

Repertorios empíricos de transformación fonológica

De correspondencias reales a estructura mensurable: el repertorio observado O , el subespacio que genera $\langle O \rangle$, y sus límites

transformations · resultados intra-sistema · v3 (tras dos rondas de revisión externa)

0. La idea en una frase

Si a cada correspondencia de sonido de una familia la describimos no como “la letra a se vuelve la letra b ” sino como qué RASGOS difieren, la variación deja de ser una lista y se vuelve un objeto mensurable: un conjunto de operadores dentro de un espacio de rasgos, con un rango, una densidad y unos ejes propios. Este documento *muestra* eso —con palabras reales, el grafo y la composición— en Indoeuropeo (IE) y Austronesio, y es explícito sobre qué es propiedad de las lenguas y qué es propiedad de la notación.

Nota (informe de investigación, v3). Dos rondas de revisión externa. La v2 separó tres niveles: (1) el álgebra de la representación, (2) el repertorio empírico observado, (3) los procesos históricos. La v3 añade, a pedido de la 2ª ronda: modelo nulo para $C(O)$; distinción rango/corango/ocupación con la dependencia concreta; métricas de cognación separadas por unidad (par vs tipo); rarefacción de ramas; y la terminología repertorio O vs subespacio $\langle O \rangle$. Las afirmaciones que no sobrevivieron se marcan retractadas.

1. Cómo se lee un operador — y qué NO es

- Sin proto-lenguaje, como propiedad del método. No anclamos las formas a una raíz reconstruida: detectamos cognados estadísticamente (LexStat) y alineamos las formas entre sí. El cálculo de operadores *no requiere* protoformas; las reconstrucciones se usan después, como interpretación independiente, nunca como insumo.
- Un operador = el conjunto de rasgos primarios que DIFIEREN entre dos segmentos alineados. Ejemplo: $d \ z$ difieren en continuidad y estridencia $\rightarrow \{\text{cont+strid}\}$.
- Es un contraste SIMÉTRICO, no un cambio dirigido. Como $\Delta(a, b) = \Delta(b, a)$, el operador no distingue $t \rightarrow s$ de $s \rightarrow t$. Es una *firma diferencial* de rasgos, una distancia orientable por una dimensión —no un operador histórico. Que “cada operador sea su propio inverso” es consecuencia de esa simetría, no un hallazgo fonológico. Para modelar cambio histórico haría falta una representación dirigida $T_{a \rightarrow b}$ con contexto, posición y lengua (ver §9).

2. Los operadores en PALABRAS REALES (Indoeuropeo)

Conjuntos de cognados detectados por LexStat; cada uno alinea las formas y deja ver qué operador actúa en cada posición.

Salvedad metodológica (importante). LexStat agrupa por señal fonética dentro de un mismo concepto y puede mezclar raíces históricamente distintas ('foot': pod- vs otras; 'tooth': el griego ὀδούς y el eslavo зѡбъ probablemente no son el mismo étimo). Cuando eso ocurre, la diferencia de rasgos entre los segmentos alineados sigue siendo un contraste formal, pero no una correspondencia genealógica. Hay que distinguir dos corpus: D_{cognado} (etimología validada) y $D_{\text{conceptual}}$ (equivalencia de significado). Lo que sigue es, con rigor, *álgebra de la variación formal asociada a conceptos*; la validación cognado-por-cognado es tarea de la fase 2.

'two' — el dos:

lengua	forma (IPA)
Rumano	/doi/
Serbocroata	/dva/
Armenio oriental	/jerk ² u/
Armenio occidental	/jergu/
Romaní	/dui/

Entre las dos variedades del armenio (*erku*, el caso célebre) aparece el operador **k' g {voi}** en posición homóloga —un contraste de un solo rasgo, sonoridad. La vocal final oscila o i u.

'tooth' (*posible mezcla de raíces — ver salvedad*): griego d frente a eslavo z da el contraste d z {cont+strid}. Es un buen ejemplo del *tipo* de operador, no evidencia de una correspondencia histórica d z mientras no se confirme cognación.

