,42

El gradiente se hace más pronunciado al quitarlas, como cabía esperar, pero no cambia de forma. Es decir: las trasplantadas aplanan el resultado sin producirlo. Se publica la tabla con todas las lenguas porque excluirlas exigiría un juicio sobre su modo de formación que este trabajo no hace.

5.6 Cinco familias

Con las cinco medidas por el mismo código y la misma definición (`ci/metricas.py`), y con el azar exacto:

Familia	Lgs	Aridad	Colisión	Precisión	Azar	× azar	Conceptos
Turco	32	2,85	59 %	25,52 %	0,118 %	216×	1.828
Indoeuropeo	207	3,35	52 %	8,73 %	0,084 %	104×	2.778
Urálico	27	3,27	59 %	6,50 %	0,083 %	78×	1.843
Austronesio	582	2,78	46 %	13,45 %	0,363 %	37×	1.823
Pama-ñungano	154	3,35	34 %	5,87 %	0,298 %	20×	1.160

Se añade la columna de conceptos porque con cinco familias ya se ve lo que con dos no se veía: el azar cae cuando el corpus tiene más conceptos, y el cociente sube por eso. Indoeuropeo y urálico tienen azares casi idénticos (0,084 % y 0,083 %) con precisiones que difieren en 1,3×, y el cociente los separa en 104× frente a 78×. Ordenar familias por esa columna sigue siendo un error, y ahora se ve mejor por qué.

La aridad predice la colisión, con una excepción que se explica. Austronesio 2,78 → 46 %, turco 2,85 → 59 %, urálico 3,27 → 59 %, pama-ñungano 3,35 → 34 %. El indoeuropeo tiene la aridad más alta empatada y una colisión del 52 %, muy por encima del pama-ñungano con la misma longitud de palabra. La diferencia está en §2.1: aquí la normalización funde sibilantes y nasales, allí conservaba cinco series coronales. La longitud de palabra da margen; la normalización decide cuánto de ese margen se aprovecha.

6. Límites

El límite mayor es el corpus, no el método — y tiene un solo culpable. Una cuarta parte de las formas no llega al catálogo, y el 81 % de esa pérdida es una sola fuente: IDS aporta 85.559 formas con glosa inglesa y no produce ni un esqueleto (§3.1). Las otras cuatro fuentes juntas pierden el 5,7 % de lo suyo.

Eso golpea a las lenguas antiguas porque son las que más dependen de esa fuente: sánscrito e inglés medio existen solo en IDS y se pierden enteros; el alto alemán antiguo y el irlandés antiguo pierden su parte de IDS y conservan el resto. En una familia cuyo interés está en buena parte en su profundidad temporal, eso no es un detalle. Este catálogo describe peor las lenguas cuyo único testimonio es IDS, ninguna cifra de §4 o §5 lo corrige, y es reparable: no es una propiedad del método sino un pase de ingesta que falta.

La colisión es del 52,5 %, y §2.1 explica por qué es alta pese a la aridad: la fusión sibilante.

Herencia y préstamo no se distinguen, y aquí eso pesa el doble. Estas lenguas llevan dos milenios en contacto documentado unas con otras y con lenguas de fuera, y comparten además una capa culta común. Un código de alcance alto puede ser cualquiera de esas cosas.

No hay reflejos cero. El procedimiento de §5.1 compara esqueletos de la misma longitud, así que no puede ver ninguna alternancia que incluya la pérdida de un segmento — y en esta familia ese es un fenómeno central.

El 19,3 % de los códigos tiene cuatro o más consonantes, y §5.3 muestra que el despiece automático no es la salida. La segmentación morfológica sigue sin resolver.

Los unarios encabezan la lista de alcance. Cuatro de los quince códigos de mayor alcance tienen una sola consonante. Quedan fuera de las cifras defendidas, pero su presencia dice que en una familia grande el instrumento produce coincidencia con facilidad.

No calibramos contra cognación experta, y aquí es una decisión y no una carencia. La fuente `iecor` la trae, en el mismo corpus, sin leer. Calibrar contra ella es exactamente el trabajo que §1.1 aplaza — pero quien quiera hacerlo tiene los dos ficheros.

Una sola familia, otra vez. Todo lo que este paper dice sobre el significado de las cifras se apoya en cinco puntos.

