El catálogo de códigos consonánticos

Un método comparativo sin genealogía, sin cognación y sin reconstrucción

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

Resumen

El método comparativo se ha aplicado con una desigualdad enorme: en el corpus sobre el que trabaja esta serie, los juicios de parentesco existen para un macrosistema de 225. Cualquier estudio que se apoye en la capa histórica del léxico medirá, sin quererlo, la densidad de la bibliografía comparatista.

Se describe un método que renuncia a la reconstrucción —no la mejora ni la sustituye: prescinde de ella— y construye un catálogo de códigos consonánticos con tres cosas y solo tres: el esqueleto consonántico de una forma, el concepto que expresa, y un umbral de tres lenguas distintas. El procedimiento no sabe que las familias tienen ramas; la genealogía entra al final y solo para leer lo que salió.

Se dan el protocolo completo, los nulos con lo que cada uno rompe y conserva, y las reglas de umbral, incluidas las tres correcciones que cinco aplicaciones obligaron a hacer: el nulo se calcula exacto y no se estima —estimarlo publicó un intervalo que medía nuestro propio sorteo durante tres papers—; batir un nulo de barajado autoriza afirmar una correspondencia y nada más, porque leerla como migración exige una protoforma que el método no tiene; y el umbral se fija acotado por dos límites, el techo alcanzable y la tasa de falsos positivos, porque el $2\times$ usado durante cinco papers resultó ser superado por azar en el 64 % de los casos para uno de los estadísticos empleados.

El criterio se pone a prueba en el único sitio donde hay respuesta conocida. Contra cognación de oro indoeuropea, concepto más esqueleto idéntico da 93,3 % de precisión frente a una base del 26,0 %, con una cobertura del 13,6 %: precisión alta y cobertura baja, que es lo que un catálogo debe ser.

Se compara además el coste del instrumento en los cinco macrosistemas aplicados, y aparece que la colisión —la medida de lo que el esqueleto no distingue— no la explica la longitud de la palabra: pama-ñungano e indoeuropeo tienen la misma longitud media y difieren en dieciocho puntos.

1 · El problema, y la apuesta

La lingüística histórica dispone de un instrumento poderoso y caro: el método comparativo. Reconstruye protoformas, establece correspondencias y produce árboles. Funciona, y lleva dos siglos funcionando.

Pero se ha aplicado con una desigualdad enorme. El indoeuropeo tiene reconstrucción densa y cognación revisada por expertos; la mayor parte de las demás familias no tiene ninguna de las dos. En el corpus sobre el que se trabaja aquí, los juicios de parentesco existen para un macrosistema de 225.

De ahí se sigue algo que conviene decir antes que nada: cualquier estudio comparativo 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.

1.1 · La apuesta

Este trabajo pregunta si se puede hacer algo útil renunciando a la reconstrucción. No mejorarla ni sustituirla: prescindir de ella y ver qué queda.

Lo que queda es un catálogo de códigos consonánticos: un objeto más modesto y mucho más barato que hace parte del trabajo de la genealogía sin hacer ninguna de sus afirmaciones. Se construye con tres cosas y solo tres —el esqueleto consonántico de una forma, el concepto que expresa, y un umbral de lenguas distintas— y no sabe que las familias tienen ramas.

1.2 · El orden importa: primero el catálogo, la genealogía al final

Es lo que distingue este método de una comparación al uso, y es fácil de perder de vista.

El catálogo se construye sin genealogía. No entra ninguna clasificación en ramas, ninguna protoforma, ninguna lista de cognados, ninguna hipótesis sobre qué desciende de qué.

La genealogía entra después, y solo para leer lo que salió. Cuando se cruzan los códigos con las etiquetas de rama, se hace sobre un catálogo ya cerrado, y la pregunta no es «¿qué códigos predice la clasificación?» sino la inversa: «¿reencuentra el catálogo, por su cuenta, algo que la clasificación reconoce?». Las protoformas que aparecen en las tablas están para que el lector compruebe qué se recuperó — el método no las vio, no las usó y no las produce.

Ese orden no es una preferencia estilística: es lo que hace que el instrumento sirva allí donde no hay reconstrucción, que es para lo que se diseñó. Un método que necesitara la clasificación para construir su catálogo no podría aplicarse justamente donde hace falta.