2b. ¿Cuánto contamina esto? — validación contra la cognación experta de IE-CoR

La salvedad anterior se puede *medir*. En IE-CoR (iecor), que trae cognación experta (gold), sobre las mismas 30 lenguas comparamos tres corpus: GOLD (alineal solo dentro de cognados expertos), LEX-STAT (lo que usa nuestro pipeline) y CONCEPTO-todo (alineal todos los pares del mismo concepto, *sin* cognación).

Hay que medir en DOS unidades distintas (lo confundía la v2): el par de formas (¿son cognados?) y el tipo de operador (¿está ese operador en el repertorio?). No son lo mismo.

Nivel cognático (unidad = par de formas): frente al gold, LexStat tiene $P_{\text{cog}} = 0.73$, $R_{\text{cog}} = 0.33$ — es *conservador* (pierde muchos pares reales) y une 2000 pares no-cognados (contaminación de instancias).

Nivel de repertorio (unidad = tipo de operador):

corpus	operadores	P_{op}	R_{op}	Jaccard espurios c/gold		C(O)
GOLD (cog- nado ex- perto)	65	—	—	—	—	0.17
LEXSTAT (nuestro pipeline)	43	1.00	0.66	0.66	0 (0 %)	0.16
CONCEPTO- todo (sin cognación)	108	0.60	1.00	0.60	43 (40 %)	0.20

El resultado interesante: el inventario de tipos es mucho más fiel que la identificación individual de cognados ($P_{\text{op}} \approx 1.00$ frente a $P_{\text{cog}} = 0.73$). Un falso cognado casi siempre produce un operador que

3. El GRAFO de operadores del sistema

4. La composición — y por qué el XOR es una identidad, no un hallazgo

$$\Delta(a, c) = \Delta(a, b) \oplus \Delta(b, c)$$

3

Honestidad (corregido en v2). Para rasgos estrictamente binarios esta igualdad es una identidad matemática —se cumple al 100% por construcción—, no un resultado empírico. Medimos 98.1% en IE. El detalle interesante es *dónde* falla: las 5/5 desviaciones son exactamente los tríos donde un rasgo toma tres valores distintos ($hi \in \{+1, 0, -1\}$: $l \quad l^j \quad r, l \quad r \quad \wedge \dots$). Es que los rasgos de panphon son ternarios, no binarios. Por tanto el XOR-break no descubre cadenas: diagnostica qué dimensiones del sistema son graduadas (aquí, la altura/hi, que debe modelarse con estados ordenados coronal $\rightarrow$ palatal $\rightarrow$ rótica, no con un bit). Cautela: eso identifica una dimensión multivalente (tres valores incompatibles con un bit), no prueba que sea una escala totalmente ordenada única — $l, l^j, r, \wedge$ varían a la vez en lateralidad, roticidad y lugar; si el orden es total, parcial o un grafo de estados, es cuestión aparte. Austronesio cumple 47/47 (100%): no se detectaron violaciones en su muestra (con $n=47$ una tasa pequeña podría no observarse), no “carece de ejes graduados”.

5. El repertorio O , los átomos, y el rango del subespacio $\langle O \rangle$

Un operador de un solo rasgo es un átomo; los de varios son moléculas. (*Rasgos primarios corregidos en v2: incluye lab —labialidad— tras notar que su ausencia sub-describía el lugar labiodental; p.ej. $t \ s=\{cont+strid\}$ pero $t \ f=\{cont+cor+strid+lab\}$, operadores distintos.*)

sistema	átomos observados	operadores-tipo $ O $	rasgos codificados	rango r	corango	$ \langle O \rangle - 1 = 2^r - 1$
Indoeuropeo	8	67	12	12	0	4095
Austronesio	3	22	11	10	1	1023

Aclaración clave (§ronda3.3): “átomos” = operadores unitarios OBSERVADOS (IE 8, AN 3); el rango se calcula sobre *todos* los rasgos que aparecen en cualquier operador —incluidos los que solo salen en moléculas—, por eso puede superar el número de átomos (IE: 8 átomos, rango 12). En IE el rango es pleno ($\dim = 12$, corango 0); en Austronesio rango 10 de 11 rasgos codificados (corango 1): en esta muestra la columna **ant** = XOR(**cor**, **lab**) —una dependencia lineal del *patrón de activación* entre operadores, no una reducción ontológica del rasgo. Lo que distingue a los sistemas no es un “número de dimensiones” sino el inventario de átomos y la ocupación: ambos usan solo 2 % del subespacio que sus operadores generan ($\rho \approx 0.016\text{--}0.022$) —el repertorio observado O es una región minúscula de $\langle O \rangle$.