7. Conclusión

El catálogo indoeuropeo agrupa el 29,5 % de las formas en 10.116 códigos sobre 1.946 conceptos, con una precisión 104 veces superior al azar. Es el catálogo más ancho de la serie en conceptos y el que contiene el código de mayor alcance de todo el corpus trabajado, `n·m|name` en 103 lenguas.

Lo que este trabajo aporta no es ese catálogo, y tampoco es un resultado sobre el indoeuropeo. Es un catálogo construido sin mirar la respuesta en la única familia donde la respuesta está escrita. Todo lo que hay en estas páginas se puede contrastar ahora con dos siglos de método comparativo, y ese contraste dirá algo que ninguna de las cuatro familias anteriores podía decir: no si el instrumento acierta —eso ya se sabe medir— sino qué clase de cosa encuentra y qué clase de cosa se le escapa, con una vara de medir independiente.

Mientras tanto, dos resultados que no dependen de esa comparación. El despiece automático empeora la ventaja sobre el azar por cuarta vez consecutiva, en cuatro morfologías distintas; ya es una propiedad del procedimiento. Y el cierre de sustituciones no produce nada por tercera vez de cinco, con una explicación candidata que el propio catálogo sugiere y que es comprobable: funciona donde la normalización conserva la dimensión sobre la que se organiza el inventario de la familia, y aquí la borra.

Un instrumento que afirma poco puede aplicarse donde no hay bibliografía. Aplicarlo donde la hay, y no leerla, es la única manera de averiguar cuánto de lo que no encuentra es culpa suya.

Referencias

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

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

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

Heggarty, P., Anderson, C., Scarborough, M. *et al.* (2023). Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. *Science* 381. [Fuente `iecor` de las formas; sus juicios de cognación no se han usado — §1.1.]

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

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

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

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

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/indo-european/output/catalogo-indo-european.tsv` — 10.116 filas
Detalle por forma	`families/indo-european/output/catalogo-indo-european-lenguas.tsv` — 65.503 filas
Higiene (§3.1)	`analysis/fases/fase0_higiene.py` → `runs/fase0-higiene.md`
Exclusiones declaradas (§3.1)	`ci/decisiones.py`

Carga, población y avisos	`ci/familia.py`
Cifras de cabecera (§4.1, §5.6)	`ci/metricas.py` — azar exacto
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`
Cierre de sustituciones (§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 hermanos (§5.2)	`analysis/fases/codigos_hermanos.py` → `runs/codigos-hermanos.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` + `analysis/fases/empalma_anexo.py`
Auditoría anti-mezcla de familias	`analysis/tools/audita_familias.py`

Fuentes de datos: Lexibank (List, Forkel, Greenhill et al. 2022, CC-BY-4.0), IDS (Key & Comrie, CC-BY-4.0), IE-CoR (Heggarty et al. 2023, CC-BY-4.0 — solo las formas), Wiktionary/kaikki (CC-BY-SA-3.0), Concepticon y Glottolog (CC-BY-4.0).

Anexo A · Cien códigos comentados

El catálogo tiene 10.116 códigos y los datos completos están en `families/indo-european/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. El cierre transitivo de las sustituciones no produce clases limpias en esta familia (§5.4), y trasplantar las del austronesio o las del pama-ñungano sería cómodo y falso: el valor de una clase es relativo a la familia que la produjo.

Y lo que gobierna la lectura entera: este anexo describe qué hay en cada bloque —cuántas lenguas, cómo se reparten, qué varía y qué no— y no dice de qué desciende ninguna forma, ni pone nombre a ninguna alternancia. No es prudencia retórica: es la decisión de diseño de §1.1. Quien conozca la familia verá aquí cosas que no decimos, y los ficheros están publicados para que las diga quien quiera.

Una advertencia que hay que llevar puesta: el 52,5 % 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

Esta tabla es más larga que en ninguna familia anterior, y la causa no es la aridad —3,35 consonantes por forma, la más alta— sino la normalización: §2.1 funde todas las sibilantes en `s` y casi todas las nasales en `n`, que son las dos zonas más pobladas del inventario. La longitud de palabra da margen y la normalización se lo gasta.