1.3 · Lo que no se intenta

El objeto es fácil de sobreinterpretar, así que conviene acotarlo con precisión.

No se reconstruyen protoformas. Ninguna forma ancestral, ni siquiera implícitamente.

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

No se distingue herencia de préstamo. Sin protoformas es imposible, y no se va a fingir lo contrario. Un grupo de palabras con el mismo código puede ser herencia común, préstamo antiguo, préstamo reciente, convergencia o coincidencia. El catálogo no elige.

No se atribuye significado a un código. Se escribe `n·m|name` y nunca `n·m = name`. El código no significa «nombre»: es la forma que «nombre» toma repetidamente.

1.4 · Qué se puede afirmar, entonces

Una sola cosa, y por eso el método es barato: que ciertas correspondencias forma–sentido se repiten entre lenguas más de lo que se repetirían por azar, con el azar medido explícitamente (§4).

Todo lo demás —qué significa esa repetición, si es herencia o contacto, qué la produjo— queda fuera. Es menos de lo que da el método comparativo, y se puede obtener donde el método comparativo no se ha aplicado.

1.5 · Dónde encaja esto

Reducir una palabra a sus consonantes no es nuevo: es el *consonant class matching* de Dolgopolsky (1964) y de Turchin, Peiros y Gell-Mann (2010), evaluado junto a SCA y LexStat en la literatura de com-

paración automática (List 2012, 2014; List, Greenhill y Gray 2017; Jäger 2018). No debe confundirse con ASJP (Brown et al. 2008; Holman et al. 2008), que conserva las vocales y compara con distancia de Levenshtein corregida por azar.

La variante que aquí se usa se diferencia en dos cosas:

No trunca. El *consonant class matching* clásico compara las dos primeras consonantes. Aquí se conserva la secuencia entera, porque la aridad es objeto de estudio y no un parámetro fijado (§2.4): que un macrosistema produzca muchos códigos de cuatro consonantes es un diagnóstico sobre su morfología, y truncando no se vería.

La salida no es una matriz de distancias sino un catálogo consultable por concepto. La diferencia no es de formato: una matriz de distancias responde «¿cuánto se parecen estas dos lenguas?» y un catálogo responde «¿qué forma toma este concepto, repetidamente, en este macrosistema?». La segunda pregunta no promedia sobre el léxico, y por eso cada código se puede inspeccionar, discutir y refutar uno a uno.

Una diferencia que NO se reclama. Sería cómodo decir que aquí no se usan clases de sonido. No es cierto: la normalización canónica del esqueleto es una capa de clases, y por eso cada aplicación publica su tabla entera antes de cualquier resultado (§2.1).

2 · El instrumento

2.1 · El esqueleto consonántico

La secuencia de consonantes de una palabra, con las vocales retiradas —no perdidas: quedan en una columna aparte— y cada consonante escrita con un representante canónico por valor cuálico.

El esqueleto no lo produce este método: es una capa del corpus, calculada con las mismas reglas para los 225 macrosistemas, y su definición, sus decisiones de normalización y su coste están descritos en el paper del corpus. Aquí se toma como entrada.

Lo que sí es de este método es la tabla de normalización que cada macrosistema publica, derivada alineando cada forma con su esqueleto y no escrita a mano. Publicarla es obligatorio, porque lo que la normalización funde deja de ser visible, y el lector tiene derecho a saber qué.

2.2 · Qué es un código

Un esqueleto que aparece, para el mismo concepto, en tres lenguas distintas o más.

El requisito del concepto no es decorativo. Sin él, el esqueleto agrupa por parecido de forma y reúne palabras sin relación; con él, el agrupamiento tiene que acertar dos cosas a la vez, y es esa doble condición la que produce toda la señal que el método puede tener.

El umbral de tres lenguas es una decisión declarada, no un descubrimiento. Con dos, cualquier coincidencia entre vecinas produce un código.

2.3 · El coste del instrumento: la colisión

Un esqueleto no distingue lo que las vocales distinguen. La medida limpia de eso, sin genealogía de por medio, es: qué proporción de formas comparte esqueleto, dentro de su propia lengua, con una forma de otro concepto.

