

La tabla periódica de los códigos

Patrones matemáticos del cambio entre los ocho sistemas indoeuropeos: un enfoque de lingüística cuantitativa

Alejandro Toledo Martínez

Libro comparativo del programa *códigos coderivativos*. Los ocho volúmenes de rama (indo-iranio, balto-eslavo, germánico, itálico, céltico, helénico, armenio, albanés) modelaron el esqueleto consonántico de cada palabra —su *código* binario o ternario— leyendo hacia atrás, de cada lengua a su proto-subsistema. Aquí no se describe una rama más: se comparan los sistemas entre sí para buscar patrones matemáticos del cambio. La tesis es que las regularidades clásicas (la partición satem/centum, las restricciones de la raíz PIE, la deriva léxica) admiten una lectura *cuantitativa* —un número, una matriz, una tabla con huecos— y que esos huecos, como en la tabla de Mendeleiev, predicen estructuras aún no observadas.

1 Qué se compara y con qué método

El programa produjo, para cada una de las ocho ramas indoeuropeas, una base reproducible con: (i) el *código* de cada raíz cognada —la secuencia de consonantes canónicas, binaria o ternaria—; (ii) las correspondencias emergentes proto → lengua; (iii) el inventario completo de códigos; (iv) un ajuste de decaimiento del código en el tiempo. Este volumen toma esas ocho bases y calcula cinco familias de patrones entre sistemas. No buscamos “reglas de conversión” evidentes entre lenguas hermanas (del tipo castellano ~ portugués), sino *invariantes cuantitativos* que se sostengan a lo largo de la familia.

El método se inscribe en la lingüística histórica cuantitativa y computacional —la regularidad neogramática del cambio fónico (Osthoff & Brugmann, 1878), la formalización de las leyes de sonido como funciones regulares (autómatas de estado finito) (Kaplan & Kay, 1994), la medida de correspondencias y distancias fonéticas (Kondrak, 2000; List, 2012), y la filogenia cuantitativa de lenguas (Swadesh, 1952; Gray & Atkinson, 2003; Bouckaert et al., 2012)—. La novedad no es el aparato, sino el *objeto*: aplicamos esos instrumentos no a las palabras, sino a los códigos consonánticos que el programa extrajo.

2 Patrón 1 · La isoglosa satem/centum como número

La división *satem* / *centum* —según si las palatovelares PIE *k̑ *ǵ se hacen sibilantes o permanecen velares (Brugmann, 1904; Meillet, 1908)— es la partición mayor del indoeuropeo. Se la enuncia siempre de forma cualitativa. Aquí la convertimos en un índice continuo. Para cada rama medimos, sobre las raíces cuya reconstrucción contiene una palatovelar, la probabilidad de que el código realice una sibilante, menos esa misma probabilidad sobre las raíces de velar simple:

$$\text{Satem(rama)} = P(\text{sib} \mid \text{palatovelar}) - P(\text{sib} \mid \text{velar simple}).$$

La resta controla la tasa de fondo de sibilantes de cada rama: un valor positivo indica que las *palatovelares en particular* sibilan, que es justo la definición de satem.

El resultado reproduce la partición clásica sin haberla supuesto: emerge de contar sibilantes en los códigos. Dos matices honestos: el albanés da un índice débil (+0,05) —coherente con que fusionó las

rama	grupo	índice satem	P(sib palv.)	P(sib velar)
Indo-Iranio	satem	+0.171	0.43	0.26
Balto-Eslavo	satem	+0.184	0.51	0.33
Armenio	satem	+0.354	0.44	0.08
Albanés	satem	+0.050	0.42	0.37
Germánico	centum	-0.052	0.27	0.32
Itálico	centum	-0.100	0.37	0.47
Céltico	centum	+0.092	0.33	0.23
Helénico	centum	-0.150	0.38	0.53

Media satem = +0.190, *centum* = -0.052; *Cohen's d* = 2.1.

Table 1: Índice satem por rama. Las cuatro satem tradicionales quedan del lado positivo; las cuatro centum, en torno a cero o negativas. La separación de medias es grande (Cohen's $d \approx 2,1$).

series de otro modo y se documenta tarde— y el céltico da un leve positivo espurio (+0,09, muestra fina). Aun así, el contraste de grupos es nítido (media satem +0,190 frente a centum -0,052). El valor de la operación es que una isoglosa cualitativa de manual se vuelve una magnitud comparable, ordenable y —en principio— proyectable a lenguas fragmentarias.