Tres casos que conviene mirar despacio:

`m·r` aparece con «mar» (62 lenguas), «morir» (38), «serpiente» (22) y «hormiga» (21). `m·s`, con «carne» (46), «mosca» (22), «hombre» (21) y «ratón» (21). Y `s·t·r`, con «estrella» (52), «cuatro» (31) y «viejo» (22). Cuatro conceptos sin relación aparente bajo dos consonantes es exactamente el coste que §2.3 declara, y en esta familia se paga cuatro veces en la primera pantalla de la tabla.

Código	Conceptos con los que aparece, y su enriquecimiento
`m·r`	sea (62, 124×) · die (38, 74×) · snake (22, 42×) · ant (21, 46×)
`m·s`	meat (46, 72×) · fly_n (22, 40×) · mouse (21, 147×) · man (21, 32×)
`s·t·r`	star (52, 120×) · four (31, 75×) · old (22, 55×)
`s·l`	salt (48, 61×) · sun (32, 40×) · year (27, 33×)
`k·r·n`	horn (33, 110×) · root (22, 73×) · meat (20, 66×)
`m·n`	man (31, 70×) · moon (30, 68×) · hand (26, 58×)
`s·n`	woman (28, 29×) · snow (23, 27×) · chest (20, 23×)

Numerales

El bloque más nítido del anexo, y el que mejor enseña lo que un código hace y lo que no.

«Tres» tiene dos códigos —`t·r` con 81 lenguas y `t·r·s` con 22—, «cuatro» tiene tres —`s·t·r` (31), `s·r` (27), `k·t·r` (20)— y «cinco» tiene dos, `p·n·s` (33) y `p·s` (29). En los tres casos los códigos del mismo

concepto difieren en una sola posición o en una consonante presente o ausente, que es la materia prima de §5.2.

Y en los tres casos las formas de cada código se parecen mucho entre sí y poco a las del código vecino. El catálogo separa; lo que separe, no lo dice.

Nótese además lo que no está aquí: «uno» y «dos» encabezan la lista de alcance de todo el catálogo, con 71 y 79 lenguas, y no aparecen en este bloque porque sus códigos son unarios —n y d— y quedan fuera por §2.4. Es el mejor recordatorio disponible de por qué esa regla existe.

t·r|three · binario · 81 lenguas · enriquecimiento 129×

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

p·n·s|five · ternario · 33 lenguas · enriquecimiento 179×

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

s·t·r|four · ternario · 31 lenguas · enriquecimiento 75×

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

p·s|five · binario · 29 lenguas · enriquecimiento 49×

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

s·r|four · binario · 27 lenguas · enriquecimiento 38×

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

t·r·s|three · ternario · 22 lenguas · enriquecimiento 112×

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

k·t·r|four · ternario · 20 lenguas · enriquecimiento 210×

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

Partes del cuerpo

El bloque más numeroso, con `n·s|nose` (79 lenguas) a la cabeza.

Dos conceptos tienen dos códigos cada uno, y los dos casos son distintos entre sí. «Mano» se reparte entre `m·n` (26 lenguas) y `r·k` (24) — dos códigos que no comparten ni una consonante, así que no son hermanos en el sentido de §5.2 sino dos formas sin parecido. «Rodilla» se reparte entre `k·n` (25) y `k·l·n` (23) — que comparten las dos consonantes y difieren en una insertada en medio.

`p·p·k|navel`, con 690×, es el enriquecimiento más alto del bloque y sus veintiuna lenguas son todas de una sola rama: bielorruso, croata, checo, casubio, macedonio, eslavo eclesiástico antiguo, polaco antiguo. Un código de alcance alto puede estar confinado a un grupo; el alcance no dice dónde.

`d·n·t|tooth` (521×) y `j·z·k|tongue` (538×) son los dos ternarios de mayor enriquecimiento de la sección, y en ambos las formas del bloque se parecen entre sí de forma evidente a simple vista.