RETRACTADO (v1 afirmaba lo contrario). El borrador anterior decía que nas, round y lo “nunca operan solos” y concluía que la nasalidad “no es un movimiento independiente del sistema”. No sobrevive a un modelo nulo. Con solo 8 operadores atómicos, la probabilidad de no ver *ninguno* igual a {nas} por puro azar (sorteando átomos $\propto$ cuánto cambia cada rasgo) es $p \approx 0.60\text{--}0.86$ (IE): no significativo. La ausencia en una muestra pequeña no demuestra dependencia. Retiramos la afirmación.

6. ¿Cuán “cerrado” trabaja el sistema? — el índice de realización composicional

El borrador hablaba de “cierre al 39%”, lo cual es contradictorio: el cierre algebraico es una propiedad binaria (un conjunto está cerrado, o no). O no está cerrado ($0 = \emptyset \notin O$). Lo que se mide es una razón de realización composicional. Como para $u \neq v$ en $\mathbb{F}_2$ siempre $u \oplus v \neq \emptyset$, el denominador es simplemente $\binom{|O|}{2}$ (se corrige la def. de la v2):

$$C(O) = \frac{|\{\{u, v\} \subseteq O : u \oplus v \in O\}|}{\binom{|O|}{2}}$$

Resultado: IE $C(O) = 0.220$, Austronesio $C(O) = 0.234$. Pero un número solo es descriptivo hasta que se compara con un modelo nulo. Generamos repertorios nulos que preservan $|O|$ y la distribución de pesos de Hamming (los tamaños de operador), muestreando cada operador como subconjunto aleatorio de ese tamaño $\propto$ la frecuencia marginal de cada rasgo, y medimos $Z = (C_{\text{obs}} - \mathbb{E}[C_{\text{null}}])/\text{sd}$:

sistema	$C(O)$	$\mathbb{E}[C_{\text{null}}]$	Z
Indoeuropeo	0.220	0.146 ± 0.017	+4.3
Austronesio	0.234	0.122 ± 0.045	+2.5

Ahora sí es un hallazgo estructural: ambos repertorios son más componibles que el azar dado su tamaño y sus tamaños de operador (IE con holgura, AN marginalmente). No es un grupo, pero tampoco un conjunto arbitrario: O se concentra donde las composiciones vuelven a caer dentro de O . (Advertencia: $C(O)$ depende de $|O|$, del rango y del umbral; dos sistemas con $C(O)$ parecido pueden tener geometrías distintas —por eso el Z , no el valor crudo, es lo comparable.)

