

Códigos consonánticos en 582 lenguas austronesias

Un catálogo de correspondencias forma–sentido construido sin genealogía ni reconstrucción

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

Resumen

Presentamos un catálogo de 10.994 códigos consonánticos que agrupan 84.884 formas léxicas de 582 lenguas austronesias. Un código es un esqueleto consonántico compartido por palabras del mismo concepto en lenguas distintas. Su construcción no usa etimología, protoformas ni juicios de cognación: solo la forma transcrita, el concepto que expresa y el requisito de que la coincidencia se dé en al menos tres lenguas.

El procedimiento agrupa el 39,1 % de las formas disponibles y su precisión supera en 41,6 veces lo esperable por azar. El 85,4 % de los códigos resultantes tiene tres consonantes o menos. Calibrado contra cognación experta indoeuropea en condiciones equivalentes, el método muestra precisión de 0,96 y recall de 0,32: encuentra una minoría fiable de las relaciones que existen.

Los códigos de mayor alcance reproducen el vocabulario panaustronesio que la lingüística comparada ya había establecido — $l \cdot m$ para «cinco» en 252 lenguas, $m \cdot t$ para «ojo» en 246, $t \cdot l \cdot n$ para «oreja» en 54— sin haber reconstruido nada. Sostenemos que este catálogo tiene tres usos: generar hipótesis genealógicas comprobables en familias donde la reconstrucción es escasa, producir un inventario de pares de sustitución sin alinear nada, y servir de base a un método de aprendizaje léxico apoyado en la predictibilidad entre lenguas emparentadas, que en austronesio alcanza una mediana del 50,6 % del vocabulario por lengua — cifra que, como se declara en §6.4, depende de la densidad de muestreo del corpus y no puede leerse como medida de divergencia.

1. Qué se busca, y por qué

Esta serie pregunta si se puede hacer algo útil renunciando a la reconstrucción, y el paper del método describe el instrumento que resulta: un catálogo de códigos consonánticos, más modesto y más barato que el método comparativo, que hace parte de su trabajo sin hacer sus afirmaciones.

1.1 Lo que este trabajo no intenta

Conviene decirlo antes que nada, porque el objeto es fácil de sobreinterpretar.

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

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

No distinguimos herencia de préstamo. Sin protoformas es imposible, y no vamos 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 atribuimos significado a un código. $l \cdot m$ no significa «cinco». Es lo que comparten las 252 lenguas que dicen cinco, bajo esas condiciones de información. Escribimos $l \cdot m | \text{five}$ y nunca $l \cdot m = \text{five}$.

1.2 Dónde encaja esto, y qué no es

El procedimiento no es nuevo. Reducir una palabra a sus consonantes e ignorar las vocales es el *consonant class matching* introducido por Dolgopolsky (1964) y formalizado por Turchin, Peiros y Gell-Mann (2010). Conviene no confundirlo con ASJP, que es otra cosa: ASJP transcribe a un alfabeto de 41 símbolos que conserva las vocales y compara con distancia de Levenshtein normalizada y corregida por azar (Brown, Holman, Wichmann y Velupillai 2008; Holman et al. 2008, 2011). La literatura de evaluación trata «Turchin» y «ASJP» como métodos distintos, y compara ambos con SCA y con LexStat (List 2012, 2014; List, Greenhill y Gray 2017, que además usa austronesio como banco de pruebas y B-cubed como métrica; Jäger 2018; Rama et al. 2018). Lo que aquí se hace es una variante del primero, con tres diferencias declaradas:

	Trabajo previo	Aquí
Longitud considerada	dos primeras consonantes de la raíz	esqueleto completo, aridad ilimitada
Salida	matriz de distancias entre lenguas	catálogo de códigos por concepto
Uso	clasificación filogenética	inventario consultable, sin clasificar nada

La primera diferencia es la sustantiva y es la que organiza §2.3 y §2.4: Dolgopolsky y Turchin et al. truncan a dos posiciones, mientras que aquí la aridad es el objeto de estudio y no un parámetro fijado. La segunda es el objeto del trabajo: no se busca un árbol, se busca una lista.

Una diferencia que no reclamamos. Sería cómodo decir que aquí no se usan clases de sonido y que por eso se agrupa con más precisión. No es cierto: la normalización canónica de §2.1 es una capa de clases — colapsa χ con x , y las africadas y postalveolares con s — y por eso se publica entera antes de cualquier resultado. La diferencia con SCA o con las clases de Dolgopolsky no es tenerlas o no tenerlas, sino que las nuestras son menos y están declaradas.

Y no es comparación multilateral. El perfil superficial de este trabajo —parecido de forma, glosa igual, tres lenguas, sin reconstrucción— coincide con el de la comparación multilateral que la disciplina rechazó, y conviene decir en qué se separa. Aquí el criterio es contable y falsable: cada agrupamiento se compara con una permutación que rompe la relación forma-sentido (§3.4), cada código lleva su enriquecimiento sobre la frecuencia base (§3.3), y no se propone ningún agrupamiento genealógico nuevo — ni entre lenguas ni entre familias. La comparación multilateral fallaba justamente por carecer de esos controles.

2. Qué es un código

2.1 El esqueleto consonántico

De cada palabra extraemos la secuencia de sus consonantes, conservando el orden y descartando las vocales. Malgache *mainity* «negro» tiene los segmentos m , $a\dot{i}$, $^n t$, i ; su esqueleto es $m \cdot n \cdot t$.

La normalización canónica agrupa cada consonante bajo un representante, y conviene tenerla delante porque decide qué códigos coinciden. Se deriva del mapeo en `ingest/recompute_skeleton.py`:

Se escriben	como	Se escriben	como
h ħ ʔ ʕ ʰ ʕ ʔ ʕ	h	n ŋ ɲ ɳ	n
l ʌ ʈ ʈ ʈ ʈ ʈ ʈ	l	r ɹ ʀ ʁ ʀ ʁ ʀ ʁ	r
x ɣ ʝ ʝ ʝ	x	g ʝ g	g
c ɟ	c	d ɗ	d · m ɱ → m
ʃ ʒ ʒ ʒ ʒ y toda africada (ts ɗz ʈʂ ɗʒ ʃʂ ʈʃ...)	s		

Cuatro consecuencias que el lector encontrará en los datos y que aquí quedan explicadas. La oclusiva glotal ʔ se escribe h (por eso *ʔepat* da h · p · t) y la nasal velar ŋ se escribe n (por eso *taliŋa* da t · l · n). Las consonantes prenasalizadas ocupan dos posiciones (ⁿt → n · t), que es la razón de que *mainty* salga ternario. Y —la de mayor alcance— las africadas se leen por su destino y se escriben s, junto con las sibilantes desplazadas: por eso m · s | star reúne *masoa* con *mat^soi* y *maʃfo*, y m · s | BE DEAD OR DIE reúne *mase* con *-meʃf*.

Esta tabla tiene dos costes que conviene ver antes que los resultados. El primero es que la afirmación «aquí no usamos clases de sonido» sería falsa, y por eso §1.2 no la hace. El segundo es que los pares que esta tabla fusiona son invisibles para §6.3: x ↔ ɣ no puede aparecer en el inventario de sustituciones porque ya está resuelto aquí, y la fila s ↔ t de esa sección se mide sobre un alfabeto en el que ʈʃ ya cuenta como s. La matriz de sustituciones mide, en parte, esta tabla.

Las vocales no se pierden: se guardan en una columna propia, junto a la plantilla silábica. La razón es práctica y se verá en §7.3: el análisis trabaja sobre el esqueleto, pero la segmentación morfológica que llevaría a la raíz necesita las vocales, porque los afijos austronesios son sílabas (ma–, ka–, –an) y en el esqueleto ma– es indistinguible de una m radical.

2.2 Un código es un esqueleto que se repite entre lenguas

Un código es un esqueleto consonántico que aparece en el mismo concepto en al menos tres lenguas distintas. Los dos requisitos importan:

- El mismo concepto, porque sin control semántico el agrupamiento colapsa. Lo verificamos: en un trabajo anterior sobre este mismo corpus, agrupar por forma compartida sin control de sentido produjo redes que encadenaban homógrafos hasta reunir palabras sin relación —96 redes con 25,9 lenguas de media y una cobertura de código desplomada al 13,1 %—. El concepto es lo que impide esa cadena.
- Al menos tres lenguas, porque dos formas de la misma lengua no son un hecho comparativo, y dos lenguas dan demasiadas coincidencias.

2.3 Códigos unarios, binarios y ternarios

La aridad de un código es su número de posiciones consonánticas.

Un código unario tiene una consonante: m, t, n. Un binario tiene dos: l · m, m · t, t · l. Un ternario, tres: b · l · n, t · l · n, k · l · t.

La aridad decide cuánto puede discriminar un código, y la razón es aritmética. Con un inventario de unas treinta consonantes, hay treinta códigos unarios posibles, novecientos binarios y veintisiete mil ternarios. Sobre 216.864 formas, eso significa que un código unario recae de media sobre miles de palabras y un ternario sobre unas pocas.

Los datos lo confirman con crudeza. El código unario n aparece en el austronesio asociado a 148 conceptos con tres lenguas o más. Una muestra, sin ninguna relación entre sí:

I, drink, he/she, this, sand, eat, mother, thou, fish, six, at, that, what, his/her, if, lie down, and, coconut, water, name, where, sleep, foot, chew, stick, tooth, mosquito, house, here, cry, blow, you, big, say...

Eso no es un código: es una consonante frecuente apareciendo por frecuencia. Medimos la precisión de cada aridad —dadas dos formas de lenguas distintas con el mismo código, ¿con qué frecuencia comparten concepto?— y el resultado es inequívoco:

Aridad	Valores distintos	Precisión	Veces por encima del azar
1 (unario)	28	2,66 %	3,7×
2 (binario)	555	6,81 %	14,7×
3 (ternario)	5.045	13,51 %	30,4×

El poder del método vive en los binarios y los ternarios. Los unarios se incluyen en el catálogo por completitud, pero no sostienen ninguna afirmación por sí solos.

Una salvedad sobre esta tabla. El 18,0 % de los códigos binarios y ternarios (1.418 de 7.881) repite alguna consonante: *s·s|breast*, *m·m|father*, *n·n|mother*, *l·l|tongue*. Son formas CVCV con la consonante repetida, sea por reduplicación —proceso gramatical productivo en austronesio—, por palabra de crianza o simplemente por la forma de la raíz: *s·s|breast* es PAN **susu*, monomorfémica, y *l·l|tongue* reúne el malgache *lela* con el polinesio *alelo*, que tampoco son reduplicación productiva. La causa varía; lo que importa aquí es la consecuencia, que es común a todas: su espacio de posibilidades es mucho menor que el de un binario con dos consonantes distintas, de modo que su probabilidad por azar es mayor. Están dentro de las cifras de arriba y no se han excluido; se declara para que quien quiera rehacer el cálculo sin ellos sepa cuántos son.

2.4 El criterio de aridad como control

Antes de generar el catálogo declaramos un criterio: los códigos deberían resultar unarios, binarios o ternarios, y la aparición de cuaternarios debía tomarse como motivo de duda sobre el propio análisis, no como un hallazgo. Un código de cuatro o cinco consonantes en una familia de palabras cortas sugiere que la palabra no se ha segmentado bien.

El resultado, sin haber filtrado nada:

Aridad	Códigos	%
1	1.509	13,7 %
2	4.797	43,6 %
3	3.084	28,1 %
4	1.200	10,9 %
5 o más	400	3,6 %

Figure 1: Reparto de los códigos por aridad. El 85,4 % tiene tres consonantes o menos.

El 85,4 % de los códigos cumple el criterio. El 14,6 % restante está en el catálogo, contado y accesible, que es lo que el criterio pedía: los cuaternarios no se ocultan, se marcan para revisión. Su origen probable es la falta de segmentación morfológica (§7.3).

3. Método

3.1 Datos

Corpus Integrativo, subconjunto austronesio: 582 lenguas, 216.864 formas con transcripción y concepto asignado. Una nota sobre ese 582, porque en el corpus conviven tres cifras y hay que decir cuál se usa: la tabla de lenguas tiene 586 filas y 584 nombres distintos —dos pares de filas son la misma lengua registrada dos veces—, y de esos 584 nombres, 582 aportan al menos una forma con esqueleto. Todas las cifras de este paper cuentan por nombre de lengua sobre esas 582, salvo donde se indique lo contrario. Las formas proceden de listas léxicas estandarizadas (Lexibank, IDS) bajo licencias CC-BY, y el concepto es un nodo de Concepticon, no una glosa libre. La cobertura de transcripción fonética es del 100 % y la de esqueleto del 93,8 %.

3.2 Construcción

Para cada concepto se agrupan las formas por esqueleto idéntico. Un grupo cuenta si reúne tres lenguas o más. No interviene ninguna otra información: ni familia, ni rama, ni geografía, ni etimología.

3.3 Enriquecimiento: separar el código de la coincidencia

Un esqueleto puede repetirse porque las palabras están relacionadas o porque está hecho de consonantes frecuentes. Para distinguirlo se compara lo observado con lo esperable:

$\text{enriquecimiento} = \text{observado} / \text{esperado}$, donde observado = lenguas del concepto c cuya forma tiene el esqueleto $e \div$ lenguas que tienen c ; esperado = formas de toda la familia con el esqueleto $e \div$ formas de toda la familia.

Un enriquecimiento de $1 \times$ significa que el esqueleto aparece en ese concepto tanto como aparece en cualquier sitio. La tabla que sigue usa una consonante en lugar de un esqueleto completo, porque el contraste se ve mejor así; el cálculo es el mismo con e de una sola posición.

El contraste es nítido:

Conjunto	Elemento	Observado	Base en la familia	Enriquecimiento
eight	w	41,2 %	2,3 %	17,8×
blood	d	25,1 %	2,2 %	11,4×
eye	m	77,5 %	15,5 %	5,0×
hundred	m	15,5 %	15,5 %	1,0×
buy	m	14,1 %	15,5 %	0,9×

Figure 2: Alcance frente a enriquecimiento. Cada punto es un código binario o ternario; el eje vertical es logarítmico. Un código puede reunir muchas lenguas y no decir nada si está hecho de consonantes frecuentes.