n·s|nose · binario · 79 lenguas · enriquecimiento 149×

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

d·n·t|tooth · ternario · 35 lenguas · enriquecimiento 521×

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

s·r|head · binario · 34 lenguas · enriquecimiento 44×

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

k·r·n|horn · ternario · 33 lenguas · enriquecimiento 110×

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

k·r|heart · binario · 27 lenguas · enriquecimiento 46×

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

m·n|hand · binario · 26 lenguas · enriquecimiento 58×

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

k·n|knee · binario · 25 lenguas · enriquecimiento 52×

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

b·n|bone · binario · 24 lenguas · enriquecimiento 72×

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

r·k|hand · binario · 24 lenguas · enriquecimiento 70×

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

k·l·n|knee · ternario · 23 lenguas · enriquecimiento 109×

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

k·s|skin · binario · 23 lenguas · enriquecimiento 38×

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

j·z·k|tongue · ternario · 23 lenguas · enriquecimiento 538×

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

h·r·n|horn · ternario · 22 lenguas · enriquecimiento 278×

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

b·k|mouth · binario · 21 lenguas · enriquecimiento 83×

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

p·p·k|navel · ternario · 21 lenguas · enriquecimiento 690×

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

s·n|chest · binario · 20 lenguas · enriquecimiento 23×

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

Parentesco y persona

n·m|name, con 103 lenguas, es el código de mayor alcance de todo el corpus trabajado — por delante del m·r|hand pama-ñungano, que llegaba a 69.

Merece mirarse por su enriquecimiento, 379×, porque explica el mecanismo. n y m son nasales, de las consonantes más frecuentes de la familia, así que la frecuencia base es alta. La cifra no sale de la rareza sino de que ciento tres de las lenguas que atestiguan «nombre» lo dicen con el mismo esqueleto: la cobertura es casi total. No hace falta una secuencia rara para llegar ahí; basta que una palabra esté bien distribuida en un corpus bien muestreado.

«Hombre» aparece con dos códigos, m·n (31) y m·s (21), y m·s es además el código de «carne», «mosca» y «ratón» del índice inverso. m·m|MOTHER da 916× con veinte lenguas, que es el enriquecimiento más alto de esta sección y sobre la secuencia consonántica menos informativa posible.

n·m|name · binario · 103 lenguas · enriquecimiento 379×

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

m·n|man · binario · 31 lenguas · enriquecimiento 70×

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

s·n|woman · binario · 28 lenguas · enriquecimiento 29×

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

m·s|man · binario · 21 lenguas · enriquecimiento 32×

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

m·m|MOTHER · binario · 20 lenguas · enriquecimiento 925×

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

Naturaleza y cielo

La sección más poblada después de las partes del cuerpo, y la que más reparte un mismo concepto entre códigos: «noche» va con `n·t` (38) y `n·s` (33), «estrella» con `s·t·r` (52) y `s·t·l` (25), «humo» con `d·m` (35) y `f·m` (27), «luna» con `l·n` (37) y `m·n` (30), «cielo» con `s·m·n` (26) y `n·b` (23).

Cinco conceptos con dos códigos cada uno en un solo bloque. En «estrella» los dos códigos difieren en una sola posición; en «luna» y «humo» no comparten ninguna consonante. El catálogo no distingue esos dos casos — §5.2 sí, y por eso existe.

Y aquí empieza lo que hay que leer con §5.5 delante. `p·r·l|APRIL` da 1131× con 23 lenguas y `m·r·s|MARCH` da 645× con 22. Son nombres de mes, y son los dos enriquecimientos más altos del

anexo entero. El bloque de abril reúne albanés, armenio, búlgaro, croata, neerlandés y alemán diciendo casi lo mismo. Un enriquecimiento de mil no dice nada sobre la antigüedad de una palabra: dice que muchas lenguas la dicen igual, y eso puede ocurrir por vías muy distintas que este instrumento no separa.