Es alta, y varía mucho por macrosistema (§5). En turco es el 59,4 %, y el ejemplo que mejor lo enseña es `k·l`, que aparece con diez conceptos: *kül* «ceniza», *kel-* «venir», *köl* «lago», *kül-* «reír». Tres se

distinguen por la vocal, que es exactamente lo que el esqueleto no registra.

Pero el cuarto es más interesante, y obliga a un matiz que el método debe llevar escrito: **`kül`** «ceniza» y **`kül-`** «reír» son homófonas ya en turco antiguo. Ni un procedimiento que conservara las vocales las separaría. Parte de la colisión no es un coste del instrumento sino un hecho de la lengua, y confundir las dos cosas sobreestima el precio del método.

2.4 · Aridad

Un código puede tener una consonante, dos, tres o más.

Un unario no sostiene ninguna afirmación por sí solo: hay pocas consonantes y cualquiera coincide por azar. Un código de cuatro o más es motivo para dudar del propio análisis, porque suele significar que no se ha separado la morfología — y en las cinco aplicaciones esa sospecha resultó fundada más de una vez.

Las cifras que un paper de esta serie defiende se restringen a los binarios y ternarios. Es una restricción declarada y cuesta: en pama-ñungano deja fuera 447 códigos de 2.092.

2.5 · Enriquecimiento, y una advertencia sobre él

Un código que aparece en muchas lenguas puede ser notable o trivial según de qué esté hecho: uno formado por las consonantes más frecuentes de la familia aparecerá en todas partes sin decir nada. El enriquecimiento corrige eso:

alcance	número de lenguas distintas en que el código aparece
cobertura	alcance ÷ lenguas que atestiguan ese concepto
frecuencia base	veces que ese esqueleto aparece en el macrosistema ÷ total de esqueletos
enriquecimiento	cobertura ÷ frecuencia base

Es un cociente de dos proporciones y no tiene unidades. El símbolo $\times$ que lo acompaña no es el mismo $\times$ de una razón sobre un nulo, que compara otras dos cosas — y la confusión entre ambos es fácil de cometer al leer una tabla.

La advertencia: los enriquecimientos no son comparables entre macrosistemas. En las cinco aplicaciones el valor mediano va de 15,0 en austronesio a 190,5 en turco, un orden de magnitud, y esa diferencia no dice que el turco tenga códigos trece veces mejores: dice que las dos magnitudes que entran en el cociente dependen del tamaño y la forma del macrosistema. Dentro de un macrosistema el enriquecimiento ordena bien; entre macrosistemas, no se debe usar sin normalizar antes, y este trabajo no lo hace.

3 · El protocolo

Cinco aplicaciones enseñaron que el orden de los pasos no es indiferente. Éste es el protocolo tal como queda, con las correcciones que cada familia obligó a hacer.

3.1 · Fase 0: higiene, antes de mirar nada

Antes de construir un catálogo se audita la población cruda: duplicados de doculecto, lenguas mal clasificadas, formas de cita, esqueletos imposibles.

Va primero porque después ya no se ve. Un doculecto duplicado infla la cobertura de todo código en que aparezca, y una vez construido el catálogo la inflación es indistinguible de una señal. En urálico la fase 0 encontró una lengua kam-sui clasificada como fínica, señalada por tres chequeos independientes: está a 6.446 km del centro de la familia, marca tono en dígitos cuando ninguna urálica lo hace, y sus formas son θaːm^{453} «tres» y ðɐm^{214} «agua». Son 1.008 formas que habrían entrado al catálogo como material urálico.

La fase 0 corre sobre la población CRUDA y publica su cifra. Lo que quita, lo declara.

3.2 · Un macrosistema por medida

Regla adoptada tras medirlo: no se mezclan macrosistemas en una misma medida.

No es purismo. En este material la varianza entre lenguas domina cualquier efecto que se busque, y agrupar produce artefactos con forma de hallazgo. En un solo día de trabajo aparecieron tres: un gradiente que iba del 63 % al 86 % se disolvió entero al mirarlo dentro de cada lengua —paradoja de Simpson—, y lo mismo pasó con la aridad y con los núcleos.

De la misma regla se sigue que el valor de una clase es relativo al macrosistema que la produjo. Trasplantar una clase descubierta en austronesio para leer turco produce cifras con forma de dato y sin contenido.