7. Comparativa: Austronesio, otra distribución

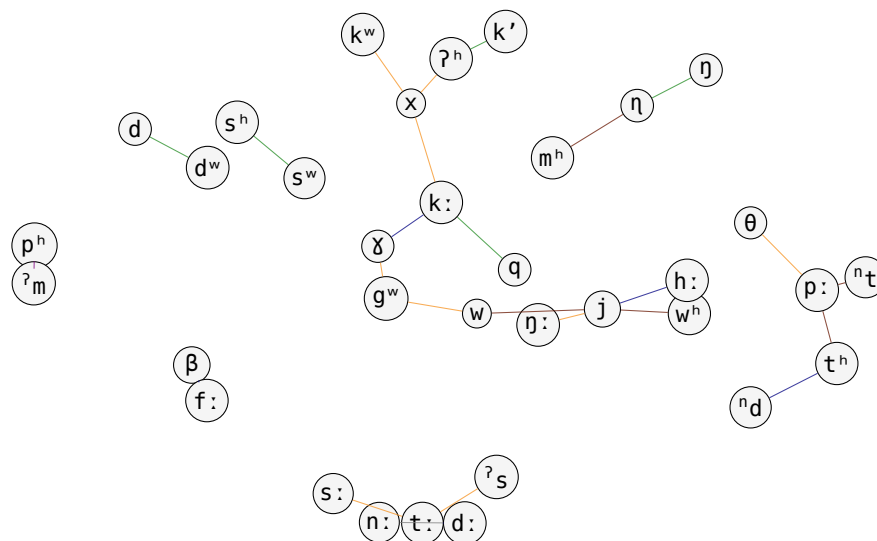


Figure 2: Grafo de operadores de Austronesio. Nodos = segmentos; aristas = correspondencias sistemáticas, con grosor $\propto$ frecuencia y color por tipo de cambio: **sonoridad**, **fricativización/lenición**, **palatalización/altura**, **lugar**, **prenasalización**.

'eye' — sobre la base **mata*:

lengua	forma (IPA)
Tongan	/mata/
Malgache	/maso/
Rotumano	/mafa/
Takia	/mala/
Hawaiano	/maka/

La t de mata* contrasta distinto en cada lengua: t s , t f , t l , t k . No son "la misma lenición". Son operadores diferentes**: t s ={cont+strid} (fricativización coronal), t f ={cont+cor+strid+lab} (fricativización + lugar labial), t k =cambio de lugar, t l =lateralización + clase. El propio modelo

de rasgos lo muestra —imponerles un proceso común sería un error. (*La salvedad de matriz de la v_l queda resuelta: con lab en el conjunto, t s y t f reciben deltas distintos.*)

En la muestra austronesia se observan 3 operadores unitarios (no 8: {voi} {cont} {hi}) y ninguna violación del XOR ($n=47$; una tasa pequeña de dimensiones graduadas podría no aparecer). Su firma es la sonorización prenasalizada t^nd , k^ng , p^nb , que empaqueta nasalidad y sonoridad —que no aparece como operador sistemático en la muestra indoeuropea analizada (no afirmamos su ausencia de todo el IE). Es una distribución distinta, y además su rango $\mathbb{F}_2$ (10) es menor que el pleno de IE (12).

8. ¿El "álgebra de la familia" es una superposición de álgebras de rama?

La objeción más seria de la revisión (§6): agrupar una familia enorme bajo *un* álgebra puede estar superponiendo varias álgebras de rama, dominadas por las ramas mejor muestreadas. Lo probamos: partimos cada familia en sus ramas genealógicas (subgrupos de Glottolog, corte del árbol por tamaño) y corremos el mismo pipeline dentro de cada rama, comparando los conjuntos de operadores.

	ramas (≥ 8 lenguas)	Jaccard medio entre ramas	ops. de familia dentro de ≥ 1 rama	núcleo ($\geq 1/2$ ramas)	de 1 sola rama
Indoeuropeo	7 (Indoiranio, Itálico, Germánico, Balto-eslavo, Céltico, Grecofrigio, Armenio)	0.42 (0.22–0.62)	67/67 = 100%	54	32
Austronesio	11 (Oceánico, Filipino Central, Barito, Luzón N., Formosano...)	0.51 (0.32–0.74)	22/22 = 100%	22	14

(*Jaccard medio = promedio no ponderado de pares de ramas; ponderarlo por tamaño de repertorio daría otra cifra para otra pregunta.*) Tres lecturas, todas con matiz:

1. No es superposición pura ni álgebra universal. El Jaccard medio entre ramas (0.42 IE, 0.51 AN) está en el medio: las ramas comparten un núcleo pero cada una tiene operadores propios. {voi} y {cont} son los únicos átomos presentes en todas las ramas de las dos muestras —candidatos a dimensión de cambio compartida (con solo dos familias no afirmamos "universal").
2. El álgebra familiar NO es un artefacto cross-rama. El 100% de los operadores de la familia aparece dentro de al menos una rama: alinear entre ramas no fabrica operadores espurios. Pero la *riqueza* aparente de la familia sí es una unión —IE acumula 32 operadores que solo viven en una rama.
3. IE parece más diferenciada por rama que Austronesio (Jaccard más bajo, más operadores de una sola rama). Es *compatible* con una mayor diferenciación histórica de la muestra IE, pero no lo demuestra: también influyen diversidad tipológica, desigualdad de ramas, cobertura y muestreo.