m·r|sea · binario · 62 lenguas · enriquecimiento 124×

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

s·t·r|star · ternario · 52 lenguas · enriquecimiento 120×

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

l·n|moon · binario · 37 lenguas · enriquecimiento 87×

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

n·t|night · binario · 37 lenguas · enriquecimiento 127×

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

d·m|smoke · binario · 35 lenguas · enriquecimiento 131×

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

n·s|night · binario · 33 lenguas · enriquecimiento 62×

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

s·l|sun · binario · 32 lenguas · enriquecimiento 40×

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

d·n|day · binario · 31 lenguas · enriquecimiento 91×

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

m·n|moon · binario · 30 lenguas · enriquecimiento 68×

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

t·r|earth · binario · 27 lenguas · enriquecimiento 41×

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

f·m|smoke · binario · 27 lenguas · enriquecimiento 222×

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

s·l|year · binario · 27 lenguas · enriquecimiento 33×

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

s·m·n|sky · ternario · 26 lenguas · enriquecimiento 178×

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

s·t·l|star · ternario · 25 lenguas · enriquecimiento 150×

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

p·r·l|APRIL · ternario · 23 lenguas · enriquecimiento 1116×

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

n·b|sky · binario · 23 lenguas · enriquecimiento 284×

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

s·n|snow · binario · 23 lenguas · enriquecimiento 27×

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

m·r·s|MARCH · ternario · 22 lenguas · enriquecimiento 645×

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

s·t·n|stone · ternario · 22 lenguas · enriquecimiento 109×

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

v·n·t|wind · ternario · 22 lenguas · enriquecimiento 303×

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

p·s·k|sand · ternario · 21 lenguas · enriquecimiento 152×

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

Animales

p·r|feather (47 lenguas) encabeza, y «pluma» tiene además un segundo código, p·n (20).

m·x|fly_N (300×) y m·s|fly_N (40×) son dos códigos para «mosca» que difieren en una posición, y m·s es el mismo esqueleto que «carne», «hombre» y «ratón». Es decir: en una sola línea del índice inverso conviven un par de hermanos y una colisión de cuatro conceptos. Distinguir una cosa de la otra es mirar las formas, y por eso el anexo las imprime.

m·r|snake (22) y m·r|ant (21) comparten esqueleto con m·r|sea (62) y m·r|die (38). Cuatro conceptos, dos consonantes.

p·r|feather · binario · 47 lenguas · enriquecimiento 90×

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

m·x|fly_N · binario · 27 lenguas · enriquecimiento 300×

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

l·s|louse · binario · 26 lenguas · enriquecimiento 57×

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

r·b|fish · binario · 25 lenguas · enriquecimiento 193×

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

m·s|fly_N · binario · 22 lenguas · enriquecimiento 40×

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

m·r|snake · binario · 22 lenguas · enriquecimiento 42×

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

m·s|MOUSE · binario · 21 lenguas · enriquecimiento 147×

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

m·r|ant · binario · 21 lenguas · enriquecimiento 46×

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

k·n|dog · binario · 20 lenguas · enriquecimiento 40×

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

p·n|feather · binario · 20 lenguas · enriquecimiento 58×

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

k·r·m|worm · ternario · 20 lenguas · enriquecimiento 183×

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

Plantas y alimento

`s·l|salt` (48 lenguas) y `s·l·t|salt` (20) son el mismo concepto con y sin una consonante final; `s·l` es además el código de «sol» (32) y «año» (27).

Y aquí está la segunda mitad de lo que §5.5 explica: `s·k·r|SUGAR` 452×, `s·p|SOUP` 295×, `f·r·t|fruit` 262×, `b·r|BEER` 189×. Junto con los meses del bloque anterior y con `s·k·l|SCHOOL` (712×) del siguiente, son los enriquecimientos más altos de todo el anexo — y son, de forma visible, vocabulario que estas lenguas comparten por vías que no son la transmisión ordinaria. El instrumento no puede distinguirlos de una palabra heredada, porque ambos son «muchas lenguas diciendo lo mismo».

Puestos en fila, son la mejor ilustración que la serie ha producido de que el enriquecimiento mide cobertura, no edad.

`s·l|salt` · binario · 48 lenguas · enriquecimiento 61×

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

`m·s|meat` · binario · 46 lenguas · enriquecimiento 72×

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

`s·p|SOUP` · binario · 30 lenguas · enriquecimiento 294×

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

`f·r·t|fruit` · ternario · 25 lenguas · enriquecimiento 262×

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

`g·r·s|grass` · ternario · 22 lenguas · enriquecimiento 115×

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

`k·r·n|root` · ternario · 22 lenguas · enriquecimiento 73×

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

s·k·r|SUGAR · ternario · 21 lenguas · enriquecimiento 452×

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

l·s·t|leaf · ternario · 21 lenguas · enriquecimiento 142×

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

s·m|seed · binario · 21 lenguas · enriquecimiento 54×

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

k·r·n|meat · ternario · 20 lenguas · enriquecimiento 66×

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

s·l·t|salt · ternario · 20 lenguas · enriquecimiento 126×

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

Verbos de acción y estado

El bloque más pequeño del anexo: cuatro códigos, todos binarios, con alcances entre 20 y 38 lenguas.