3.3 · El despiece de la forma de cita

Las fuentes no citan igual. En turco, `ids` da los verbos casi siempre en infinitivo —el 85 % lleva `–mAK`— mientras otras fuentes los dan en raíz. El mismo verbo sale entonces con dos esqueletos distintos, `j·ʃ·m·q` frente a `j·ʃ`, y no forma código: son dos posiciones consonánticas que no son de la palabra sino del diccionario.

Quitarlas es legítimo y es peligroso, así que va con tres condiciones:

1. Se declara por familia. No hay despiece automático. Recortar finales «que parecen sufijos» es la vía rápida a fabricar exactamente los códigos que uno quiere ver.
2. Doble condición. Debe coincidir la ortografía y la cola del esqueleto. Si solo coincide una, no se toca — así una raíz que acabe en `–mak` por casualidad sobrevive. Es exclusión por construcción, no por criterio.
3. Se mide y se publica. Cuántas formas se tocaron y cuántos códigos se ganaron o perdieron. Un despiece que no rinde se retira.

3.4 · Declarar antes de medir

Todo estudio de esta serie escribe, antes de correr nada: qué se mide, sobre qué población, con qué nulo, con qué umbral, y qué contaría como falsación. Después se corre una vez y se publica salga como salga.

No es una formalidad. En los estudios de tono de esta serie, tres medidas quedaron por debajo de su bar y las tres se publicaron; y en dos ocasiones el protocolo impidió un resultado que habría salido de mover el umbral tras ver los datos.

El orden es: dispersión → nulo solo → techo → umbral → observado, una vez. El paso de dispersión —¿hay varianza que medir?— se añadió tras saltárselo dos veces y descubrir después que una de las dos variables no variaba, con lo que la medida no tenía respuesta posible.

3.5 · La puerta única

Todos los scripts leen los datos de una familia por una sola función, que aplica siempre las decisiones declaradas y cuenta lo que aplicó.

Existe porque no existía: el catálogo aplicaba todas las decisiones y el anexo solo algunas, así que el catálogo excluía una lengua y el anexo la enseñaba, en dos páginas seguidas del mismo paper.

Y esa consulta lleva **`ORDER BY`** terminado en una columna única. Sin él, PostgreSQL no promete orden y no lo daba: cinco de seis scripts eran no deterministas, y la tabla de sustituciones por par de ramas publicaba `b p` a 0,5× o `b n` a 1,9× según la corrida.

4 · Los nulos

Un catálogo sin nulo no dice nada: cualquier procedimiento que agrupe produce grupos. La pregunta es siempre cuántos habría producido el azar.

4.1 · Qué debe romper un nulo, y qué debe conservar

Un nulo debe romper exactamente el vínculo que se afirma, y conservar intacto todo lo demás que sea real.

El nulo de esta serie baraja los conceptos dentro de cada lengua. Rompe la relación forma–sentido —que es lo que se afirma— y conserva el inventario de cada lengua, su número de formas y la longitud de sus palabras, que son hechos.

Los nulos que se descartaron enseñan más que el que se usa:

- Colocar un elemento en una posición al azar. Descartado: destruye una restricción real y ajena a la hipótesis, y produce un nulo trivial de batir — un cociente grande por un motivo aburrido.
- Comparar, dentro de una lengua, formas marcadas contra no marcadas. Descartado tras comprobarlo: en vietnamita el 42 % «sin tono» era una fuente al 0 % frente a otra al 58 %. Compararía fuentes, no estados de lengua.

4.2 · El nulo se calcula exacto, no se estima

Durante tres papers el nulo se estimó barajando ochenta mil veces, y su intervalo se publicó como si midiera la incertidumbre de los datos. Medía la del sorteo. Se descubrió al quitar seis formas y ver que el nulo se movía un 14 %: una sensibilidad imposible si el intervalo describiera la población.

La forma cerrada es

```
(Σ_c n_c² − Σ_cl n_cl²) / (N² − Σ_l n_l²)
```

y al sustituirla las cifras publicadas cambiaron: 225× → 218×, 77× → 78×, 39× → 37×.

Regla: si el nulo tiene forma cerrada, se usa la forma cerrada; y se valida contra el barajado por fuerza bruta, porque una forma cerrada sin validar es exactamente lo que publicó un nulo equivocado durante tres papers.

4.3 · Qué autoriza batir un nulo

Esto es lo más fácil de exagerar y conviene enunciarlo estrechamente.