La misma consonante m es informativa en «ojo» y vacía en «comprar». En el primer caso aparece cinco veces más de lo esperable; en el segundo, menos que por azar. El enriquecimiento es lo que convierte una constancia en un hecho.

Pero no mide antigüedad, y conviene saberlo antes de leer las tablas. Un enriquecimiento alto dice que una secuencia rara se repite; nada más. En §6.2 se muestra que los códigos de mayor enriquecimiento

son sistemáticamente los más locales —conglomerados de lenguas vecinas— y que los candidatos a reconstrucción tienen enriquecimiento moderado. Ordenar por enriquecimiento no es ordenar por profundidad histórica.

3.4 Nulos

Todo resultado se compara con una permutación que rompe la relación forma–sentido conservando las distribuciones marginales. Concretamente: dentro de cada lengua se barajan los conceptos, de modo que cada lengua conserva su inventario de esqueletos y se destruye solo la correspondencia entre lenguas.

Esta elección no es cosmética. Un nulo más obvio —barajar los esqueletos dentro de cada concepto— produce un artefacto: el 40,7 % de los pares (lengua, concepto) tiene más de una forma, y en 32.495 casos esas formas repiten esqueleto dentro de la misma lengua. Al barajar, esos duplicados intralingüísticos se reparten entre lenguas distintas y fabrican grupos que no existían. Con ese nulo, el azar «agrupaba» más que los datos reales y el método parecía inservible. Lo dejamos anotado como advertencia metodológica: un nulo mal construido no produce ruido, produce un resultado invertido.

3.5 Calibración

El método se calibró contra IE-CoR (Heggarty et al. 2023), cognación indoeuropea establecida por expertos: 2.618 conjuntos, 22.554 formas, 170 conceptos, 152 lenguas. La calibración se corrió en condiciones equivalentes a las de este trabajo —usando solo forma, concepto y esqueleto, sin etimología ni protoformas— y con la métrica B-cubed, estándar en detección automática de cognados.

Criterio	Precisión	Recall	F1
Esqueleto idéntico	0,963	0,317	0,477
<i>nulo</i>	0,620	0,150	0,241

Por qué no se calibró en la propia familia. El austronesio tiene cognación experta publicada —la ABVD de Greenhill, Blust y Gray (2008), con juicios sobre unos 210 conceptos en más de mil lenguas— y calibrar contra ella habría sido preferible. No se hizo porque la ABVD no está en el Corpus Integrativo: el subconjunto austronesio del corpus no contiene ni un solo conjunto de cognados. Ingerirla es trabajo pendiente, y hasta entonces la extrapolación desde indoeuropeo es lo que hay, con el descuento que eso merece.

Cuando el método agrupa dos palabras, acierta el 96,3 % de las veces; encuentra el 31,7 % de las que existen. Ese perfil define su uso: no reconstruye la partición completa de un léxico, propone grupos fiables. Con una advertencia: IE-CoR es vocabulario básico indoeuropeo, el caso favorable. Fuera será peor.

Y una parte de ese 0,963 no es mérito del esqueleto. La calibración se corre dentro de conjuntos de concepto ya fijado, así que la restricción semántica está haciendo un trabajo que la cifra no separa. El reverso importa más: al exigir nodo idéntico de Concepticon, el método no puede encontrar ninguna relación que implique desplazamiento de sentido —del tipo *piojo* → *pulga*, o las cadenas *mano/cinco* y *ojo/cara*—, y ahí vive buena parte del trabajo etimológico. El recall de 0,32 incluye, por construcción, cero cognados con cambio de significado.

Un segundo resultado de la calibración explica el recall bajo y no es un defecto del método. Partiendo de los conjuntos de cognados verdaderos, solo el 12,8 % tiene esqueleto homogéneo, y el esqueleto más frecuente cubre de media el 52,9 % del conjunto. Los cognados reales divergen en la forma: el cambio fonético hace su trabajo. Lo que un código puede recuperar es la parte que todavía se parece.

4. Resultados

4.1 El catálogo

Códigos (≥ 3 lenguas)	10.994
Conceptos con al menos un código	543 de 1.823
Formas catalogadas	84.884 de 216.726 (39,2 %)
Lenguas	582
Binarios y ternarios	7.881 (71,7 %)
Enriquecimiento mediano (bin. y tern.)	14×
Lenguas por código (mediana)	4

Figure 3: Distribución del alcance: cuántos códigos comparten cuántas lenguas. La mediana es de cuatro lenguas por código.

El agrupamiento cubre el 39,1 % de las formas con una precisión 41,6 veces superior al azar (15,92 % frente a 0,383 % del nulo). La mediana de cuatro lenguas por código indica que la mayoría son pequeños y locales; los de gran alcance son una minoría diminuta.

4.2 Los códigos de mayor alcance

Concepto	Código	Aridad	Lenguas	Enriquecimiento
five	l·m	2	252	120×
eye	m·t	2	246	53×
three	t·l	2	241	86×
be dead or die	m·t	2	204	47×
breast	s·s	2	184	95×
bird	m·n	2	172	36×
louse	k·t	2	135	65×
father	t·m	2	105	76×
seven	p·t	2	101	92×
stone	b·t	2	98	113×
hand	l·m	2	93	41×
six	n·m	2	77	61×
eight	w·l	2	75	152×
moon	b·l·n	3	71	149×
sleep	m·t·r	3	54	90×
ear	t·l·n	3	54	78×
lightning	k·l·t	3	51	137×
nine	s·j·m	3	49	531×

Figure 4: Los veinte códigos binarios y ternarios de mayor alcance, por número de lenguas.

Estos códigos coinciden con el vocabulario panaustronesio establecido por la lingüística comparada. Lo señalamos como observación, no como validación: no se ha puntuado contra ninguna reconstrucción, porque el corpus no contiene proto-austronesio.

5. Ejemplos comentados

5.1 **l·m|five** — un código binario de máximo alcance

252 lenguas. Cobertura del 48,3 % de las que tienen la palabra; enriquecimiento 120×.

Lengua	Forma	Lengua	Forma
Puyuma (Formosa)	<i>lima</i>	Aklanon (Filipinas)	<i>lima</i>
Kalinga	<i>limá</i>	Tolomako (Vanuatu)	<i>lima</i>
Ambrym Norte	<i>lim</i>	Ifugao	<i>lema</i>
Naman	<i>iləm</i>	Rano	<i>i-lim</i>
Atta	<i>li:ma:</i>	Nulnesa	<i>lima</i>

El código **l·m** sobrevive a la variación vocálica (*lima*, *lema*, *lim*, *iləm*), a la pérdida de la vocal final y a la prefijación (*i-lim*). Eso es exactamente lo que un esqueleto consonántico hace: ignorar lo que varía rápido y conservar lo que varía despacio.

Obsérvese que *hand* «mano» tiene el mismo código **l·m** en 93 lenguas. Son dos conjuntos distintos —cada uno definido por su concepto— que resultan compartir código. El catálogo no los une, y hace bien: unirlos sería una hipótesis, no un dato. Pero los deja visibles el uno junto al otro, y esa vecindad es información. En austronesio esa coincidencia concreta es conocida —«cinco» y «mano» están relacionadas, cosa nada rara en sistemas numerales de base cinco— y el catálogo la expone junto a todas las demás coincidencias entre conceptos, sin afirmarla. Es un generador de preguntas, no de respuestas.

5.2 **t·l·n|ear** — un ternario a través de diez mil kilómetros

54 lenguas, enriquecimiento 78×.

Lengua	Forma	Región
Arta	<i>taliŋa</i>	Filipinas
Itawis	<i>taliŋa</i>	Filipinas
Bare'e	<i>taliŋa</i>	Célebes
Chamorro	<i>talaŋa</i>	Micronesia
Bali-Vitu	<i>taliŋa</i>	Nueva Bretaña
Kove	<i>taliŋa</i>	Nueva Guinea
Gitua	<i>taliŋa</i>	Nueva Guinea
Gedaged	<i>tiliŋa</i>	Nueva Guinea
Lou	<i>teŋiŋa</i>	Almirantazgo
Lote	<i>talŋa-</i>	Nueva Bretaña
Itbayaten	<i>tilwan</i>	Batanes
Label	<i>talŋa</i>	Nueva Irlanda
Nias	<i>taliŋa</i>	Sumatra
Pukapuka	<i>taliŋa</i>	Islas Cook

Un código ternario con tres posiciones que acertar, sostenido a lo largo de toda la dispersión austronesia. La variación vocálica es amplia (*a-i*, *a-a*, *i-i*, *e-i*) y el esqueleto no se mueve. En *Lote* y *Label* la vocal media desaparece (*talŋa*) y el código sigue intacto, porque el código no cuenta vocales.

Las dos últimas filas dan la medida: entre Nias, frente a Sumatra, y Pukapuka, en las Islas Cook, hay 10.755 km de océano, y la palabra es la misma. Ésa es la distancia que da título a la sección; las doce primeras filas, todas del arco Filipinas–Melanesia, se quedan en unos 4.200 km, y sin las dos últimas el título no se sostendría.

El mismo alcance —54 lenguas— tiene *m·t·r|sleep*, que llega a 12.682 km entre el sakalava de Madagascar y el mwerlap de Vanuatu. No son, por tanto, un par que separe alcance de distancia: los dos son de largo alcance y de larga distancia, y decir otra cosa sería fabricar un contraste que los datos no dan.

Y *m·t·r|sleep* trae su propia advertencia, que conviene poner aquí y no esconderla en los límites. El étimo es PMP *ma- + *tuduR: prefijo estativo más raíz. La *m* de ese código no es material de la raíz, es morfología — exactamente el defecto que §7.3 y la sección «Cualidades» del Anexo A identifican como el peor del método. El código funciona porque el prefijo se ha heredado junto con la raíz, no porque el instrumento haya sabido separarlos.

5.3 *s·j·m|nine* — el enriquecimiento más alto

49 lenguas, 531×. La secuencia *s·j·m* es rara en austronesio, así que su repetición en cuarenta y nueve lenguas para el mismo concepto es extraordinariamente improbable por azar. Compárese con un código de igual alcance formado por consonantes frecuentes, que se quedaría en 1-2×.

Formas: *siyam* (bikol, binukid, ibanag, dupanangan), *siyám* (alangan, cuyonon, hanunóo), *syam* (aklanon, bolinao), *si'yam* (balangaw).

Este es el caso que mejor ilustra por qué el enriquecimiento es imprescindible. Un código con 49 lenguas puede ser trivial o extraordinario según de qué esté hecho, y solo la comparación con su frecuencia base lo distingue.

Y también ilustra su límite, que §6.2 mide — hasta el punto de invertir la lectura del caso. Las lenguas de este código no son «todas filipinas» por casualidad de la muestra: lo son todas, y la razón es conocida. El proto-austronesio para «nueve» es *Siwa; *siam/*siyam es una innovación de nivel filipino, no un étimo heredado del nodo profundo. Es decir: el enriquecimiento más alto de todo el catálogo corresponde a una innovación de un solo subgrupo. Lejos de debilitar el resultado, es la mejor ilustración posible del criterio de §6.2 — enriquecimiento alto no significa antigüedad, significa que una secuencia rara se repite entre vecinos— y por eso el ejemplo se mantiene en su sitio en vez de sustituirse por otro más cómodo.

Una dependencia de transcripción, que este código enseña mejor que ningún otro. *s·j·m* existe porque la fuente escribe *siyam* con glide; escrito *siam*, el mismo étimo daría *s·m*, un binario banal de enriquecimiento bajo. Nueve de las dieciséis formas del bloque del Anexo A llevan la *y*, tres no. El código no es más frágil que otros: es que en este el punto de fractura queda a la vista.

5.4 *m·t* — un código que agrupa de más, y qué se aprende de ello

El código binario *m·t* aparece asociado a varios conceptos: *eye* (246 lenguas), *be dead or die* (204), *snake* (95), *right* (71), *black* (39). Merece examen, porque no todos son el mismo fenómeno.

«Ojo» y «morir» son dos étimos distintos que convergen en el mismo esqueleto. Es homofonía de código, y el catálogo la refleja correctamente: son dos conjuntos separados, cada uno con su concepto.

«Negro» es otra cosa. Sus formas son *maeto*, *mainity*, *mainte*, *maitem*, *maitum*, *metan*, *maʔitim*. Comparadas con las de «ojo» —*mata*, *matá*, *mata*— se ve que no son paralelas:

Concepto	Forma	Esqueleto	Análisis
eye	<i>mata</i>	<i>m·t</i>	raíz completa
black	<i>maitem</i>	<i>m·t·m</i>	prefijo + raíz + nasal final
black	<i>metan</i>	<i>m·t·n</i>	prefijo + raíz + nasal final
black	<i>maʔitim</i>	<i>m·h·t·m</i>	con la glotal conservada

Concepto	Forma	Esqueleto	Análisis
black	<i>mainty</i>	m · n · t	con nasal prenasalizada

En «negro», la *m* inicial no es parte de la raíz: es un prefijo. Y el esqueleto no puede distinguirlo, porque *ma-* y *m* radical son ambos *m*. Solo las lenguas que conservan más material —*maitem*, *maʔitim*— lo revelan.

Esto es un límite real del método y conviene no disimularlo. Un código puede fundir un prefijo con una raíz. La solución sería la segmentación morfológica, que no está resuelta (§7.3). Mientras tanto, el catálogo lo señala: los conjuntos de «negro» aparecen repartidos entre *m · t*, *m · t · m*, *m · t · n*, *m · h · t · m* y *m · n · t*, y esa dispersión es visible y es información.

5.5 Un caso de variación regular: *t · l* y *t · r* para «tres»

«Tres» aparece con dos códigos: *t · l* en 241 lenguas y *t · r* en 43. No son dos hechos independientes: son la alternancia *l/r*, una de las correspondencias más regulares del austronesio.

Nuestro análisis de sustituciones (§6.3) la detecta como el par más frecuente de toda la familia: *l* ↔ *r* con 4.040 casos en el conjunto de validación, 3,3 veces por encima de lo esperable. Si se tratan *l* y *r* como equivalentes, los dos códigos se funden en uno que cubre 284 lenguas (los dos conjuntos son disjuntos: ninguna lengua aparece en ambos).