Que los verbos rindan tan poco comparado con los sustantivos es constante en las cinco familias de la serie, y la explicación candidata es la de siempre: las listas de las que sale el corpus citan los verbos en formas flexionadas o con afijos que dependen de la fuente, y §5.3 muestra que quitarlos automáticamente empeora el resultado. Parte de lo que aquí falta es morfología que no sabemos separar.

m·r|die · binario · 38 lenguas · enriquecimiento 74×

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

m·l|grind · binario · 22 lenguas · enriquecimiento 68×

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

k·m|come · binario · 20 lenguas · enriquecimiento 93×

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

d·t|give · binario · 20 lenguas · enriquecimiento 105×

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

Objetos y cultura

s·k·l|SCHOOL con 712× es el enriquecimiento más alto de la sección y el tercero del anexo. Le siguen t·n|TONE (MUSIC) (309×), k·r·s|CROSS (268×), k·n|CANOE (236×) y l·n|LINE (194×).

Cinco conceptos de cultura material o institucional, cinco enriquecimientos altísimos, y ninguna sorpresa una vez leído §5.5. k·n es además el código de «rodilla» y «perro».

h·s|house (195×) y k·r|bark (38×) son los dos del bloque que no siguen ese patrón, y k·r comparte esqueleto con «corazón».

h·s|house · binario · 27 lenguas · enriquecimiento 195×

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

k·r·s|CROSS · ternario · 25 lenguas · enriquecimiento 268×

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

l·n|LINE · binario · 25 lenguas · enriquecimiento 195×

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

s·k·l|SCHOOL · ternario · 25 lenguas · enriquecimiento 712×

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

t·n|TONE (MUSIC) · binario · 24 lenguas · enriquecimiento 313×

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

k·r|bark · binario · 21 lenguas · enriquecimiento 38×

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

l·t·r|ALTAR · ternario · 20 lenguas · enriquecimiento 939×

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

Cualidades

La sección donde el reparto de un concepto entre varios códigos llega más lejos.

«Lleno» tiene tres códigos: `p·r` (25 lenguas), `p·l·n` (23) y `f·l` (21). «Nuevo» tiene dos que difieren en una sola posición, `n·v` (37) y `n·w` (33). «Seco» tiene dos, `s·k` (33) y `s·x` (27), que también difieren en una. «Verde» tiene dos sin ninguna consonante en común, `g·r·n` (23) y `z·l·n` (21). Y «bueno» tiene dos, `b·n` (27) y `d·b·r` (21).

Cinco conceptos repartidos, y los cinco de una manera distinta. Es el bloque que mejor enseña por qué §5.2 existe como sección aparte: el catálogo produce estos pares por millares —3.647 en total— y separarlos en «dos formas de escribir lo mismo», «dos palabras distintas» y «la misma raíz con distinta flexión» es un trabajo que el catálogo no hace solo.

`s·t·r|old` (22) comparte esqueleto con «estrella» y «cuatro». Es la cuarta vez que esta advertencia aparece en el anexo, y con 10.097 códigos sobre 1.944 conceptos no será la última.

`d·r|far` · binario · 42 lenguas · enriquecimiento 81×

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

`n·v|new` · binario · 37 lenguas · enriquecimiento 267×

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

`s·k|dry` · binario · 33 lenguas · enriquecimiento 61×

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

`n·w|new` · binario · 33 lenguas · enriquecimiento 355×

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

`k·r·t|short` · ternario · 29 lenguas · enriquecimiento 172×

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

`s·x|dry` · binario · 27 lenguas · enriquecimiento 144×

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

b·n|good · binario · 27 lenguas · enriquecimiento 78×

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

p·r|full · binario · 25 lenguas · enriquecimiento 46×

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

p·l·n|full · ternario · 23 lenguas · enriquecimiento 186×

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

g·r·n|green · ternario · 23 lenguas · enriquecimiento 137×

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

g·r·m|hot · ternario · 23 lenguas · enriquecimiento 282×

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

l·n·g|long · ternario · 23 lenguas · enriquecimiento 350×

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

s·t·r|old · ternario · 22 lenguas · enriquecimiento 55×

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

f·l|full · binario · 21 lenguas · enriquecimiento 56×

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

d·b·r|good · ternario · 21 lenguas · enriquecimiento 391×

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

z·l·n|green · ternario · 21 lenguas · enriquecimiento 560×

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

l·s|smooth · binario · 20 lenguas · enriquecimiento 47×

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

Other

f·g·r|IDOL · ternario · 20 lenguas · enriquecimiento 1838×

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