> Batir un nulo de barajado demuestra una CORRESPONDENCIA. Nada más.

Leer esa correspondencia como *migración*, como *herencia* o como *cambio* exige afirmar que existió una forma anterior de la que ambas descienden. Eso es una protoforma, y este método no las tiene. El nulo no la fabrica.

Los nulos no fallan por construcción; falla lo que uno se permite decir al batirlos. Es el error de alcance característico de este trabajo: afirmar más escala de la que se ha ganado.

4.4 · El umbral se fija antes, y contra dos límites

Un umbral elegido después de ver el resultado no es un umbral. Pero fijarlo antes tampoco basta: hay que fijarlo contra los dos límites que lo acotan, y ninguno es evidente.

Por arriba, el techo alcanzable. Toda medida tiene un máximo —para una información mutua, `min(H(a), H(b))`— y un umbral por encima de él declara la medida imposible de antemano. Hay que calcularlo y publicarlo.

Por abajo, los falsos positivos. Un nulo tiene media y también dispersión, y la dispersión no la da la forma cerrada. Comprobado: con un umbral de $2\times$ sobre la media del nulo, el 64 % de los pares lo superaba por azar en más de uno de cada veinte barajados. El $2\times$ que esta serie usó durante cinco papers no vale para una información mutua.

La regla, formulada tras corregirla una vez: el techo mínimo es un tope, no un objetivo; dentro de él se elige el valor más bajo cuya tasa de falsos positivos quede por debajo del 0,5 %. La primera formulación —«el más alto que siga siendo alcanzable»— funcionó por casualidad en un macrosistema donde los dos límites se cruzaban en un solo valor, y en el siguiente daba $15\times$, que ningún efecto real alcanza.

4.5 · Pseudorreplicación

Un conjunto de pares de lenguas no es un conjunto de observaciones independientes si las mismas lenguas aparecen en muchos pares. En un estudio de esta serie, 243 pares salían de once lenguas recombinadas.

Es el error que invalidó una versión anterior de un paper de esta misma serie. El resumen se hace por lengua, no por par, y el recuento de pares se reporta como lo que es, no como tamaño muestral.

5 · Lo que cuesta el instrumento, medido en cinco macrosistemas

El método se ha aplicado a cinco. La tabla sale de los catálogos publicados y de la puerta única, con las mismas decisiones que aplicó cada paper:

macrosistema	lenguas	formas	C/palabra	colisión	códigos	bin+ter	≥4	enriq. mediano
Austronesio	582	216.728	2,78	46,2 %	10.990	7.881	1.600	15,0
Turco	32	54.875	2,85	59,4 %	1.515	1.150	180	190,5
Urálico	27	80.032	3,28	59,1 %	1.474	1.175	133	154,9
Indoeuropeo	207	319.155	3,35	52,5 %	10.097	7.445	1.947	138,1
Pama-ñungano	154	45.829	3,35	34,0 %	2.092	1.563	447	40,3

5.1 · La colisión no la explica la longitud de la palabra

Sería lo esperable —palabras más cortas, más colisión— y no se cumple. Pama-ñungano e indoeuropeo tienen exactamente la misma longitud media, 3,35 consonantes, y su colisión difiere en dieciocho puntos: 34,0 % frente a 52,5 %.

De modo que el precio del instrumento no depende solo de cuántas consonantes tiene una palabra, sino de cómo se reparten las consonantes disponibles: un inventario grande y usado uniformemente produce menos choques que uno pequeño con unas pocas consonantes muy frecuentes.

Pama-ñungano es el macrosistema donde el instrumento cuesta menos de los cinco medidos, y esta comparación reproduce por su cuenta el veredicto de su paper, que se escribió sin ella.

5.2 · El enriquecimiento no es comparable entre macrosistemas

Ya avisado en §2.5 y aquí está la cifra: la mediana va de 15,0 a 190,5. Trece veces.

No dice que el turco tenga códigos trece veces mejores. Dice que las dos proporciones del cociente dependen del tamaño y la forma del macrosistema, y que el enriquecimiento ordena bien dentro de uno y no significa nada entre dos. Es la clase de cifra que invita a una tabla comparativa que no debe hacerse.