6. Qué se puede hacer con esto

6.1 Generar hipótesis genealógicas comprobables

El catálogo no afirma parentesco. Pero produce candidatos ordenados por fuerza, y eso es precisamente lo que un programa de reconstrucción necesita para saber por dónde empezar.

El caso más claro es el de los códigos que cruzan la división primaria de la familia. Conviene enunciar bien esa división, porque de ella depende que el procedimiento signifique algo. La clasificación establecida (Blust 2009, 2013) no divide el austronesio en dos, sino en una decena de subgrupos primarios: nueve formosanos y el malayo-polinesio. «Formosano» es una etiqueta geográfica, no un clado — y ésta es precisamente la razón por la que contar ramas formosanas funciona: si dos lenguas de Taiwán pertenecen a ramas primarias distintas, su ancestro común ya es el proto-austronesio. Un étimo presente en varias ramas formosanas es, por tanto, el candidato clásico a reconstrucción para ese nodo.

Nuestro catálogo los identifica sin conocer esa regla. Los denominadores, que mandan sobre todo lo demás: de esos nueve subgrupos formosanos, el corpus contiene cinco —Atayálico, Formosano Oriental, Formosano Noroccidental, Tsóuico y Austronesio de las Llanuras Occidentales—, así que «cinco ramas» significa *todas las que hay aquí*, no todas las que existen. 46 códigos binarios o ternarios cruzan tres ramas o más. La tabla muestra los cinco que cruzan cuatro o más —la lista completa por arriba— y, de los que cruzan tres, los que alcanzan 40 lenguas, para que las filas grandes sean comparables entre sí:

Concepto	Código	Ramas formosanas	Lenguas formosanas	Malayo-polinesias	Otras	Total
three	t · r	5	9	32	2	43
four	s · p · t	5	12	0	3	15
seven	p · t	4	11	86	4	101
five	r · m	4	6	28	3	37
I	j · k	4	5	4	0	9
five	l · m	3	6	243	3	252
three	t · l	3	6	232	3	241
father	t · m	3	4	99	2	105
stone	b · t	3	5	92	1	98
six	n · m	3	6	67	4	77
four	p · t	3	5	64	3	72

La segunda fila es el mejor resultado de la tabla, y merece leerse despacio. s · p · t | four cruza las cinco ramas formosanas del corpus y tiene cero lenguas malayo-polinesias. El catálogo no sabe por qué; la disciplina sí: el étimo es PAN *Sepat, esa s inicial es el reflejo de *S, y el proto-malayo-polinesio pierde *S — de ahí PMP *epat, que en nuestro catálogo son las otras dos entradas de «cuatro», h · p · t (36 lenguas, con la glotal inicial) y p · t (72). Es decir: el procedimiento separó por sí solo el nivel PAN del nivel PMP en el numeral, usando uno de los fonemas diagnósticos con que se define esa división, y sin conocer ni el fonema ni la división. Que los numerales sean además el dominio desde el que Sagart (2004) argumenta el subagrupamiento formosano hace la coincidencia más útil todavía: hay con qué contrastar.

La tabla se restringe a códigos binarios y ternarios, y esa restricción es una regla, no una limpieza. §2.3 estableció que los unarios no sostienen ninguna afirmación por sí solos; admitirlos aquí habría producido, entre otros, n | mother como candidato a reconstrucción — un par que §2.3 usa como ejemplo canónico de ruido, porque ese mismo código unario aparece con 148 conceptos distintos. Un criterio que se declara en una sección tiene que gobernar las demás, y la columna «Otras» —lenguas sin rama asignada en el corpus— está para que el lector pueda comprobar que las cifras suman.

Estos son hipótesis, no conclusiones: el catálogo dice que hay una correspondencia forma–sentido que atraviesa la división más profunda de la familia, y que por tanto merece que alguien la examine con el método comparativo. No dice que sea herencia. Podría ser préstamo antiguo, y de hecho los numerales son notoriamente prestables.

Pero la lista es corta, está ordenada por fuerza y es reproducible. Para una familia sin reconstrucción, eso es la diferencia entre saber por dónde empezar y no saberlo.

Obsérvese que la lista se ordena por número de ramas formosanas, no por enriquecimiento. La razón está en §6.2: el enriquecimiento premia lo local, y ordenar por él pondría arriba justamente los peores candidatos.

Y un uso complementario, que quizá sea más valioso: la lista de lo que no encaja. De los 10.994 códigos, el 96,8 % no sale de una sola macroárea. Las macroáreas reales son tres —Papunesia (561 lenguas), África (22) y Eurasia (1)— y dos lenguas del corpus no tienen ninguna asignada; se excluyen del recuento, y la cifra no cambia. Los 349 que sí cruzan (3,2 %) son la lista corta de anomalías. Su desglose dice más que el porcentaje:

Macroáreas que cruza	Códigos
África ↔ Papunesia	298
Eurasia ↔ Papunesia	48
las tres	2
África ↔ Eurasia	1

Es decir: la lista de anomalías es, en un 85 %, el malgache, la lengua austronesia de Madagascar cuyo

pariente próximo está en Borneo. El instrumento no sabe eso y lo encuentra igual. Con solo tres macroáreas el listón es bajo, así que la cifra vale por la lista corta que produce, no por el porcentaje en sí; pero esa lista corta contiene el caso de dispersión de larga distancia más famoso de la familia, y sin el catálogo no habría manera de encontrarlo entre 216.864 formas.

6.2 ¿Coinciden los códigos con el parentesco de las lenguas que agrupan?

La pregunta se puede medir, y su respuesta obliga a corregir cómo se lee el enriquecimiento.

Aquí se usa la distancia geográfica entre lenguas, y hay que decir con franqueza que es un proxy y por qué se aceptó. La razón que dábamos —que la etiqueta de Glottolog no discrimina, porque 567 de las 582 lenguas caen en Malayo-Polinesio— solo vale para el nivel primario. Glottolog no da una etiqueta por lengua: da una ruta de clasificación anidada de quince o más niveles en la parte oceánica, y de ahí se obtiene la medida estándar, que es la profundidad del nodo común más reciente. Para austronesio existe además una filogenia publicada con longitudes de rama (Gray, Drummond y Greenhill 2009). Ninguna de las dos se ha usado aquí, y eso es una carencia de este trabajo, no una propiedad de la familia. §6.2 mide proximidad geográfica; la versión con profundidad de nodo está pendiente.

Y el proxy es especialmente malo *en esta familia*. El austronesio es el caso de manual de dispersión de larga distancia: el malgache está a unos 6.500 km de su pariente próximo en Borneo (Dahl 1951; Adelaar 1995), los *outliers* polinesios están incrustados en Melanesia, el chamorro y el palauano son ramas malayo-polinesias primarias en Micronesia. El propio §5.2 celebra un código que une Madagascar con el Pacífico. Es decir: hemos elegido como proxy de parentesco la variable que esta familia falsifica en sus casos más citados. Los resultados que siguen se sostienen —la señal es grande y va en la dirección predicha—, pero léanse como cercanía, no como parentesco, y cuéntese con que la versión con árbol podría mover las cifras.

Las lenguas que comparten más códigos están más cerca, y la relación es monótona:

Códigos compartidos	Pares	Distancia media	÷ azar
1	64.297	3.373 km	0,94
2	22.447	2.950 km	0,82
3–5	16.505	2.341 km	0,65
6–10	4.853	891 km	0,25
11–20	2.744	553 km	0,15
<i>dos lenguas al azar</i>	40.000	3.582 km	1,00

En promedio, los grupos que forma cada código son más compactos de lo esperable para su tamaño: entre 0,40 y 0,65 veces la dispersión de un grupo aleatorio de igual número de lenguas. El código capta estructura real. Lo que no puede decir es de qué tipo: herencia y contacto predicen exactamente lo mismo aquí.

Ese promedio, sin embargo, esconde lo interesante, y las dos tablas que siguen lo sacan: los códigos más enriquecidos bajan a 0,24 y los candidatos a reconstrucción suben por encima de 1,00, es decir, se dispersan más que el azar. La compacidad media es una descripción del montón, no de lo que importa.

El enriquecimiento mide localidad, no antigüedad

Al cruzar las dos magnitudes aparece una relación negativa ($r = -0,25$). Aquí cambia la población y hay que decirlo: las dos tablas que siguen se calculan sobre los 1.563 códigos binarios y ternarios presentes en ocho lenguas o más, no sobre los 7.881 del resto del paper, porque por debajo de ocho la

dispersión geográfica de un grupo es demasiado ruidosa para compararla. Los cortes de la tabla son los cuartiles de esa población:

Enriquecimiento	Dispersión ÷ azar
2–8×	0,60
8–21×	0,54
21–71×	0,41
71–2208×	0,24

Y en los extremos es inequívoco:

	Código	Lenguas	Ramas formosanas	Enriquecimiento	Dispersión
Máximo de esta población	‘g·l·s	ice’	9	0	2208×
0,17					
	‘x·m·n	beard’	9	0	1468×
0,06					
	‘p·p·s	turmeric’	10	0	700×
0,07					
Candidatos de §6.1	‘t·r	three’	43	5	20×
1,25					
	‘r·m	five’	37	4	39×
1,21					
	‘b·t	stone’	98	3	113×
0,26					

(El máximo de todo el catálogo, sin el filtro de ocho lenguas, es de 13.554× y corresponde a códigos de tres o cuatro lenguas, que es justamente el rango que ese filtro excluye por ruidoso.)

Los códigos de mayor enriquecimiento no tienen ni una rama formosana y forman conglomerados locales. Y el caso extremo lo dice mejor que la tabla: **g·l·s|ice**, con 2208×, son doce variedades malgaches diciendo *gilasy*, *glasy*, *glatsy* — el francés *glace*, un préstamo colonial de hace un siglo. El segundo, **x·m·n|beard**, son nueve lenguas de Nueva Georgia con el mismo *yumi-na*, posesivo incluido. Nótese además que «hielo» no es vocabulario básico ni heredable en lenguas tropicales: el propio concepto ya avisaba.

Es decir: el código más enriquecido de toda la familia austronesia es un galicismo. Los candidatos a reconstrucción de §6.1, en cambio, tienen enriquecimiento moderado y están más dispersos que el azar.

Esto corrige una lectura que el resto del paper favorece. El enriquecimiento se presenta en §3.3 como lo que separa un código de una coincidencia, y lo es; pero un enriquecimiento alto no indica profundidad histórica. Indica que una secuencia rara se repite, y una secuencia rara repetida entre vecinos es exactamente lo que produce un préstamo areal o una innovación reciente. Ordenar candidatos a reconstrucción por enriquecimiento sería ordenarlos casi al revés.

Dos ejes, y hacen falta los dos

De aquí sale un criterio de lectura que el catálogo no tenía:

	Compacto	Disperso
Enriquecimiento alto	innovación o préstamo local	<i>raro; merece examen</i>
Enriquecimiento moderado	poco informativo	candidato a retención antigua

Advertencia sobre el eje geográfico. Mide kilómetros, no profundidad temporal, y los dos ejes pueden contradecirse: $s \cdot p \cdot t \mid \text{four}$ cruza las cinco ramas formosanas del corpus —empatado con $t \cdot r \mid \text{three}$ en lo más alto de §6.1— y sin embargo sale compacto ($0,04 \times$), porque doce de sus quince lenguas están en Taiwán, que es una isla pequeña. Un código puede ser antiguo y geográficamente concentrado si su familia lo es. Es el contraejemplo más limpio a la lectura de esta sección, y viene del mismo catálogo.

Dos advertencias más, que afectan a todas las cifras de §6.2. La primera es el problema de Galton, que §6.4 enuncia y que aquí opera igual: la compacidad de un grupo depende de dónde el corpus es denso, y este corpus tiene quince variedades malgaches y decenas de lectos de Vanuatu. La segunda es la lista fuente: los datos vienen de decenas de encuestas distintas, y dos lenguas de la misma encuesta comparten inventario de conceptos, convención de transcripción y vecindad geográfica a la vez. Parte de «compartir códigos → estar cerca» es «venir de la misma fuente». Ninguna de las dos está controlada aquí, y la corrección correcta es conocida: el nulo debería muestrear de las lenguas que atestiguan ese concepto, estratificando por fuente, en lugar de las 582; y la pregunta genealógica se responde con un test de Mantel parcial —código contra geografía, controlando distancia en el árbol, y al revés—. Queda pendiente.

6.3 Pares de sustitución, descubiertos sin alinear nada

Dentro de cada conjunto se pueden tomar los pares de lenguas cuyos esqueletos tienen la misma longitud y difieren en exactamente una posición. Esa diferencia es una sustitución ya alineada.

ROPE	Taiof	afon	→ f·n	Tolai	vinau	→ v·n	⇒ f ↔ v
HAIR	Emae	furu	→ f·r	Nalik	vur	→ v·r	⇒ f ↔ v

Acumuladas sobre la familia (206.856 sustituciones limpias) y validadas en una mitad de lenguas que no participó en el descubrimiento, las más frecuentes son (los recuentos son de la mitad de validación):

No son «correspondencias» en el sentido técnico. Una correspondencia es regular, direccional y condicionada por el entorno, entre dos lenguas concretas. Esto es una matriz de sustitución no ordenada y agregada sobre 582 lenguas, sin entorno ni dirección. Frecuencia no es regularidad. Lo que produce son candidatos: el punto de partida de un análisis de correspondencias, no su resultado.