Cautela de muestreo (revisión ronda 3). Los recuentos brutos de operadores por rama son artefacto de muestreo: bajo rarefacción a $n=8$ lenguas/rama el ranking de "riqueza" se desordena —Armenio (≈ 14) deja de ser la más pobre; Itálico (≈ 12) e Indoiranio (≈ 9) quedan por debajo. Por eso no interpretamos la riqueza por rama; lo que sí es robusto al muestreo es el solapamiento de conjuntos (Jaccard) y el núcleo compartido. Reproducible: `TF_RAREFY=3 make superposition FAMILY="X"`.

9. Qué sobrevive, en claro

1. Cada sistema ocupa un subconjunto propio del espacio de rasgos, más componible que el azar: $C(O) = 0.22$ (IE, $Z = +4.3$), 0.23 (AN, $Z = +2.5$) frente a un nulo que preserva tamaño y pesos.
2. Los sistemas difieren en su distribución —IE (8 átomos, rango F_2 12, corango 0) vs Austronesio (3 átomos, rango 10, corango 1, prenasalización); no en un "número de dimensiones" sino en inventario y densidad.
3. El XOR-break localiza dimensiones multivalentes (en IE, la altura), que piden una geometría distinta de $\mathbb{F}_2$. (*diagnóstico de la representación, no descubrimiento de una escala fonológica*)
4. Los operadores son contrastes simétricos, no cambios dirigidos; la historia necesita la capa dirigida $T_{a \rightarrow b}$ con contexto.
5. El álgebra de familia es un núcleo compartido + extensiones de rama, no superposición pura ni universal; su riqueza agregada es en parte una unión dominada por las ramas mejor muestreadas.

Retractado: que la nasalidad/redondeo sean "movimientos no independientes" (no pasa el nulo).
Marco corregido: el grupo abeliano bajo XOR es propiedad de la *representación*, no un hallazgo empírico.

10. El núcleo científico: repertorio observado O vs subespacio generado $\langle O \rangle$

Terminología (revisión ronda 3): O no es un subespacio —no está cerrado bajo XOR ni contiene el 0. Es el repertorio (región, soporte) observado. El subespacio es solo $\langle O \rangle = \text{span}_{\mathbb{F}_2}(O)$. La distinción es productiva porque separa dos objetos:

- $O_L \subseteq \mathbb{F}_2^n$ — el repertorio de firmas diferenciales observadas del sistema L ;
- $V_L = \langle O_L \rangle$, con $\dim V_L = r_L$ — la capacidad algebraica generada;
- $\rho_L = |O_L|/(|V_L| - 1)$ — la ocupación ($\approx 2\%$ en ambas familias): cuánto del subespacio se usa de verdad;
- $C(O_L)$ — la realización composicional (con su Z frente al nulo);
- $K = \bigcap_r O_r$ (núcleo entre ramas) y $E_r = O_r \setminus \bigcup_{s \neq r} O_s$ (extensiones exclusivas);
- y, para historia, la capa dirigida $T = (a, b, c, p, \ell_1, \ell_2, w)$ con firma $\sigma(T) = \Delta(a, b)$ pero $T_{t \rightarrow s} \neq T_{s \rightarrow t}$.

Esto abre preguntas nuevas y verdaderas: ¿qué operadores generables ($V \setminus O$) nunca se observan? ¿una rama expande el rango o solo añade densidad dentro del mismo span (innovación *dimensional* vs *combinatoria*)? ¿dos familias con el mismo rango ocupan regiones distintas? La tesis, en una frase:

Las lenguas no "forman un grupo" de cambios; sus correspondencias observadas ocupan de manera desigual y medible un espacio algebraico que proporciona la representación de rasgos.

Restante: modelos nulos también para átomos, rango y exclusividad de rama; intervalos por bootstrap; metadatos de reproducibilidad; grafos con leyenda, umbral y coordenadas articulatorias.

Todo intra-sistema, sin presuponer historia ni universalidad. Reproducible: `make family FAMILY="X", make algebra FAMILY="X", make chains FAMILY="X", make superposition FAMILY="X", make cognate-eval, bash docs/build.sh`.