5.3 · La aridad alta señala morfología sin separar

Los códigos de cuatro consonantes o más van del 12,0 % en turco al 21,4 % en pama-ñungano, y en todos los casos donde se investigó resultaron ser morfología no separada. En turco, el despiece de la forma de cita (§3.3) los bajó del 29 % al 12 %.

Un porcentaje alto de aridad $\geq$4 es un diagnóstico, no un resultado: dice que hay que volver a la fase 0.

5.4 · Qué se aprende de aplicarlo cinco veces

Cada macrosistema enseñó algo distinto sobre el instrumento, y ninguna de las cinco lecciones se podía anticipar desde el diseño:

- Austronesio —el primero— mostró que el catálogo existe: que renunciar a la reconstrucción deja algo.
- Turco mostró qué significan sus cifras, por tener historia conocida contra la que contrastar. Y mostró el coste de la normalización: fundir toda sibilante en `s` hace invisible el lambdacismo, que es una de las correspondencias que define la división primaria de la familia.
- Urálico fue donde el instrumento por fin pudo hacer lo que promete, con la razón de nulo más alta.
- Indoeuropeo se hizo a ciegas, sin mirar la reconstrucción, para poder comparar después.
- Pama-ñungano es donde menos cuesta, y sugiere que el precio depende del inventario y no solo de la longitud.

6 · El criterio, puesto a prueba donde hay respuesta conocida

Todo lo anterior descansa en una apuesta: que concepto compartido más esqueleto idéntico identifica palabras emparentadas suficientemente a menudo. Es una apuesta comprobable en un sitio y solo en uno — el indoeuropeo, único macrosistema con juicios de cognación de experto.

Se usa `iecor-gold` (152 lenguas, 170 conceptos de léxico básico, juicios controlados) y se pregunta: de los pares que el criterio propone, ¿cuántos son cognados de verdad?

6.1 · El resultado

base — pares del mismo concepto que son cognados	389.457 / 1.495.154	26,0 %
código — de ésos, los que comparten esqueleto entero	52.780 / 56.559	93,3 %
cobertura — pares cognados que el criterio propone	56.559 / 389.457	13,6 %

El criterio es de precisión alta y cobertura baja. Nueve de cada diez códigos son un grupo de cognados real; y encuentra uno de cada siete de los que hay.

Eso es exactamente lo que un catálogo debe ser. Un instrumento que renuncia a la reconstrucción no puede aspirar a encontrarlo todo; lo que no puede permitirse es que lo que encuentre sea ruido. Con 93,3 %, no lo es.

6.2 · Una tensión con §2.4, y no se resuelve a favor nuestro

Por número de consonantes del esqueleto compartido:

aridad	precisión
1	92,8 %
2	93,8 %
3	92,6 %
4+	93,7 %

Es plana. Un código unario acierta tanto como un ternario, y eso contradice §2.4, que dice que un unario no sostiene ninguna afirmación por sí solo.

No se va a disimular la contradicción, pero tampoco se resuelve a favor de los unarios, por una razón que está en la población medida: todas estas formas tienen juicio de cognación, es decir, pertenecen ya a algún conjunto. La medida dice cuántos de los pares propuestos son cognados entre formas que están en el inventario de cognados, no entre formas cualesquiera. Y la base de partida —26 %— ya es altísima, porque el léxico básico de una familia bien atestiguada es denso en cognados.

La lectura prudente: en léxico básico indoeuropeo, una sola consonante compartida más el concepto predice cognación al 93 %. No autoriza usar unarios en un macrosistema donde la densidad de cognados es desconocida, que es el caso de los otros 224. La restricción de §2.4 se mantiene, ahora como decisión declarada y no como hecho medido, que es una posición más honesta que la anterior.

6.3 · Y lo que esta prueba no puede decir

`iecor-gold` son 170 conceptos de léxico básico en un macrosistema con dos siglos de trabajo detrás. El criterio podría comportarse peor en léxico no básico, y no hay manera de saberlo: no existe cognación de oro fuera de ahí.

Ésa es la asimetría del programa tal como está hoy integrado el corpus, y conviene decirlo así: fuera del indoeuropeo existen conjuntos de cognación publicados —el *Austronesian Comparative Dictionary*, las bases urálicas, el trabajo bantú y sino-tibetano— que no están integrados aquí. Integrarlos ampliaría el banco de pruebas, y es trabajo pendiente antes que un límite del campo.