Par	Casos	Veces sobre el azar	Lectura (<i>glosa nuestra, no resultado</i>)
f ↔ v	1.073	13,9×	sonoridad labial
v ↔ β	438	13,4×	labial
p ↔ φ	133	11,6×	labial
s ↔ z	466	8,0×	sonoridad sibilante
f ↔ p	897	7,1×	labial
k ↔ x	1.478	5,1×	espirantización dorsal
f ↔ h	1.029	4,8×	debucalización *p > f > h
l ↔ r	4.040	3,3×	líquidas
s ↔ t	3.321	2,7×	palatalización *t > s

De los pares a las clases. Si se toman los pares que aguantan en la mitad de validación y se cierran transitivamente, la estructura que aparece depende del umbral, y el barrido entero es más informativo que cualquiera de sus filas:

Umbral	Clases que salen
$\geq 2\times$	{b c d f g h j k l n p r s t v w x z $\emptyset$ β } — todo encadenado
$\geq 3\times$	{b f g h j k p s t v w x z $\emptyset$ β } · {l r}
$\geq 4\times$	{b f h p v w $\emptyset$ β } · {k s x z}
$\geq 6\times$	{b f p v $\emptyset$ β } · {s z}
$\geq 10\times$	{f p v $\emptyset$ β }

Las tres clases que el Anexo A escribe como 1, 2 y 3 no son la salida de un solo umbral, y conviene no fingir que lo son:

Segmentos		Sale a
1	b f h p v w $\emptyset$ β	labiales, más h $\geq 4\times$
2	k s x z	dorsales y sibilantes $\geq 4\times$
3	l r	líquidas $\geq 3\times$

Las líquidas se desprenden antes que nada ($a \geq 3\times$), pero a ese umbral las otras dos todavía forman un solo bloque indiferenciado; a $\geq 4\times$ ese bloque se parte limpiamente en dos y $l \leftrightarrow r$ —que está a $3,3\times$, como se ve en la tabla de arriba— ya no aparece. El trío publicado es, por tanto, una lectura del barrido completo, no una ejecución. Decirlo así cuesta un párrafo; declarar «umbral de $4\times$ » habría producido una tabla que la fila $l \leftrightarrow r$ de esta misma sección desmiente.

Ninguno de estos pares se buscó. Salieron de contar diferencias en pares alineados. Y son de los que cualquier descripción del austronesio menciona, incluida la debucalización $*p > f > h$ que da el hawaiano.

Pero hay que decir qué no puede ver este procedimiento, porque es justamente donde vive el problema clásico de la familia. El método toma pares de esqueletos de la misma longitud que difieren en una posición; por construcción, no puede ver ningún reflejo cero. Y los conjuntos de correspondencias más citados del austronesio son precisamente los que incluyen $\emptyset$: PAN *R, cuyos reflejos son $r, l, g, y, h, \emptyset$ según la rama; PAN *q ($?, h, k, \emptyset$); y el mismo PAN *S ($s, h, \emptyset$) que produce el $s \cdot p \cdot t$ | four de §6.1. Ninguna de las tres clases emergentes captura *R ni *q. Dicho de otro modo: estos pares son los que cualquier descripción menciona en el capítulo de fonética; los que menciona en el capítulo de subagrupamiento son otros, y este instrumento es ciego a ellos. A eso se suma el aviso de §2.1: los pares que la normalización canónica ya fusionó — $x \leftrightarrow \text{ɣ}, s \leftrightarrow \text{ʃ}, s \leftrightarrow \text{tʃ}$ — tampoco pueden aparecer aquí.

Con esas dos cegueras declaradas, para una familia sin descripción histórica este procedimiento produciría un primer inventario de pares candidatos en cuestión de minutos.

6.4 Aprendizaje de lenguas

Este es el uso que menos esperábamos y el que tiene cifras más favorables.

Si un estudiante conoce el código de un concepto, conoce la forma aproximada de esa palabra en todas las lenguas del grupo. Aprender $t \cdot l \cdot n$ | ear da acceso simultáneo a *taliŋa*, *talaŋa*, *tiliŋa*, *teŋiŋa*, *talŋa* y *tilwan* en doce lenguas de ocho regiones. No es la palabra exacta —faltan las vocales— pero es el armazón, y el armazón es lo que se conserva.

Medimos qué proporción del vocabulario de cada lengua tiene código compartido con otras:

Lengua	Entradas	Con código compartido	%
Merina (Madagascar)	619	592	95,6 %

Lengua	Entradas	Con código compartido	%
Masikoro	204	194	95,1 %
Penrhyn (Polinesia)	384	364	94,8 %
Vezo	613	570	93,0 %
Betsileo	817	754	92,3 %
...			<i>mediana 50,6 %</i>
Tomoip	749	105	14,0 %
Atayal (Formosa)	1.144	132	11,5 %

La mediana es del 50,6 %: la mitad del vocabulario básico de una lengua austronesia media es predecible desde el código si se conocen otras lenguas de la familia. En el grupo malgache, que es muy cohesionado, supera el 92 %. La mediana está sujeta a la misma advertencia que los extremos, y el párrafo siguiente la enuncia: es una mediana sobre una muestra desequilibrada, no sobre la familia.

Los extremos bajos parecen igual de informativos —el atayal, formosano, se queda en el 11,5 %— pero esa lectura no se sostiene sin un control que aquí falta. El corpus contiene una quincena de variedades malgaches (merina, betsileo, bara, antanosy, antankarana, masikoro, vezo, sakalava...) y ninguna lengua hermana cercana del atayal. Las lenguas se cuentan como si fueran observaciones independientes y no lo son: es el problema de Galton. Así que el 95,6 % frente al 11,5 % mide, en primera instancia, cuántos parientes próximos tiene cada lengua en este corpus, y solo después algo sobre divergencia. Llamarlo «medida de distancia» exigiría controlar por densidad de muestreo, y no se ha hecho. La advertencia vale para las tres cifras, mediana incluida: si el 95,6 % no puede leerse como cohesión ni el 11,5 % como divergencia, el 50,6 % tampoco puede leerse como una propiedad del austronesio. Lo que sí sostiene es la afirmación pedagógica, que es más modesta y no necesita el control: para un hablante de una lengua concreta con parientes en el corpus, esa proporción de su vocabulario es predecible desde el código, y eso es cierto por muy sesgada que esté la muestra.

Para un método de enseñanza esto sugiere un orden: enseñar primero los códigos de mayor alcance —los que cubren doscientas lenguas— y después las reglas de realización de cada lengua concreta. El estudiante aprende una vez y aplica muchas veces.

7. Límites

7.1 Lo que el catálogo no puede decir

Herencia o préstamo. No hay manera de distinguirlos sin protoformas, y en austronesio los estratos de préstamo son masivos y conocidos: el malayo como donante en todo el sudeste asiático insular, el sánscrito y el árabe sobre él, y el malayo comercial en Filipinas oriental y Melanesia. Varios códigos del catálogo caen en esa zona de debate —b·l·j|house (*balay/balai*), d·g·m|needle, s·l|path— y el catálogo los contiene sin marcar. Los numerales del §6.1 son el caso típico: cruzan la familia entera y son también notoriamente prestables.

Antigüedad. El catálogo no data nada.

Parentesco entre familias. El espacio de códigos es pequeño y las coincidencias entre familias son esperables. Comparar códigos entre familias exige medir antes la equivalencia, y no se ha hecho.

Significado. Un código no significa. Agrupa.

7.2 El sesgo de la fuente

Las formas proceden de listas léxicas estandarizadas. No son listas de vocabulario básico, y decirlo así sería inexacto: el corpus cubre 1.823 conceptos, entre ellos «hielo», «cúrcuma» y «aguja de coser». Son decenas de encuestas de campo distintas, cada una con su recorte de conceptos, su transcriptor y su ámbito regional; el núcleo de todas ellas sí es léxico básico, pero la cola no. Eso tiene dos consecuencias. El sesgo hacia el léxico conservador existe y no dice nada del resto de la lengua: la cifra de 39,1 % de cobertura se refiere a ese subconjunto, no al léxico austronesio. Y la procedencia de cada lista es una variable de confusión no controlada: dos lenguas de la misma encuesta comparten conceptos, convención de transcripción y vecindad geográfica a la vez, lo que afecta directamente a §6.2, donde queda declarado.

7.3 La segmentación morfológica, sin resolver

Es la limitación más seria. El análisis trabaja sobre palabras enteras: no se ha separado la raíz de los afijos. En una familia con la afijación y la reduplicación del austronesio, eso tiene dos consecuencias: Subestima el agrupamiento. Dos formas de la misma raíz con afijos distintos no se reconocen. El 39,1 % es un suelo.

Y funde prefijos con raíces, como se vio en «negro» (§5.4).

Intentamos resolverlo con un detector distribucional de afijos y fracasó de forma instructiva. Aplicado al esqueleto, en hawaiano marcó siete de las ocho consonantes de la lengua como prefijo y como sufijo: no había descubierto morfología, había descubierto el inventario. La causa es que el criterio de separabilidad —«se puede quitar y el resto existe por su cuenta»— se satura en lenguas de inventario pequeño y sílaba CV.

Repetido sobre la forma con vocales, el mismo detector encontró morfemas reales: *te-* (artículo definido maorí), *-ŋa* (nominalizador, 154 casos), *ka-/ke-/nā-* (artículos hawaianos). El despiece necesita las vocales que el código descarta, y ese es el camino del trabajo siguiente.

Y hay una segunda razón para no usarlo, que solo se ve mirando el nulo. El detector sube el agrupamiento en bruto —de 62,2 % a 77,2 %— y esa cifra, sola, parece una ganancia grande. Pero el mismo recorte que junta las palabras emparentadas junta también las que no lo están, así que el nulo sube con ella, de 39,0 % a 70,4 %:

Paso	En grupo	Nulo	Ventaja	κ
Palabra entera	62,2 %	39,0 %	23,2	0,380
+ reduplicación	64,8 %	—	—	—
+ afijos descubiertos	77,2 %	70,4 %	6,8	0,230

La ventaja sobre el azar cae de 23,2 puntos a 6,8, y el agrupamiento pasa a ser apenas un 10 % mejor que barajar los conceptos ($1,60\times \rightarrow 1,10\times$). La columna κ es la corrección de Hubert-Arabie, ($\text{obs} - \text{nulo} / (1 - \text{nulo})$), que normaliza por el techo alcanzable y permite comparar entre familias: $\Delta\kappa = -0,151$.

(La primera versión de esta tabla llamaba «palabra entera» al paso que ya había quitado la reduplicación —64,8 %— y comparaba desde ahí. Los otros siete papers de la serie parten del paso anterior, así que la cifra de esta familia no era comparable con las suyas. Corregido el 2026-09-09; la fila de la reduplicación se publica aparte, como en el paper indoeuropeo.)

Es la mayor caída de las ocho familias medidas, y va en la misma dirección en todas ellas —turco $-0,009$, afroasiático $-0,004$, sino-tibetano $-0,020$, urálico $-0,035$, indoeuropeo $-0,094$, pamañungano $-0,102$, trans-neoguineano $-0,118$ —. El catálogo de este paper no usa el despiece

automático, y ésta es la medición que lo justifica.

Lo que la serie completa no confirma es la explicación que este párrafo daba antes: que «cuanto mayor es la ganancia bruta, más negativa es la real». El sino-tibetano, con la palabra más corta de las ocho, cae $-0,020$, menos que el indoeuropeo. Ver §5.1 del paper sino-tibetano.

7.4 Onomatopeyas

Algunos códigos de enriquecimiento altísimo son onomatopéyicos: «gallina» = $k \cdot k \cdot r \cdot k$, del tipo *kokorako*. El método los agrupa con razón —comparten forma y sentido— pero no son herencia. El catálogo no los distingue.

7.5 Una familia

Todo lo anterior es austronesio. Que el procedimiento funcione aquí no garantiza que funcione en familias con otra tipología. Medimos el perfil de forma de trece familias y las diferencias son grandes: el indoeuropeo tiene 4,65 consonantes por palabra y el sino-tibetano 2,22. En familias de palabra corta y tono alto —sino-tibetano, tai-kadai— un esqueleto de dos consonantes discrimina muy poco, y además entre el 33 % y el 45 % de sus segmentos son tonales y quedan fuera del esqueleto. Para esas familias este instrumento es insuficiente, y decirlo forma parte del resultado.

8. Conclusión

Renunciando a la reconstrucción se puede construir un objeto útil. El catálogo de códigos consonánticos agrupa el 39,1 % de las formas austronesias con una precisión 41,6 veces superior al azar, reproduce el vocabulario panaustronesio conocido sin haberlo buscado, y produce tres cosas que para esta familia no existían en forma consultable: una lista ordenada de candidatos a reconstrucción, un inventario de pares de sustitución descubierto sin alinear nada —candidatos a correspondencia, no correspondencias, en el sentido que §6.3 delimita— y una medida de predictibilidad léxica entre lenguas que puede fundamentar un método de enseñanza.

Lo hace afirmando muy poco. Un código no es una protoforma, no implica parentesco, no distingue herencia de préstamo y no significa nada. Es una correspondencia entre forma y sentido que se repite entre lenguas, con su alcance y su enriquecimiento medidos.

Sostenemos que esa modestia es la condición de su utilidad. Un instrumento que afirma poco puede aplicarse donde no hay bibliografía que lo respalde. Con el alcance que §7.5 delimita: en familias de palabra larga y sin tono, que es donde el perfil de forma lo permite — no en todas.