3 Patrón 2 · La tabla periódica de los códigos

Éste es el corazón del libro. Reunimos todos los códigos *binarios* de las ocho ramas ($N = 490$ códigos-rama distintos) y construimos la matriz de ocupación $M[c_1][c_2]$: cuántas ramas realizan cada combinación primera-consonante $\times$ segunda-consonante. Es, literalmente, una tabla periódica: las filas y columnas son las “valencias” consonánticas y cada celda, un elemento posible.

	r	n	s	l	t	w	d	j	k	g
p	4	2	2	6	5	3	4	6	3	2
k	2	2	5	3	2	6	1	5	Ø	Ø
s	1	5	Ø	4	4	4	4	1	5	1
n	2	1	5	1	2	6	2	1	1	6
g	4	7	6	6	Ø	1	4	4	1	Ø
w	4	2	4	2	2	Ø	7	2	Ø	3
m	6	4	5	3	4	1	2	1	Ø	3
d	4	3	2	3	1	7	Ø	4	Ø	3
t	4	4	Ø	2	Ø	Ø	Ø	Ø	3	2
b	6	4	1	3	2	1	1	2	2	0

Table 2: Matriz de ocupación de los códigos binarios (10 consonantes más frecuentes en cada posición; número de ramas que realizan cada código). En **rosa** los huecos: celdas cuya frecuencia esperada bajo independencia es ≥ 1 pero que ninguna rama realiza.

La pregunta de Mendeleiev es: ¿los huecos son ruido, o los predice una ley? Definimos como *hueco* toda celda con ocupación observada nula pero frecuencia esperada ≥ 1 bajo independencia ($\hat{e} = N P(c_1)P(c_2)$). Hay 33 huecos así; y 24 de ellos (73 %) los predice una restricción fonotáctica conocida.

código	esperado	espejo obs.	restricción que lo predice
s·s	3.66	0	geminado (anti-geminación / OCP)
g·t	2.57	2	dos oclusivas (co-ocurrencia PIE)
w·w	2.42	0	geminado (anti-geminación / OCP)
l·r	2.40	0	dos sonorantes (OCP-sonorante)
t·s	2.35	4	sibilante+coronal
l·n	2.30	1	dos sonorantes (OCP-sonorante)
d·d	2.29	0	geminado (anti-geminación / OCP)
k·k	2.23	0	geminado (anti-geminación / OCP)
k·g	2.23	1	dos oclusivas (co-ocurrencia PIE)
l·l	2.16	0	geminado (anti-geminación / OCP)
g·g	1.94	0	geminado (anti-geminación / OCP)
w·k	1.89	6	—
m·k	1.89	2	—
t·t	1.89	0	geminado (anti-geminación / OCP)

Table 3: Los huecos de mayor frecuencia esperada, la ocupación de su *espejo* $c_2 \cdot c_1$, y la restricción que los explica.

Los huecos no son azar: son leyes. Todos los códigos *geminados* (s·s, w·w, d·d, k·k, l·l, g·g) están vacíos —y su espejo también—: es la anti-geminación, el Principio de Contorno Obligatorio que prohíbe dos segmentos idénticos contiguos (Leben, 1973; McCarthy, 1986) y que, como restricción de la raíz, se remonta a la teoría de la raíz indoeuropea de Benveniste (1935) y Meillet (1937). Los huecos de dos oclusivas (g·t, k·g) reproducen la *ley de co-ocurrencia de oclusivas* de la raíz PIE, que veda ciertas combinaciones de sonoridad (Iverson & Salmons, 1992). Y los huecos sonorante+sonorante (l·r, l·n) reflejan la OCP extendida a sonorantes (Dolgopolsky, 1999). La tabla periódica de los códigos tiene la misma virtud que la de los elementos: sus vacíos están *estructurados*.

Más aún: las asimetrías del espejo codifican hechos históricos concretos. El hueco w·k está vacío pero su espejo k·w lo realizan seis ramas —porque la labiovelar *k^w es una *unidad* real del PIE, y *wk no—. El hueco t·s está vacío pero s·t aparece cuatro veces —la direccionalidad de los grupos con *s* *móvil* y los antiguos “grupos thorn”—. La matriz no solo dice qué falta: dice *en qué dirección* el sistema permite construir.

El uso predictivo (analogía Mendeleiev)

Igual que el hueco entre el silicio y el estaño anticipó el germanio, un hueco *no* explicado por restricción marca una casilla que una lengua indoeuropea perdida o fragmentaria *podría* haber ocupado. Distinguimos así dos clases de vacío: huecos-ley (prohibidos: nunca se llenarán) y huecos-vacantes (permitidos pero no atestiguados: candidatos a reconstrucción). Formalmente, para una lengua de la que solo tengamos fragmentos, la matriz de ocupación de la familia da una *distribución previa* sobre qué códigos esperar —un dato para reconstruir lo no documentado (anatolio, tocario, mesápico, frigio) por encima del azar.