Lo que sí es asimétrico de raíz es que el instrumento se diseñó para los macrosistemas sin tradición comparativa, y ésos son justamente los que no pueden servir de banco de pruebas. La respuesta no es fingir que la calibración transfiere, sino declarar el alcance de lo comprobado y aplicar el instrumento sabiendo qué se ignora.

7 · Falsadores y límites

7.1 · Qué falsaría este método

Un método que no se puede falsar no es un método. Éstos son los resultados que lo tumbarían, y ninguno ha ocurrido todavía:

Que el catálogo no batiera su nulo. Si el agrupamiento real cubriera aproximadamente lo mismo que el barajado, no habría correspondencia forma–sentido que medir. En las cinco aplicaciones el real supera al nulo por un factor de tres a siete.

Que el criterio no discriminara donde se puede comprobar. Si concepto + esqueleto no elevara la tasa de cognados por encima de su base en indoeuropeo, todo lo construido sobre él sería ruido. Da 93,3 % frente a 26,0 % (§6).

Que los códigos aparecieran igual dentro y fuera de rama. Se comprobó en un estudio de esta serie: la razón sobre el nulo es 1,29× dentro de rama y 0,99× fuera — exactamente el nulo. Que la distinción de rama aparezca como salida de un diseño que no la usa como entrada es lo que autoriza aplicar el instrumento donde no hay ramas.

Que la aridad alta no se explicara por morfología. Se explicó: el despiece de la forma de cita bajó los códigos de cuatro o más del 29 % al 12 % en turco.

7.2 · Los límites que no se van a resolver

No distingue herencia de préstamo, ni lo hará. Es consustancial a renunciar a las protoformas.

No es comparable entre macrosistemas más allá de la colisión. El enriquecimiento no lo es (§2.5, §5.2), y las clases descubiertas en uno no valen en otro.

La cobertura es baja por diseño. Encuentra uno de cada siete grupos de cognados que existen (§6.1). Un catálogo no es un inventario.

Depende de una capa que no controla. El esqueleto lo produce el corpus; sus decisiones de normalización —qué se funde y qué no— determinan qué correspondencias pueden verse. En turco, fundir las sibilantes hace invisible el lambdacismo, que es la correspondencia que define la división primaria de la familia. Ese coste se declara en cada aplicación y no se puede eliminar, solo medir.

Y no se puede calibrar donde hace falta. §6.3.

7.3 · Lo que este método no es

No es una alternativa al método comparativo ni una versión barata de él. Produce menos, afirma menos y cuesta menos, y su única ventaja es que se puede aplicar donde el otro no se ha aplicado. Donde hay reconstrucción, el método comparativo es mejor en todo.

8 · Cómo se aplica a un macrosistema nuevo

El protocolo completo, en el orden en que hay que hacerlo. Cada paso está donde está porque saltárselo costó algo en alguna de las cinco aplicaciones.

1 · Fase 0, sobre la población cruda. Duplicados de doculecto por glotocódigo, lenguas mal clasificadas, formas de cita, esqueletos imposibles. Se publica lo que se quita y por qué. *Va primero porque después la inflación es indistinguible de la señal.*

2 · Declarar las decisiones de la familia. Qué doculectos se excluyen y con qué regla mecánica; si hay despiece de forma de cita, con su doble condición. *Nada automático: una regla que nadie escribió no se aplica.*

3 · Publicar la tabla de normalización, derivada y no escrita a mano, con el coste que impone: qué correspondencias deja de poder ver esta familia por lo que la normalización funde.

4 · Construir el catálogo sin genealogía. Esqueleto, concepto, umbral de tres lenguas. Nada más.

5 · Comprobar la aridad. Si los códigos de cuatro o más pasan del 15 %, volver al paso 2: es morfología sin separar, no un hallazgo.

6 · El nulo, exacto y validado. Barajar conceptos dentro de cada lengua. Si tiene forma cerrada, forma cerrada — y contrastada contra el barajado por fuerza bruta.

7 · El umbral, antes de mirar lo observado, acotado por arriba por el techo alcanzable y por abajo por la tasa de falsos positivos medida sobre el propio nulo.

8 · Correr una vez. Y publicar el resultado sea cual sea.