Referencias

- Adelaar, K. A. (1995). Borneo as a cross-roads for comparative Austronesian linguistics. En P. Bellwood, J. Fox y D. Tryon (eds.), *The Austronesians*. ANU. — [origen barito del malgache]
- Aikhenvald, A. Y. y Dixon, R. M. W. (eds.) (2001). *Areal Diffusion and Genetic Inheritance*. Oxford University Press.
- Bagga, A. y Baldwin, B. (1998). *Algorithms for scoring coreference chains*. LREC Workshop on Linguistic Coreference. — [métrica B-cubed]

- Blust, R. (2013). *The Austronesian Languages*. Edición revisada. Asia-Pacific Linguistics, ANU. — [clasificación de referencia de la familia]
- Blust, R. y Trussel, S. *Austronesian Comparative Dictionary*. En línea.
- Bowern, C. (2013). Relatedness as a factor in language contact. *Journal of Language Contact* 6. — [préstamo en léxico básico]
- Brown, C. H., Holman, E. W., Wichmann, S. y Velupillai, V. (2008). Automated classification of the world's languages. *Sprachtypologie und Universalienforschung* 61. — [método ASJP: LDN/LDND]
- Dahl, O. C. (1951). *Malgache et Maanjan*. Egede-Instituttet.
- Dolgopolsky, A. B. (1964). Gipoteza drevnejšego rodstva jazykovyx semej Severnoj Evrazii s verojatnostnoj točki zrenija. *Voprosy Jazykoznanija* 2. — [clases consonánticas]
- Gray, R. D., Drummond, A. J. y Greenhill, S. J. (2009). Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323. — [filogenia austronesia con longitudes de rama]
- Greenhill, S. J., Blust, R. y Gray, R. D. (2008). The Austronesian Basic Vocabulary Database: from bioinformatics to lexomics. *Evolutionary Bioinformatics* 4.
- Hammarström, H., Forkel, R., Haspelmath, M. y Bank, S. *Glottolog*. Max Planck Institute for Evolutionary Anthropology.
- Heggarty, P. *et al.* (2023). Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. *Science* 381. — [IE-CoR]
- Jäger, G. (2018). Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data* 5.
- Kessler, B. (2001). *The Significance of Word Lists*. CSLI Publications. — [parecido por azar en listas léxicas]
- 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. — [LexStat]
- List, J.-M., Greenhill, S. J. y Gray, R. D. (2017). The potential of automatic word comparison for historical linguistics. *PLOS ONE* 12(1). — [evaluación de Turchin, SCA y LexStat con B-cubed; austronesio entre los bancos de prueba]
- List, J.-M., Forkel, R., Greenhill, S. J. *et al.* (2022). Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data* 9.
- List, J.-M., Rzymski, C., Greenhill, S. J. *et al.* *Concepticon*. Max Planck Institute for Evolutionary Anthropology.
- Nichols, J. (1992). *Linguistic Diversity in Space and Time*. University of Chicago Press. — [confusión entre geografía y genealogía]
- Rama, T., List, J.-M., Wahle, J. y Jäger, G. (2018). Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? *NAACL-HLT*.
- Ranacher, P., Neureiter, N., van Gijn, R. *et al.* (2021). Contact-tracing in cultural evolution: a Bayesian mixture model to detect geographic areas of language contact. *Journal of the Royal Society Interface* 18.
- Ross, M. (1988). *Proto Oceanic and the Austronesian Languages of Western Melanesia*. Pacific Linguistics.

Ross, M., Pawley, A. y Osmond, M. *The Lexicon of Proto Oceanic*. Pacific Linguistics.

Sagart, L. (2004). The higher phylogeny of Austronesian and the position of Tai-Kadai. *Oceanic Linguistics*

1. — [subagrupamiento formosano argumentado desde los numerales]

Ringe, D. (1992). On calculating the factor of chance in language comparison. *Transactions of the American Philosophical Society* 82.

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.

Wichmann, S., Holman, E. W. y Brown, C. H. (eds.). *The ASJP Database*. — [base de datos; el método está en Brown et al. 2008 y Holman et al. 2008]

Nota: las referencias se dan con autor, año y publicación. Antes de cualquier envío conviene completarlas con páginas y DOI, y verificar las ediciones citadas.

Datos y reproducibilidad

Catálogo	output/catalogo-austronesio/catalogo-austronesian.tsv — 10.994 filas
Detalle por forma	output/catalogo-austronesio/catalogo-austronesian-lenguas.tsv — 84.884 filas
Construcción	analysis/fases/catalogo_familia.py
Agrupamiento y nullos (§4.1)	analysis/fases/agrupamiento_familia.py → docs/AUSTRONESIO-PRIMERA-APLICACION.md
Calibración (§3.5)	analysis/estudios/fase_c_calibracion.py → docs/FASEC-CALIBRACION.md
Despiece morfológico (§7.3)	analysis/fases/fase2_despiece.py → docs/FASE2-DESPIECE.md
Pares y clases (§6.3)	analysis/fases/fase6_clases_emergentes.py → docs/FASE6-CLASES-EMERGENTES.md
Peldaño por conjunto	analysis/estudios/peldano_por_conjunto.py — material de apoyo, no citado en el cuerpo
Códigos y parentesco (§6.2)	analysis/fases/codigos_vs_parentesco.py
Anexo A	analysis/fases/anexo_codigos.py

Tres avisos sobre esta tabla, porque una tabla de reproducibilidad que no es fiel es peor que ninguna. (i) Buena parte de las cifras del cuerpo no vive en la salida de un script sino en los informes de ejecución de docs/, que por eso se listan al lado. (ii) Los recuentos de `agrupamiento_familia.py` y los del catálogo no usan el mismo criterio: el primero agrupa desde dos lenguas y cuenta por `lect_id`, el catálogo exige tres y cuenta por nombre; de ahí que la cobertura del 39,1 % y la precisión del 15,92 % de §4.1 vengan de poblaciones distintas aunque se citen en la misma frase. (iii) Las cifras de ramas formosanas de §6.1 y las de macroáreas se recomputan directamente desde los dos TSV publicados, no desde `codigos_vs_parentesco.py`, que no las calcula.

Fuentes de datos: Lexibank (List, Forkel, Greenhill et al. 2022, CC-BY-4.0), IDS (Key & Comrie, CC-BY-4.0), Concepticon y Glottolog (List et al.; Hammarström et al., CC-BY-4.0), IE-CoR (Heggarty et al. 2023, CC-BY-4.0). Todas redistribuibles con atribución.

Anexo A · Cien códigos comentados

El catálogo tiene 10.994 códigos y los datos completos están en `output/catalogo-austronesio/`. Lo que sigue es una muestra legible: los cien códigos binarios y ternarios de mayor alcance, cada uno con dieciséis lenguas y sus palabras.

Tres cosas sobre la presentación. Las dieciséis lenguas están repartidas a lo largo de la lista alfabética, no tomadas de su cabeza: dieciséis lenguas empezadas por A ocultarían la dispersión geográfica, que es justamente lo que el ejemplo debe mostrar. Los nombres de lengua se recortan cuando son largos; las formas no se alteran nunca. Y la agrupación temática se declara a mano como orden de lectura: no sale de los datos y no es una clasificación semántica.

Los códigos emergentes se escriben 1, 2 y 3 por las tres clases de §6.3 —1 = b f h p v w φ β, 2 = k s x z, 3 = l r—. El comentario que precede a cada sección señala qué se puede ver en el bloque. Es observación, no análisis: se espera que el lector mire las palabras.

Índice inverso: códigos que reaparecen en varios conceptos

El orden temático que sigue agrupa por sentido, y con ello dispersa justo lo contrario: un mismo esqueleto recayendo sobre conceptos sin relación entre sí, que es el fenómeno del §2.3 y la razón de que los códigos cortos no sostengan afirmaciones. Esta tabla lo devuelve a la vista antes de empezar.

Léase con el enriquecimiento al lado: cuando un código aparece en cinco conceptos y en todos con cifras bajas, no es un código, es una consonante frecuente.

Código	Conceptos con los que aparece, y su enriquecimiento
m · t	eye (246, 53×) · be dead or die (204, 48×) · snake (95, 23×) · right (71, 16×) · die (43, 56×) · black (39, 8×)
t · n	cry (109, 27×) · earth (92, 22×) · year (59, 16×) · mother (56, 13×) · intestines (47, 13×) · burn (40, 10×)
m · n	bird (172, 36×) · drink (81, 17×) · laugh (58, 13×) · tongue (38, 8×) · male person (37, 8×)
r · n	hear (108, 39×) · day (68, 26×) · leaf (40, 14×) · rain (precipitation) (36, 14×)
n · m	mosquito (104, 44×) · six (77, 61×) · drink (69, 28×) · house (62, 26×)
l · m	five (252, 120×) · hand (93, 41×) · come (35, 16×)
t · l	three (241, 86×) · rope (70, 26×) · egg (47, 18×)
n · n	mother (76, 19×) · eat (36, 9×) · yellow (36, 10×)
m · m	father (58, 24×) · chew (51, 35×) · piss (48, 159×)

Numerales

Los numerales son el mejor caso del método y conviene empezar por ellos.

El mismo étimo aparece repartido en varios códigos, y las diferencias entre ellos son correspondencias regulares. «Cuatro» sale como p · t (72 lenguas), h · p · t (36) y v · t (35): la misma palabra con la glotal inicial conservada —que el esqueleto escribe h, §2.1— y con p > v. Hay un cuarto código de «cuatro» que no entra en este anexo porque solo tiene 15 lenguas, y sin embargo es el más interesante de los cuatro: s · p · t, que cruza las cinco ramas formosanas del corpus y ninguna malayo-polinesia (§6.1). Los tres de arriba son el nivel PMP; el que falta es el nivel PAN. Que un criterio de alcance deje fuera al mejor candidato es, en sí mismo, la lección de este bloque. «Cinco» como l · m (252) y r · m (37); «tres» como t · l (241) y t · r (43) — la alternancia l/r de §6.3, aquí visible palabra a palabra. «Seis» como n · m (77) y h · n · m (40).

El catálogo no los une, y eso es deliberado: unirlos sería proponer una correspondencia, que es trabajo del análisis y no del inventario. Lo que hace es dejarlos contiguos, con sus cifras, para que la correspondencia se vea.

Obsérvese también el enriquecimiento: **s·j·m|NINE** alcanza 531× con solo 49 lenguas, mientras **p·t|four** se queda en 34× repartido entre 72 de ellas. Más lenguas no es más señal.

l·m|five · binario · 252 lenguas · enriquecimiento 120× · emergente 3·m

Agta lima	As lim	Bikol li'ma	Dupaningan Agta lima
Ibanag lima	Kairiru lim	Koronadal Blaán limi	Lonwolwol lim
Manobo 'lima	Merei lima	Nasarian i-ləme	Piamatsina lima
S: Raga: Avavanva lima	Simbo lima	Tagbanwa lima	Uripiv: Durpif e-lim

t·l|three · binario · 241 lenguas · enriquecimiento 86× · emergente t·3

'Oyaoya tolu	Antambohoaka telo	Banggi taolu	Bolaang Mongondow tolu
Futuna tolu	Itawis tallu	Kavalan F69 tulu	Lelepa tolu
Mamanwa tolu	Merina_Malagasy telo	Navut tolu	Nume i-tol
Saaroa F69 to:lo	South Efate itol	Tiale tolu	Vaeakau-Taumako tolu

p·t|SEVEN · binario · 101 lenguas · enriquecimiento 92× · emergente 1·t

Agta pitu	Arta pitu	Bantik pitu	Binukid 'pitu
Cebuano pito	Hiligaynon pito	Ilokano pito	Itawis pitu
Kalinga pitú	Kapampangan pitú	Manobo pitu	Masbatenyo pito
Rano: Chinambong ε-pət	Sangil 'pitu	Surigaonon pito	Tonsea pitu

n·m|SIX · binario · 77 lenguas · enriquecimiento 61×

Agutyanen enem	Ayta Abellen anım	Basay TsYM91 anəm	Bolaang Mongondow onom
Cebuano onom	Hiligaynon anem	Ibanag annam	Iraya anúm
Kalinga onom	Kinaray-a anəm	Manide anum	Moken nam
Romblon unóm	Sarangani Blaán nam	Tagalog anem	Ughele ɔnoɔmo

w·l|EIGHT · binario · 75 lenguas · enriquecimiento 151× · emergente 1·3

Agta walu	Arta walu	Bantik walu	Bolaang Mongondow wolu
Gaddang walu	Hiligaynon walo	Inagta Alabat wa'lu	Isneg walu
Kalinga walú	Koronadal Blaán walu	Manide walo	Maranao walo
Ratagnon waló	Sambal wa'lo	Subanon walu	Tontemboan walu

p·t|four · binario · 72 lenguas · enriquecimiento 34× · emergente 1·t

Agutyanen epat	Ayta Abellen apat	Bare'e pata	Bontok opat
Cebuano upát	Guiangan appat	Inati apat	Itawis appat
Kalinga opat	Lelepa pati	Misima epat	Ratagnon apat
Sinauna a'pat	Subanon pat	Thao B96 pat	Tungag / Tungak / apuat

s·j·m|NINE · ternario · 49 lenguas · enriquecimiento 531× · emergente 2·j·m

Aklanon – Bisayan syam	Ayta Mag–indi siyam	Bikol siyam	Bolinao syam
Cuyonon siyám	Hanunóo sijam	Inibaloi siyam	Itawis siyam
Kagayanen syam	Kalinga siyam	Kinaray-a siyam	Manobo siyam
Ratagnon siyám	Sambal siyam	Subanon siyam	Tagalog siyam

t·r|three · binario · 43 lenguas · enriquecimiento 20× · emergente t·3

Angavae i-εtJ	Babuza Ts82 tarua	Emae toru	Ifira-Mele e-toru
Kalinga tuļu	Kavalan F69 turu	Mangareva toru	Mansaka turu
Nukumanu toru	Puyuma tero	Rarotongan toru	Seediq F69 teru
Tarpia tor	Tikopia toru	Uliveo i-tor	Yami atoro

h·n·m|SIX · ternario · 40 lenguas · enriquecimiento 308× · emergente 1·n·m

Agta ʔʌnnʌm	Balangaw ʔi'nɪm	Gaddang ʔanɪm	Hanunóo ʔunum
Ifugao ʔenem	Inibaloi ʔinɪm	Itneg ʔi'nɪm	Kalagan ʔinɪm
Kamayo ʔunúm	Kapampangan ánam	Manobo ʔinɪm	Manobo ʔʌnnʌm
Palawan Batak ʔinɪm	Sakizaya ʔənəm	Tagalog ánim	Tagbanwa ʔinɪm

r·m|five · binario · 37 lenguas · enriquecimiento 39× · emergente 3·m

Arosi rima	Bauro erima	Emae rima	Futuna-Aniwa rima
Kanakanabu F69 rima	Kavalan F69 rima	Maori rima	Nukumanu rima
Paiwan rima	Port Sandwich ε-rim	Rarotongan rima	Saaroa F69 rima
Tahitian rima	Tikopia rima	Tuamotuan rima	Uruava ruma

h·p·t|four · ternario · 36 lenguas · enriquecimiento 226× · emergente 1·1·t

Agta ʔʌppat	Bikol ʔə'pat	Dumagat ʔi'pat	Hanunóo úpat
Ifugao ʔepat	Inibaloi ʔipat	Itbayaten ʔapat	Kagayanen ʔáppat
Kalinga ʔo'pat	Kamayo ʔupát	Mamanwa ʔopat	Manobo ʔipat
Palawan Batak ʔépat	Sambal aʔpat	Tagalog ápat	Tagbanwa ʔipat

v·t|four · binario · 35 lenguas · enriquecimiento 44× · emergente 1·t

Aore a_vati	Bilur vauat	Burmbar e-vet	Kandas vat
Lunga Lunga aivat	Magori vati	Mbirao vati	Molima veyata
Mpotovoro vāt	Nese -vat	Piva evati	Sisingga vati
Siviti i-vat	Sowa ivêt	Tasmate vati	Tolomako vati

Partes del cuerpo

Aquí aparece un patrón que se repetirá en todo el anexo: el mismo étimo con y sin sufijo posesivo produce dos códigos distintos.