4 Patrón 3 · Una filogenia desde los códigos

Si dos ramas comparten muchos códigos, están cerca. Medimos la distancia de Jaccard entre inventarios de códigos, $d(A, B) = 1 - |A \cap B| / |A \cup B|$, y tomamos el vecino más próximo de cada rama. Es el mismo principio —distancia léxica → árbol— de la glotocronología (Swadesh, 1952) y de la filogenia bayesiana del indoeuropeo (Gray & Atkinson, 2003; Bouckaert et al., 2012), pero calculado sobre *códigos consonánticos*, no sobre listas de palabras.

rama	vecino más próximo	¿mismo grupo?
Indo-Iranio	Balto-Eslavo	sí
Balto-Eslavo	Céltico	no
Armenio	Helénico	no
Albanés	Céltico	no
Germánico	Céltico	sí
Itálico	Céltico	sí
Céltico	Germánico	sí
Helénico	Balto-Eslavo	no

Table 4: Vecino más próximo por inventario de códigos. Recupera tres agrupaciones indoeuropeas reales: indo-iranio – balto-eslavo (satem), el núcleo centum del noroeste (germánico – céltico – itálico) y el greco-armenio.

El resultado es notable: solo con la *estructura consonántica* de las raíces reaparecen (i) el par indo-iranio – balto-eslavo, columna vertebral del satem; (ii) el “núcleo centum del noroeste” que reúne germánico, céltico e itálico (Ringe et al., 2002); y (iii) el vínculo greco-armenio, una de las subagrupaciones más discutidas del indoeuropeo (Clackson, 1994). Que tres hipótesis filogenéticas independientes emerjan del solo esqueleto consonántico es un argumento fuerte de que el código porta señal histórica.

5 Patrón 4 · La ley de decaimiento del código

Cada volumen ajustó la conservación del código en el tiempo a una curva de decaimiento $d(t) = d_{\infty} (1 - e^{-\lambda t})$, con su vida media $t_{1/2} = \ln 2 / \lambda$ —trasladando a los códigos el modelo exponencial que la glotocronología (Swadesh, 1952; Lees, 1953) tomó de la desintegración radiactiva—. Comparadas las ocho ramas, los códigos binarios son sistemáticamente más longevos que los ternarios: menos “masa” consonántica, menos superficie de erosión.

rama	binario $t_{1/2}$ (ka)	ternario $t_{1/2}$ (ka)
Indo-Iranio	2.67	1.82
Balto-Eslavo	1.82	0.23
Armenio	5.78	3.47
Albanés	6.93	4.95
Germánico	0.29	0.23
Itálico	0.23	0.23
Céltico	0.23	0.23
Helénico	1.02	3.47

Table 5: Vidas medias por rama y aridad (miles de años). Advertencia: en armenio y albanés el ajuste no es evaluable (sin pares ③ de sinónimo no-cognado); sus cifras se dan solo como referencia.

Entre las ramas con ajuste evaluable, el gradiente binario > ternario es constante; y las cifras absolutas sugieren —con cautela por el muestreo de etapas— ramas más conservadoras (indo-iranio, balto-eslavo) frente a otras de recambio rápido (itálico, céltico, germánico). Se ofrece como hipótesis cuantificada, no como resultado cerrado.

6 Patrón 5 · La especularidad recurrente

El programa registró *pares especulares*: códigos $c_1 \cdot c_2$ cuyo espejo $c_2 \cdot c_1$ también existe. Cruzando las ocho ramas, algunos espejos recurren en casi toda la familia.

par especular / mirror pair	ramas/branches
d·w / w·d	6
g·n / n·g	5
m·s / s·m	4
n·s / s·n	4
d·g / g·d	3
l·s / s·l	3
k·s / s·k	3
g·r / r·g	3
g·j / j·g	2
k·l / l·k	2
g·l / l·g	2
j·k / k·j	2

Table 6: Pares especulares presentes en dos o más ramas (número de ramas). Encabeza d·w / w·d —el código de **duo*- ‘dos’ y su espejo— presente en seis ramas.

La recurrencia de d·w / w·d en seis ramas (el numeral ‘dos’, **duo*-, y su reflejo) muestra que la especularidad no es un accidente de rama sino un rasgo pan-indoeuropeo del espacio de códigos. Con la cautela metodológica del programa: la especularidad es un hecho *de la forma*; su lectura semántica (la “inversión cuántica”) pertenece al registro hermenéutico, no al comparativo.