9 · Solo ahora, la genealogía, y solo para leer lo que salió: ¿reencuentra el catálogo algo que la clasificación reconoce?

10 · Escribir qué enseñó esta familia sobre el instrumento, que suele ser lo más útil del ejercicio y no se puede anticipar. Las cinco enseñaron algo distinto (§5.4).

Dónde está el material, y cómo se pide

Todo —el pipeline, los catálogos, los cinco papers y las declaraciones previas de cada estudio— está en el repositorio del programa. Los datos vienen del Corpus Integrativo, descrito aparte.

Los dos repositorios son privados hoy y el acceso se da por solicitud, escribiendo a alex@pixelspace.com con quién se es y para qué. Quien lo recibe tiene lectura completa y puede bifurcarlos a su propia cuenta.

La razón está en el paper del corpus: mientras las versiones de las fuentes no estén fijadas, un clon público prometería una reproducibilidad que aquel documento declara que todavía no existe. Es un estado, no una política, y se levanta cuando esas versiones se fijen.

La licencia no depende de eso: CC-BY-SA-4.0 sobre el contenido, y ShareAlike se hereda.

9 · Referencias

Este apartado faltaba. El texto citaba por autor y año —Dolgopolsky, List, Jäger, ASJP— sin que ninguna de esas citas tuviera entrada, que es el mismo defecto que una revisión externa marcó como bloqueador en el paper del corpus. Las entradas identifican la obra, no la versión que se usó; para el corpus del que salen los datos, la huella exacta de los ficheros está en su manifiesto de fuentes.

9.1 · Los otros dos trabajos de esta serie

- Toledo Martínez, A. (2026). *Corpus Integrativo: una base léxica multi-capa de 225 macrosistemas, con procedencia por fila y su aparato de verificación.* — El recurso sobre el que corre todo lo de aquí. El catálogo de este trabajo no construye datos: los lee de ahí, y el reparto entre los dos está en §1.5.
- Toledo Martínez, A. (2026). *Meulemans: una capa de consulta de solo lectura sobre un corpus léxico multi-capa de 225 macrosistemas.* — El otro consumidor del mismo corpus, y el contraste útil: enseña el linaje que este método se prohíbe mirar.

9.2 · El esqueleto consonántico y la comparación automática

- Dolgopolsky, A. B. (1964). Gipoteza drevnejšego rodstva jazykovyx semej Severnoj Evrazii s verojatnostnoj točki zrenija. *Voprosy Jazykoznanija* 2, 53–63. — El *consonant class matching* del que desciende el esqueleto de §2.1.
- Turchin, P., Peiros, I. & Gell-Mann, M. (2010). Analyzing genetic connections between languages by matching consonant classes. *Journal of Language Relationship* 3, 117–126. — La versión que trunca a las dos primeras consonantes, de la que §1.5 se separa a propósito.
- List, J.-M. (2012). LexStat: Automatic detection of cognates in multilingual wordlists. *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, 117–125.
- List, J.-M. (2014). *Sequence Comparison in Historical Linguistics*. Düsseldorf: Düsseldorf University Press.
- List, J.-M., Greenhill, S. J. & Gray, R. D. (2017). The potential of automatic word comparison for historical linguistics. *PLoS ONE* 12(1), e0170046.
- Jäger, G. (2018). Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data* 5, 180189.

9.3 · ASJP, con el que no debe confundirse

- Brown, C. H., Holman, E. W., Wichmann, S. & Velupillai, V. (2008). Automated classification of the world's languages: A description of the method and preliminary results. *STUF – Language Typology and Universals* 61(4), 285–308.
- Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A. & Bakker, D. (2008). Explorations in automated language classification. *Folia Linguistica* 42(3–4), 331–354.
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10(8), 707–710. — La distancia que ASJP corrige por azar, y que §1.5 distingue de lo que se hace aquí.

9.4 · Los nulos

- Vinh, N. X., Epps, J. & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research* 11, 2837–2854. — El sesgo de la información mutua con muestra finita y su corrección en forma cerrada, que es lo que implementa `cod/metricas.py` y lo que §4.2 generaliza como regla.

9.5 · El patrón de referencia de §6

- 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, eabg0818. — IE-CoR, de donde sale `iecor-gold`: los juicios de cognación contra los que se calibra el instrumento.