«Ojo» da m·t (246 lenguas) y m·t·n (65). «Pecho» da s·s (184) y s·s·n (50). Ese -n final es el posesivo de tercera persona, y en muchas lenguas viene marcado con guion en el propio dato: *mata-n*, *susu-na*. El código lo trata como una consonante más, porque no sabe que es morfología.

Esto no es un error del catálogo: es información morfológica emergiendo de él. Que un concepto aparezca sistemáticamente con y sin una consonante final en la misma posición es exactamente la pista que llevaría a identificar un sufijo, y es el camino de la segmentación pendiente (§7.3).

«Lengua» es el caso más disperso: d·l (52), l·l (42), m·n (38) y d·l·h (37). d·l y d·l·h son el mismo étimo con y sin la glotal final —*dila* frente a *dilaʔ*—, y l·l es compatible con ellos. m·n es otra palabra (*mea*, *mea-na*, *mena*-), y su enriquecimiento de 8× lo delata: es de los más bajos del anexo y el paper enseña a leerlo como ruido. De paso, *mea* / *mea-na* vuelve a mostrar el sufijo posesivo con que abre esta sección.

m·t|eye · binario · 246 lenguas · enriquecimiento 53×

Agta mʌtʌ	Arta mata	Bantoanon mata	Bughotu mata
Emae mata	Hiw məta-	Isneg mata	Kavalan F69 mata
Malalamai mata	Mansaka mata	Mota matai	Palawan Batak mata
Ririo mat	Sinaugoro mata-	Tangoa məta-	Vaghua mata-

s·s|BREAST · binario · 184 lenguas · enriquecimiento 95× · emergente 2·2

'Are'are susu	Anuki susu	Bantik susu	Buli sus
Hanunóo susu	Itbayaten soso	Kankanaey soso	Lau susu-
Malo susu	Marino sus	Na:ti səsə-	Pangasinan suso
Roria susu-	Sewa susu	Tagalog su:su	Torau susu

l·m|hand · binario · 93 lenguas · enriquecimiento 41× · emergente 3·m

Akolet lama	Babuza Ts82 lima	Bontok lima	Fortsenal lima-
Ibanag lima	Kandas a_lama-	Lakalai lima	Lunga Lunga a_lima
Marino lima-i	Navut lima-	Pangasinan lima	Puyuma llima
S: Raga: Avavanva lima	Sinauna alíma	Tiale lima	Western Fijian lima

m·t·n|eye · ternario · 65 lenguas · enriquecimiento 77×

Alor m'atʌŋ	Bilur omatana	Burmbar mata-n	Gedaged matan
Iduna matana	Kusaghe mata-na	Likum mita-n	Malango mata-na
Misima mata na	Nahavaq mətɛ-n	Nggela mata-na	Roviana mata-na
Siviti mətə-n	Tesmbol mətɛ-n	Ubir mata n	Uripiv: Durpif mətɛ-n

t·l·n|ear · ternario · 54 lenguas · enriquecimiento 78× · emergente t·3·n

Alor t'iluŋ	Bare'e taliŋa	Biliau taləŋa-u	Gitua taliŋa
Kagayanen taliŋa	Kove taliŋa	Malalamai taliŋa	Mangap-Mbula talŋa-
Mutu taliŋa	Nias taliŋa	Pukapuka talinga	Simbo taliŋa
Tigak taləŋa	Uruava talina	Vangunu taliŋi	Western Bukidnon teləŋa

d·l|tongue · binario · 52 lenguas · enriquecimiento 92× · emergente d·3

Alangan dila	Ayta Ambala dila	Balangaw 'dila	Batangan dila
Capisano dila	Gaddang dila	Hanunóo dila	Ifugao di:la
Isinay dila	Itneg 'dila	Kalinga dela	Kallahan dila
Kinaray-a dila	Pangasinan dila	Sangil 'dila	Tadyawan dilâ

s·s·n|BREAST · ternario · 50 lenguas · enriquecimiento 133× · emergente 2·2·n

'Are'are susu-na	Bauro susu-na	Dori'o susu-na	Koro süsu-ŋ
Lau susu-na	Lunga Lunga a_susuna	Marino susuŋa	Mbirao čuču-na
Nduke susu-na	Pangkumu susi-n	Roviana susu-na	Shark Bay sisia-ŋ
Talise čuču-na	Tawaroga susu-na	Ubir susu n	Uruava susú-na

t·n|INTESTINES · binario · 47 lenguas · enriquecimiento 13×

Ambrym teña	Aore tine-	Buol tinae	Cebuano tiná-i
Gedaged taen	Kapingamarangi tinae	Lihir atien	Malo tine
Masbatenyo tinâi	N: Havai: Lombaha tinae	Orkon tine-	Sangil ti'nai
Sowa tène-	Surigaonon tina-i	Tangoa tine-	Uvea tinae

l·l|tongue · binario · 42 lenguas · enriquecimiento 18× · emergente 3·3

Antaimoro lela	Antambohoaka lela	Antankarana lela	Bara lela
Buhutu lelo	Itbayaten lila	Luangiua alelo	Masikoro lela
Mikea lela	Niue alelo	Pukapuka alelo	Sakalava lela
Sihanaka lela	Tikopia alelo	Tsimihety lela	Wuna lela

m·n|tongue · binario · 38 lenguas · enriquecimiento 8×

'Are'are mea-na	Ambrym mə-ŋ	Burmbar mə-ŋ	Dori'o mea-na
Ghanongga mea-na	Iamalele meana	Kusaghe mea-na	Kwara'ae mea-n
Luqa mea-na	Marovo mea-na	Namagir mena-	Port Sandwich mə-n
Sa'a mea-na	Tawala mena-	To'ambaita mea-na	Uliveo mea-n

d·l·h|tongue · ternario · 37 lenguas · enriquecimiento 196× · emergente d·3·1

Aklanon – Bisayan dí:la?	Binukid 'dila?	Bolinao dila?	Dumagat di'la?
Iban dilah	Inati dila?	Kalagan dila?	Mamanwa dila?
Manobo 'dila?	Manobo dila?	Mansaka dila?	Romblon díla'
Sambal 'dila?	Subanon dilah	Surigaonon dila?	Tiruray dila?

Parentesco y persona

El parentesco separa dos cosas que conviene no confundir.

«Padre» aparece con cuatro códigos, y sus enriquecimientos los ordenan solos: t·m·n (153×), t·m (76×), m·m (24×) y t·t (22×). Los dos primeros son un étimo con y sin sufijo; los dos últimos son palabras de crianza —*mama*, *tata*— del tipo que aparece en familias sin ninguna relación entre sí y que el método comparativo trata aparte por esa razón.

El enriquecimiento los distingue sin saber nada de eso: una palabra de crianza está hecha de las consonantes más frecuentes y repetidas, así que su enriquecimiento se queda muy por debajo del de un étimo heredado: 24× y 22× frente a los 76× y 153× de las dos primeras. Es un caso donde la cifra hace el trabajo que haría el criterio del comparatista.

t·m|FATHER · binario · 105 lenguas · enriquecimiento 76×

'Auhelawa tama	Anejom etma-	Bariai tama	Buhutu tama
Gitua tama-	Kiribati táma	Luqa tama	Manam tama
Morouas tama-	Navut tama-	Papitalai tama-	Saliba tama
South Efate tem	Tangoa taäma-	Tolai tama-	Varisi tama-

k·m|YOU · binario · 103 lenguas · enriquecimiento 70× · emergente 2·m

Akei ikomiu	Atchin: Orap kami	Bare'e komi	Butuanon kamo
Hanunóo kamu	Koro kimi	Löyöp kimi	Masbatenyo kamo
Mwerlap kami	Ninde kamo	Ratagnon kamó	Satawalese kámi
Tadyawan kam	Tasmate kami	Tunjung kam	Vurës ekimi

n·n|MOTHER · binario · 76 lenguas · enriquecimiento 19×

Ali nyany	Antanalana nene	As nen	Biliau nin
Futuna-Aniwa nana	Hula ina-na	Koluwawa inana	Mahafaly nene
Mikea nene	Nauru inen	Orkon nina	Sakalava neny
Talise na-na	Tawaroga ena-na	Tsimihety niny	Western Fijian Nene

t·m·n|FATHER · ternario · 70 lenguas · enriquecimiento 153×

Ambrym temæ-ŋ	Bellona tamana	Bunama tamana	Gedaged tamán
Haku tamana	Koro tama-ŋ	Lonwolwol timæ-n	Marovo tama-na
Muyuw taman	Nauna tama-n	Nombotkote atimæn	Roviana tama-na
Sikaiana tamana	Tape etmen	Ughele tama-na	Vaghua tama-na

m·m|FATHER · binario · 58 lenguas · enriquecimiento 24×

'Are'are mama	Amis mama	Avava mom	Bughotu mama
Fifti: Khatbol mama	Gitua mama	Inati mama	Kokota mama
Leleghia mama	Löyöp mam	Mindiri mamə	Mota mama
Neverver mama	Papitalai momo	Rukai mamaa	Tomoip mimma

t·n|MOTHER · binario · 56 lenguas · enriquecimiento 13×

Akei tina-	Baluan tina	Buhutu tina	Dehu thin
Kiribati tina	Luqa tina	Maleu tna	Marovo tina-
Motu tina-	Nukumanu tinna	Piamatsina tina-	Samoan tinā
Sinaugoro tina-	Sudest tinae	Tangoa tina-	Tutuba tina-

n·t|CHILD · binario · 55 lenguas · enriquecimiento 31×

Akei natu-	Aore natu-	Fortsenal natu-	Iamalele natu
Kandas nat	Mafea natu-	Manam natu	Minaveha natu
Morouas natu-	Namakir nati-	Nese nat-	Patpatar natu
Sivisa Titan nat	Sye nete-	Tiale natu-	Tungag / Tungak / nat

t·t|FATHER · binario · 51 lenguas · enriquecimiento 22×

Alavas: Alavas tete	Avok tata	Burmbar tata	Ctl: Nalemba: Nar tata
Kapampangan tata	Maii tata	Mpotovoro tete	Nahavaq tatai
Ninde tatai	Nombotkote tata	Patpatar tata	Rano: Chinambong tete
Siviti tate	Tasmate tata	Tirax tata	V'ënen Taut tate

k·m|WE · binario · 47 lenguas · enriquecimiento 103× · emergente 2·m

Alangan kami	Ayta Ambala kami	Bare'e kami	Bikol kami
Bunun F69 kam	Cebuano kami	Gaddang ikkami	Inati ikám
Kalinga kami	Kapampangan ikami	Maloh ikam	Mansaka kami
Palawan Batak kámi	S.W. Palawano kami	Surigaonon kamo	Tausug kami

n·k|CHILD · binario · 41 lenguas · enriquecimiento 34× · emergente n·2

Ambel nikwë	Ayta Ambala anak	Babuyan anak	Bantoanon anak
Bolinao anak	Cebuano anak	Hiligaynon anak	Iban anak
Itbayaten anak	Kankanaey anak	Lala naku	Masbatenyo anak
Pangasinan anak	Sambal anak	Surigaonon anak	Waray-Waray anák

n·r|THEY · binario · 39 lenguas · enriquecimiento 27× · emergente n·3

Agutyanen nira	Ambae nɪər	Antaisaka anareo	Araki nira
Banoni nari	Halia nori	Lelepa naara	Morouas nira
N: Havai: Lombaha gere	Nakanamanga: Sesa nara	Piamatsina injire	Rano: Chinambong nir
Tiang ner	Uripiv: Durpif nir	Vunapu inire	Wailapa inira

s·n|WHO · binario · 38 lenguas · enriquecimiento 16× · emergente 2·n

Amara seiŋa	Babuyan sino	Bikol sin-o	Hiligaynon sin-o
Ilokano sinno	Kalinga 'sinu	Kankanay si'no	Kwamera sin
Mansaka sini	Ratagnon sino	Sa sien	Siviti sena
Sori sen	Surigaanon sin-o	Tambotalo sin	Unua asigo

m·n|MALE PERSON · binario · 37 lenguas · enriquecimiento 8×

'Are'are mane	Ambel man	As -man	Bughotu mane
Gimán maón	Kokota mane	Longgu mwane	Malango mane
Minyaifuin man	Mortlockese mwæ:n	Numfor man	Oroha mane
Puluwatese mʷáán	Satawalese mʷáán	Tawaroga mwane	Waropen mano

n·k|THOU · binario · 36 lenguas · enriquecimiento 35× · emergente n·2

Akei iniko	Araki niko	Baetora niko	Koro nik
Lemerig næk	Marino iniko	Mwerlap iniok	N: Leonda nieko
Nokuku nik	Orkon nak	Rano: Chinambong nik	Tangoa eʷko
Tigak nuk	Tonsea ni ko	Vera'a enike	Wailapa iniko

l·l·k|MALE PERSON · ternario · 35 lenguas · enriquecimiento 169× · emergente 3·3·2

Agta lálaki	Ayta Mag-antsi lalaki	Bikol lalake	Bolaang Mongondow lolaki
Capisano lalake	Cuyonon lalaki	Hanunóo lalaki	Ibanag lalaki
Inati lalaki	Itawis lalaki	Kalinga lálaki	Kapampangan lalaki
Masbatenyo lalaki	S.W. Palawano lelaki	Sambal la'laki	Tagbanwa lalaki

Naturaleza y cielo

«Piedra» con tres códigos —b·t (98), v·t (49), f·t (38)— es la serie labial de §6.3 desplegada en un solo concepto: *b, v, f* en la misma posición, en tres grupos de lenguas.