7 Síntesis e hipótesis derivadas

Cinco magnitudes de una familia. (1) La isoglosa satem/centum es un índice con $d \approx 2,1$ de separación. (2) La tabla periódica de los códigos tiene huecos *estructurados*: el 73 % los predice una restricción de la raíz PIE, y sus asimetrías codifican la unidad de **k*^w y la *s* móvil. (3) La distancia entre inventarios de códigos recupera tres subagrupaciones IE reales, incluido el greco-armenio. (4) El decaimiento del código es más lento cuanto menor es la aridez. (5) La especularidad d·w/w·d es pan-indoeuropea.

De aquí salen hipótesis comprobables, en la línea del pedido —“crear un tipo de hipótesis”, predecir lo perdido—:

H1 (predictiva). Para una lengua IE fragmentaria (frigio, tracio, mesápico, ilirio), la matriz de ocupación de la familia da una *previa* sobre sus códigos probables; un código atestiguado que caiga en un *hueco-ley* señala error de lectura o préstamo, no herencia.

H2 (cuantitativa). El índice satem debería ordenar también las lenguas de posición discutida (albanés, y en su día el anatolio) sobre un continuo, en vez de forzar un binario.

H3 (estructural). Los huecos-ley (geminados, oclusiva+oclusiva de sonoridad vetada) deben mantenerse vacíos al añadir cualquier rama nueva; su llenado falsaría la universalidad de la restricción de la raíz dentro del indoeuropeo.

H4 (evolutiva). Si el decaimiento binario > ternario es real, las lenguas de mayor profundidad temporal deberían mostrar códigos más largos conservados que las de recambio rápido; comprobable al añadir etapas intermedias.

Este es el *primer paso* comparativo con lo ya disponible. Las precisiones previstas —matriz de transición proto → proto entre ramas como operador algebraico, tabla periódica ternaria, e incorporación

de las ramas ultrafinas (anatolio, tocario) como test predictivo— quedan para la segunda iteración.

Reproducibilidad

`python3 src/cc/compare.py` lee las ocho `coderviv.<rama>.db` y escribe `docs/paper-cc-compare/data.json`; `gen_compare.py` genera las tablas de este documento. Índice satem por detección de palatovelares en la reconstrucción; tabla periódica sobre códigos binarios agrupados; distancia de Jaccard entre inventarios; decaimiento de los ajustes por rama. Datos primarios: IE-CoR.

References

- Benveniste, É. (1935). *Origines de la formation des noms en indo-européen*. Maisonneuve.
- Bouckaert, R. et al. (2012). Mapping the origins and expansion of the Indo-European language family. *Science* 337, 957–960.
- Brugmann, K. (1904). *Kurze vergleichende Grammatik der indogermanischen Sprachen*. Trübner.
- Clackson, J. (1994). *The Linguistic Relationship between Armenian and Greek*. Blackwell.
- Dolgopolsky, A. (1999). *The Nostratic Macrofamily and Linguistic Palaeontology*. McDonald Institute.
- Gray, R.D. & Atkinson, Q.D. (2003). Language-tree divergence times support the Anatolian theory. *Nature* 426, 435–439.
- Iverson, G.K. & Salmons, J.C. (1992). The phonology of the Proto-Indo-European root structure constraints. *Lingua* 87, 293–320.
- Kaplan, R.M. & Kay, M. (1994). Regular models of phonological rule systems. *Computational Linguistics* 20(3), 331–378.
- Kondrak, G. (2000). A new algorithm for the alignment of phonetic sequences. *Proc. NAACL*, 288–295.
- Leben, W. (1973). *Suprasegmental Phonology*. PhD thesis, MIT.
- Lees, R.B. (1953). The basis of glottochronology. *Language* 29, 113–127.
- List, J.-M. (2012). LexStat: automatic detection of cognates in multilingual wordlists. *Proc. LINGVIS/UNCLH*, 117–125.
- McCarthy, J.J. (1986). OCP effects: gemination and antigemination. *Linguistic Inquiry* 17, 207–263.
- Meillet, A. (1908). *Les dialectes indo-européens*. Champion.
- Meillet, A. (1937). *Introduction à l'étude comparative des langues indo-européennes*. 8e éd. Hachette.
- Osthoff, H. & Brugmann, K. (1878). *Morphologische Untersuchungen*, vol. I. Hirzel.
- Ringe, D., Warnow, T. & Taylor, A. (2002). Indo-European and computational cladistics. *Transactions of the Philological Society* 100, 59–129.
- Swadesh, M. (1952). Lexico-statistic dating of prehistoric ethnic contacts. *Proc. American Philosophical Society* 96, 452–463.