«Mar» ilustra el otro extremo: t·s (57 lenguas, 68×) y d·g·t (49, 248×) no son variantes del mismo étimo, son dos palabras distintas. El enriquecimiento del segundo lo delata: una secuencia rara repetida en cuarenta y nueve lenguas.

k·l·t|LIGHTNING (137×) y b·l·n|moon (149×) son ternarios de libro: tres posiciones que acertar, alcance amplio y enriquecimiento alto.

b·t|stone · binario · 98 lenguas · enriquecimiento 113× · emergente 1·t

Agta batu	Ayta Ambala bato	Balangaw ba'to	Bikol ba'to
Bujan bat	Dupaningan Agta bitu	Hoanya Ts82 ba:tú:	Ilokano bato
Itbayaten bato	Kalinga ba'tu	Kapampangan batu	Mamanwa bato
Manobo bato	Pazih F69 batu	Sambal ba'to	Tadyawan bató

t·n|earth · binario · 92 lenguas · enriquecimiento 22×

Ambae tan	Antaimoro tany	Antanosy tany	Bantik tana
Betsimisaraka tane	Ctl: Nalemba: Nar tano	Gitua tano	Lonwolwol tan
Mangap-Mbula toono	Merina_Malagasy tàny	N: Havai: Lombaha tano	Nulnesa tono
Sa tan	Sio tono	Tsimihety tany	Wab ton

b·l·n|moon · ternario · 71 lenguas · enriquecimiento 149× · emergente 1·3·n

Agutyanen bolan	Bahasa Indonesia bulan	Bikol bo'lan	Bolinao bulan
Cuyonon bulán	Hoanya Ts82 bu-lan	Ifugao bu:lan	Iraya bulán
Kalinga bu'lan	Kavalan F69 bu:lán	Manide bilan	Manobo bolan
Pangasinan bulan	Sambal bolan	Tadyawan bulán	Tboli bulon

l·n|wind · binario · 69 lenguas · enriquecimiento 30× · emergente 3·n

Akei lani	Aore laḡ	Butmas-Tur laḡ	Eton laḡ
Lamenu laḡ	Lonwolwol leḡ	Malo laḡi	Mwerlap laḡ
Narango lan	Nulnesa laḡ	Port Vato leḡ	S: Raga: Avavanva laḡi
Siviti laḡ	Tasmate lan	Uripiv: Durpif liḡ	Vurès leḡ

r·n|day · binario · 68 lenguas · enriquecimiento 26× · emergente 3·n

Akei rani	Apma ren	Ctl-E: Tawet: Nav ran	Ghanongga rane
Larevat iren	Malo rani	Mokilese reen	N: Leonda rani
Nduke rane	Piamatsina ran	Roro rani	Sakao ɛren
Sivisa Titan raḡ	Tambotalo ranja	Tolomako ran	Vaghua ranḡ

t·n|year · binario · 59 lenguas · enriquecimiento 16×

Akei tauni	Antandroy taogne	Berawan_Long_Tera taun	Bikol táon
Buol taun	Inibaloi tao-en	Malo tauni	Merei tauon
Morouas taone	Numfor taun	Roria taon	Sihanaka taona
Tambotalo taunua	Tiale taoni	Tsimihety tagno	Wailapa taun

t·s|sea · binario · 57 lenguas · enriquecimiento 68× · emergente t·2

Akei tasi	Angavae ɛ-tʰaːs	Bare'e tasi	Ctl: Nalemba: Nar tas
Ghari tasi	Kandas tasi	Loniū tas	Mafea tasi
Mbirao tasi	Narango tas	Nokuku tps	S: Raga: Avavanva tasi
Sissano tias	Tambotalo tasia	Tape tes	Tiang tas

k·l·t|LIGHTNING · ternario · 51 lenguas · enriquecimiento 137× · emergente 2·3·t

Agta kilat	Ayta Abellen kilat	Bikol kil-at	Buol kilat
Dumagat ki'lat	Guiangan kilat	Ifugao ki'lat	Iranun kilat
Kalagan kilat	Kinaray-a kilat	Maloh kilat	Manobo kilat
Maranao kilat	Ratagnon kilát	Subanon kilat	Tausug kilat

d·g·t|sea · ternario · 49 lenguas · enriquecimiento 248×

Aklanon – Bisayan dá:gat	Ayta Mag-antsi dagat	Bikol dagat
Bolaang Mongondow dagat	Capisano dagat	Dumagat di'get
Hiligaynon dagat	Kagayanen dá:gat	Kamayo dá:gat
Mamanwa dagat	Manobo 'dagat	Manobo dagat
S.W. Palawano dagat	Subanon dagat	Tadyawan dagát
Tausug dagat		

v·t|stone · binario · 49 lenguas · enriquecimiento 61× · emergente 1·t

Angavae ε-vet	Antambohoaka vato	Antankarana vato	Avava a-vat
Betsimisaraka vato	Iaai veto	Kavalan F69 vato	Mahafaly vato
Meramera vatu	Mikea vato	Namagir vat	Raga vatu
Sihanaka vato	Talise vatu	Tesmbol i-vet	Vezo vato

b·n|night · binario · 41 lenguas · enriquecimiento 41× · emergente 1·n

Andra biŋ	Babuza Ts82 bi-ini	Bariai boŋ	Bughotu boŋi
Lau boŋi	Malalamai boŋ	Marshallese bʷéŋʷ	Mokaren biŋ
Neve'ei buŋ	Pak bu:ŋ	Ponam biŋ	Saipan Carolinian bʷooŋ
Simbo boŋi	Tawaroga bwŋi	Tiang biŋ	Tulu bI:n

m·s|star · binario · 39 lenguas · enriquecimiento 16× · emergente m·2

Akei matˢoi	Avok mʷaŋi	Buninga: Moso masoe	Fortsenal matˢoi
Koro mʷese	Mafea masoa	Mbwenelang mʷose	Nisvai mʷaŋi
Orkon m̥iso	Piamatsina matˢoe	Ske mso	Sye mosi
Tesmbol mʷose	Unua matʃo	V'enen Taut məsi	Vunapu matˢoi

f·t|stone · binario · 38 lenguas · enriquecimiento 50× · emergente 1·t

Ambae fat	Amblong fatu	Bontok fiato	Ctl: Nalemba: Nar fat
Futuna fatu	Ifira-Mele fatu	Leo Tuvalu fatu	Marino fato
Narango fat	Polonombauk fat	Samoan fatu	South Efate fat
Takuu fatu	Thao B96 fatu	Tsou T63 fatu	Uvea fatu

n·b·n|night · ternario · 37 lenguas · enriquecimiento 72× · emergente n·1·n

Ambae ʱboiŋ	Bieria iʱboŋ	Ctl: Nalemba: Nar ʱbʷoŋ	Ghanongga mbŋi
Kusaghe mboŋi	Longgu mboŋi	Maii iʱboŋ	N: Havai: Lombaha ʱboŋi
Narango ʱbon	Nese ʱbun	Pangkumu aʱboŋ	S: Raga: Avavanva ʱboŋi
Talise mbŋi	W: Walaha ʱboŋi	camuki ʱboen	ʰMoaveke ʱboen

r·n|RAIN (PRECIPITATION) · binario · 36 lenguas · enriquecimiento 14× · emergente 3·n

Antambohoaka orana	Arosi raŋi	Ayta Mag-antsi uran	Basay TsYM91 uran
Bikol uran	Cuyonon urán	Ibanag uran	Iranun uran
Kapampangan uran	Kinaray-a uran	Masbatenyo uran	Merina_Malagasy orana
Sakalava orana	Sihanaka orana	Tanala orony	Tsimihety oragna

Animales

m·n|bird reúne 172 lenguas con 36× y es el binario de mayor alcance de la sección. Conviene compararlo con **m·t|snake** (95 lenguas, 23×), que comparte esqueleto con **m·t|eye** (246, 53×) y **m·t|BE DEAD OR DIE** (204, 47×): son conjuntos separados —cada uno con su concepto— pero el enriquecimiento de «serpiente» es el más bajo de los tres, y la cifra avisa de que **m·t** es demasiado común para sostener nada por sí solo.

Y hay un segundo «serpiente» que enseña más. **n·m·t|snake** (35 lenguas, 72×) es el mismo étimo con un prefijo nominal que el propio dato trae separado con guion: *na-mat*, *nə-mat*, *na-mate*, *n-mat*.

El catálogo lo registra como código distinto porque no sabe que na- es morfología — el mismo fenómeno que §5.4 describe para «negro», pero aquí en un sustantivo y con el límite visible en la transcripción. Quien quiera ver de un vistazo por qué falta la segmentación, este bloque lo muestra mejor que cualquier explicación.

m·n|bird · binario · 172 lenguas · enriquecimiento 36×

'Are'are manu	Arop manu	Bauro amanu	Ctl: Nalemba: Nar manu
Hiw mɔn	Kove manu	Lihir mon	Mangseng mɛn
Moene ^m benj meni	Nakanamanga: Sesa manu	Nulnesa ma:nu	Rapa Nui manu
South Efate manu	Tarpia manu	Tongan manu	Vurës m ^w on

k·t|louse · binario · 135 lenguas · enriquecimiento 65× · emergente 2·t

Agta kutu	Apma kut	Basay TsYM91 kutu	Buol kutu
Futuna kutu	Ibanag kutu	Itbayaten koto	Kankanaey koto
Lelepa kuutu	Manide kuto	Maranao koto	Namagir kit
Penrhyn kutu	Sikaiana kutu	Tagbanwa kutu	Uvea kutu

n·m|MOSQUITO · binario · 104 lenguas · enriquecimiento 44×

Agutyanen namo	Axamb nam	Bughotu ñamu	Ctl-E: Tawet: Nav nam
Futuna-Aniwa namo	Kilivila nim	Lemerig næm	Maewo na:mu
Misima nimu	Nalik namu	Nukuoro namu	Puluwatese -nóm ^w
Samoan namu	Sobei namu	Tuamotuan namu	Vera'a nam

m·t|snake · binario · 95 lenguas · enriquecimiento 23×

'Oyaoya mata	Anuki mota	Bebeli mɔtɔ	Bulu mata
Eton m ^w at	Kove mota	Loniu m ^w at	Malasanga mota
Minaveha mwata	Mota m ^w ata	Nggela umata	Saliba mwata
Sivisa Titan moat	Talise muata	Tigak muata	Vunapu mata

t·l|egg · binario · 47 lenguas · enriquecimiento 18× · emergente t·3

Ambae tol	Ambel talo	Antanosy atoly	Biga tolo
Ctl-E: Tawet: Nav tol	Ifira-Mele atolu	Lengo tolu	Mahafaly atoly
Marino toli	Mota toliu	N: Leonda tol	Nggela tolu
S: Raga: Avavanva toli	Sakao atæl-	Tunjung talo	Vao tol

k·l|dog · binario · 35 lenguas · enriquecimiento 16× · emergente 2·3

Ambrym kuli	Apma kuli	Baki kuli	Bieria kuliu
Daa kaka kuli	Futuna kulii	Lamenu kuli	Leo Tuvalu kulii
Nahavaq kilu	Nembao kuli	Niue kulii	Pukapuka kulī
Sikaiana kulii	Tanna kuli	Uvea kulii	Vaeakau-Taumako kuli

Plantas y alimento

«Beber» da m·n (81) y n·m (69). Nótese que son las mismas dos consonantes en orden inverso. El catálogo los trata como códigos distintos, y hace bien: el orden es parte del código. Pero la coincidencia salta a la vista y merece examen — puede ser metátesis, o pueden ser dos palabras.

b·n·g|fruit (133×) muestra el contraste habitual con los verbos vecinos: los conceptos concretos tienden a enriquecimientos altos, los muy generales (n·n|eat, 9×) a enriquecimientos que rozan el azar.

m·n|drink · binario · 81 lenguas · enriquecimiento 17×

Ambrym -mün	Antanosy mino	Babuza Ts82 man o	Bierebo -munia
Ctl-E: Tawet: Nav min	Gimán imin	Lelepa munu	Maii -mun
Na'ahai -mən	Nasarian -min	Nisvai -mun	Pangkumu -min
Sa imini	Sori mwiŋ	Tape -min	Uripiv: Durpif -mini

n·m|drink · binario · 69 lenguas · enriquecimiento 28×

'Auhelawa numa	As -nim	Ayta Mag-indi inum	Buhutu nom
Hawaiian inu-mia	Itawis inúm	Kinaray-a inəm	Maori inu-mia
Mindiri ŋa-mi	Mpotovoro -num	Numbami -numu	Ramoaaia inim
Sambal inom	Surigaonon inom	Tigak inum	Wampar nom

k·n|eat · binario · 69 lenguas · enriquecimiento 25× · emergente 2·n

Akolet ken	Arop a-kani	Axamb -kan	Ayta Mag-indi ikan
Bolaang Mongondow kaan	Fifti: Khatbol -kan	Itbayaten kan	Kiribati kana
Maguindanaon kan	Maleu kan	Maranao kan	Na'ahai -kan
Nitita -kan	Rano: Chinambong ku-ain	Sambal kan	Surigaonon kaon

s·n|salt · binario · 51 lenguas · enriquecimiento 30× · emergente 2·n

Alangan asín	Ayta Mag-indi asin	Batangan asín	Bikol asín
Butuanon asin	Chuukese san	Ibanag asin	Inibaloi asin
Itawis asin	Kalinga asín	Kapampangan asin	Masbatenyo asin
Romblon asín	Sangir asinŋ	Tadyawan asín	Tontemboan asinŋ

m·m|CHEW · binario · 51 lenguas · enriquecimiento 35×

Ansus emama	Bali-Vitu mama	Bikol mama	Gedaged mam
Isneg mama	Kalinga mu'ma	Kwaio meme-a	Lunga Lunga mamai
Namakir mama	Numfor mam	Roro momo	Sikaiana maamaa
Tetum mam:a	Torau mama	Vitu mama	Windesi Wandamen mam

b·n·g|fruit · ternario · 41 lenguas · enriquecimiento 133× · emergente 1·n·g

Atta bunga:	Bantoanon bunga	Binukid 'bunga	Capisano bunga
Guiangan bungo	Ifugao bu:nga	Inati búnga	Isinay bunga
Kakiduge:n Ilongo bunga	Kalinga búnga	Koronadal Blaan bingɿ	Masbatenyo bunga
Pangasinan bunga	Sambal 'bonga	Sinauna bunga	Tagalog bunga

r·n|leaf · binario · 40 lenguas · enriquecimiento 14× · emergente 3·n

Akei rau-na	Arop rau-na	Axamb raun	Doura rauna
Lou rei-n	Malasanga rau-na	Mor ranu	Nasarian a-raon
Pangkumu raon	Rano: Chinambong roin	Sissano ro-ñ	Tape ron
To'ambaita rea-na	Unua a-raon	Vunapu rauna	Wusi rauna

n·n|eat · binario · 36 lenguas · enriquecimiento 9×

Amblong anian	Atchin: Orap an-en	Baluan ŋan	Fortsenal aniani
Kandas i_uəŋon	Loniu ñani	Mafea anan	Merei anian
Morouas aniani	Namakir ŋan	Navut anani	Nisvai -nin
Sihanaka inana	Torau aniani	Ubir i ani n	Vangunu ŋəŋau

Verbos de acción y estado

La sección donde mejor se ve la variación regular.

«Morir» da m·t (204) y m·s (47): la palatalización *t > s de §6.3. «Cavar» da k·l (77) y x·l (49): la espirantización *k > x. «Oír» da r·n (108) y l·n (47): las líquidas otra vez. Tres correspondencias distintas, cada una visible en un concepto y confirmada en el recuento global.

m·t·r|sleep (90×) y m·t·k|fear (94×) son ternarios que empiezan por m·t, como «ojo» y «morir». La tercera posición es lo que los separa, y es un buen recordatorio de por qué los binarios discriminan menos (§2.3).

m·t|BE DEAD OR DIE · binario · 204 lenguas · enriquecimiento 48×

'Oyaoya mate	Aria mete	Basay TsYM91 matai	Chamorro matai
Haku mete	Kankanaey matey	Lelepa mate	Malango mate
Mengen mate	N: Havai: Lombaha mate	Nitita -mat	Pukapuka mate
Sarangani Blaen mati	Sursurunga mat	Tikopia mate	Vaeakau-Taumako mate

t·n|cry · binario · 109 lenguas · enriquecimiento 27×

Akei tani	Anejom tañ	Bariai taŋ	Fortsenal tani
Kove taŋi	Levei e teŋ	Malalamai taŋ	Marquesan tani
N: Leonda taŋ	Nggela taŋi	Penrhyn tangi	Roinji u-taŋ
Sobei e-tan	Tangoa taŋi	Uruava tani	Vurës teŋ

r·n|hear · binario · 108 lenguas · enriquecimiento 39× · emergente 3·n

'Are'are rono	Anuta roŋo	Butmas-Tur roŋ	Faghani roŋo
Kwara'ae roŋpa	Lemerig rəŋ	Marino roŋo	Morouas ronea
Nahavaq -roŋ	Nggela roŋo	Piamatsina roŋo-i	Sa roŋo
Ske -roŋo	Tawaroga roŋo	Tutuba rŋoa	Vunapu ronoa

m·r|BE ALIVE · binario · 77 lenguas · enriquecimiento 35× · emergente m·3

'Are'are emauri	Anuta ma-uri	Axamb -maur	Emae mauri
Ifira-Mele mouri	Langalanga mouri	Longgu mauri	Mandri -maur
Motu mauri	Nasarian -maor	Nggela mauri	Nombotkote -maor
Port Sandwich -maur	Sa mur	Tawaroga məri	Unua -maur

k·l|dig · binario · 77 lenguas · enriquecimiento 38× · emergente 2·3

Akolet keyl	Ambae keli	Avava ki-il	Bantik kali
Bola keli	Kandas i_kil	Lamenu -kil	Lwepe kel
Maranao kali	Niue keli	Palawan Batak kali	Ramoaaina kil
Satawalese kkel	Tagbanwa kali	Tungag / Tungak / kel	Western Fijian keli

t·r|stand · binario · 62 lenguas · enriquecimiento 30× · emergente t·3

Akei turi	Aore turu	Fortsenal turi	Label tur
Malo turu	Morouas turi	Na'ahai -tur	Nduke turu
Neverver -tur	Patpatar tur	Roviana turu	Sursurunga tur
Tasmate turi	Tolai tur	Uripiv: Durpif -tur	Waropen otaro

m·n|laugh · binario · 58 lenguas · enriquecimiento 13×

Akei mana	Amblong mana	Apma man	Baetora mana
Ctl: Nalemba: Nar mana	Hoanya Ts82 munē	Maewo mana	Marino man
N: Havai: Lombaha mana	Narango man	Orkon mæn	Port Sandwich -man
Sa man	Tasmate mana	Uliveo -man	Vunapu mana

m·t·r|sleep · ternario · 54 lenguas · enriquecimiento 91× · emergente m·t·3

Ambae matur	Antanosy matory	Bara matory	Ctl: Nalemba: Nar matir
Fifti: Khatbol -matur	Lelepa maturu	Malo maturu	Merina_Malagasy matòry
N: Leonda matur	Namagir matir	Neverver -matur	Papitalai metir
S: Raga: Avavanva maturu	Sivisa Titan matir	Tanala matory	Uripiv: Durpif -matur

l·l|vomit · binario · 52 lenguas · enriquecimiento 23× · emergente 3·3

Akei lulua	Arop i-lulu	Bali-Vitu lualua	Ctl: Nalemba: Nar lulua
Lamenu -lulua	Löyöp lolo	Malalamai lulua	Mpotovoro -lulue
Na'ahai i-lulu	Nasarian -lalu	Neverver -lialu	Nombotkote -lualua
Siviti -lolo	Tasmate lulua	Uvea lulua	Vera'a luluo

x·l|dig · binario · 49 lenguas · enriquecimiento 106× · emergente 2·3

Baetora xil	Ctl: Nalemba: Nas xəl	Koro xil	Lemerig xil
Maewo xili	Marino xili	Mwotlap xil	Na'ahai i-xil
Navwien 162-87 -xəl	Neverver -xil	Nombotkote -xil	S: Raga: Avavanva xeli
Siviti -xil	Tangoa xeli	Tirax -xeli	V'ënen Taut -xəli

m·m|PISS · binario · 48 lenguas · enriquecimiento 159×

'Are'are mimi	Asumboa mimi	Bughotu mimi	Ghari mimi
Kwaio mimi	Lau mimi	Luangiua mimi	Maori mimi
Mbaelelea mimi	Mbirao mimi	Nggela mimi	Roviana mimi
Simbo mimi	Tanema mimi	Tongan mimi	Vangunu mimi

m·s|BE DEAD OR DIE · binario · 47 lenguas · enriquecimiento 21× · emergente m·2

Anejom mas	Atchin: Orap -mats̃	Babuza Ts82 macha	Fifti: Khatbol i-mas
Kusaie mis̃	Mandri -mæt̃	Motu mase	Nahavaq -mes̃
Navwien 162-87 -mæt̃	Nisvai -mats̃	Pangkumu -met̃	Port Sandwich -maɸ̃
Siviti -mes̃	Tesmbol -mes̃	Unua -mit̃	Whitesands -imis̃

l·n|hear · binario · 47 lenguas · enriquecimiento 20× · emergente 3·n

Amara loŋo	Arop a-loŋo	Bariai -loŋo	Biliau loŋ
Buma leŋi	Hiw ʔloŋ	Kapingamarangi longo	Kwaio loŋa-a
Lihir lan	Malasanga i-luŋ	Mangap-Mbula -leŋ	Minyaifuin loŋo
Sarangani Blaen liŋi	Takia -loŋ	Tungag / Tungak / loŋo	Vaeakau-Taumako longo

m·t|die · binario · 43 lenguas · enriquecimiento 56×

Alor m'atɛɛ	Antaisaka maty	Antandroy mate	Anuta mate
Bara maty	Betsimisaraka mate	Kusaghe mate	Lote mete
Malango mate	Merina maty	Mikea mate	Nggela mate
Sihanaka maty	Simbo mate	Tanala maty	Ughele mate

m·t·k|fear · ternario · 42 lenguas · enriquecimiento 94× · emergente m·t·2

Ambae matakū	Angavae -matak	Bantik matakū	Bieria -matak
Ctl: Nalemba: Nas matak	Futuna matakū	Ifira-Mele matakū	Leo Tuvalu matakū
Maori matakū	Namakir matak	Nombotkote -metak	Nukumanu matakū
Rapa Nui matakū	South Efate mtak	Tikopia matakū	Uvea matakū

t·n|burn · binario · 40 lenguas · enriquecimiento 10×

Amara tin	Anuta tungia	Bellona tuungia	Chamorro tunu
Gedaged tun i	Iban tunu	Kapingamarangi tungi	Lunga Lunga tunia
Minaveha tunu	Navut taon	Nukeria ttuni	Nyelâyū thîni
Sengseng tun	Takuu tuni	Tiang tan	Tuamotuan tuni

k·t|see · binario · 39 lenguas · enriquecimiento 18× · emergente 2·t

Anuki kita	Ayta Ambala akit	Ayta Mag-indi akit	Dawawa kita
Futuna kite	Inati kitæ	Mamanwa kita	Manobo kita
Nukeria kite	Pazih F69 kita-	Rapa Nui kite	Saliba kita
Sikaiana kite	Tagalog ki:ta	Tuamotuan kite	Uvea kitea

s·l|burn · binario · 35 lenguas · enriquecimiento 25× · emergente 2·3

Akei sulia	Aore sulia	Avava -sulai	Gimán slai
Lou sulī	Mafea sulia	Marovo sulua	Morouas sulia
Narango sula	Navut sulu	Neve'ei suli	Roviana sulu-a
Siviti -sul	Tangoa sulī	Tiale sului	Tutuba sulia

l·m|come · binario · 35 lenguas · enriquecimiento 16× · emergente 3·m

'Auhelawa laoma	Buhutu laoma	Ghanongga lame	Halia lama
Langalanga lamae	Longgu lamai	Mbaengguu lea_mai	Mumeng lam
Neverver -lɛm	Puyuma a_lamo	Sa'a lae_mai	Saliba lao-ma
Suau laoma	Tanema loma	Tubetube laoma	Vaghua lama

n·r|cry · binario · 35 lenguas · enriquecimiento 24× · emergente n·3

'Are'are nara	Baetora ŋara	Ctl: Nalemba: Nar ŋara	Fifti: Khatbol -ŋar
Maewo ŋara	Marino ŋar	Mpotovoro -ŋar	N: Havai: Lombaha ŋara
Nduke ŋire	Neverver -ŋar	Roria ŋar	Sakao ŋor
Siviti -ŋar	Talise ŋara	Tutuba nara	Valpei ŋara

Objetos y cultura

«Casa» con **n·m** (62 lenguas, 26×) y **b·l·j** (42, 271×) es el caso más nítido de dos palabras distintas para el mismo concepto, no dos variantes de una. Las cifras lo dicen: el segundo es un ternario raro con enriquecimiento diez veces mayor.

Lo mismo en «camino»: **s·l** (56, 39×) y **d·l·n** (38, 208×).

d·g·m|NEEDLE alcanza 564× con solo 41 lenguas — y §7.1 lo señala, junto a **b·l·j**|house y **s·l**|path, como candidato a préstamo: §6.2 explica por qué el enriquecimiento alto apunta a lo local antes que a lo antiguo. Una secuencia de tres consonantes poco frecuentes repetida para el mismo objeto en cuarenta y una lenguas es prácticamente imposible por azar.

t·l|ROPE · binario · 70 lenguas · enriquecimiento 26× · emergente **t·3**

Agutyanen tali	Antanosy taly	Balangaw ta'le	Buol tali
Guiangan toli	Iranun tali	Kalinga tali	Kankanay ta'li
Maloh tali	Maranao tali	Nasarian a-tal	Pukapuka taula
Sangil 'taḷi	Subanon tali	Toga tel	Vezo taly

n·m|house · binario · 62 lenguas · enriquecimiento 26×

'Are'are nima	Angavae ne-imʷə	Axamb na-im	Dori'o nima
Kayupulau [Kajupu numa	Magori numa	Mpotovoro ne-ime	Namakir næim
Nese naime	Nombotkote ne-imʷə	Port Sandwich na-im	Saliba numa
Suau numa	Tanimbili nomʷa	Tesmbol ne-ma	Vao ne-me

s·l|PATH OR ROAD · binario · 56 lenguas · enriquecimiento 39× · emergente **2·3**

Akei sala	Angavae a-sɛɛ	Baetora sala	Ctl-E: Tawet: Nav sala
Fortsenal sala	Maewo sala	Manam jala	Mbirao sala
Narango sal	Navwien 162-87 sal	Port Vato sili	Rotuman sala
Ske sal	Tambotalo sala	Tomoip sal	Wailapa sala

t·p|THATCH OR ROOF · binario · 46 lenguas · enriquecimiento 66× · emergente **t·1**

Alangan atop	Ayta Abellen atip	Babuyan atep	Bolaang Mongondow atop
Butuanon atop	Cuyonon atip	Hanunóo atop	Ilokano atep
Isinay atop	Itbayaten atep	Kankanaey atep	Maranao atep
Romblon atóp	Surigaonon atop	Tonsea atep	Western Bukidnon atep

b·l·j|house · ternario · 42 lenguas · enriquecimiento 271× · emergente **1·3·j**

Agutyanen balay	Atta balay	Bikol balay	Binukid 'balay
Cuyonon baláy	Guiangan bilaʷy	Hiligaynon balay	Ilokano balay
Isneg balay	Kalagan balay	Kinaray-a balay	Manobo balɨy
Masbatenyo balay	Sinauna baláy	Subanun balay	Tagbanwa balay

k·j|stick · binario · 41 lenguas · enriquecimiento 132× · emergente **2·j**

Agta kɛyu	Babuyan kayu	Binukid 'kayu	Gaddang kayo
Inati káyo	Itawis kayu	Kaidipang kayu	Kalinga 'kayu
Maloh kaju	Manobo kayu	Manobo kayo	Nauna kɨy
S.W. Palawano kayo	Sarangani Blaan kayo	Tagbanwa kayu	Western Bukidnon kayu

d·g·m|NEEDLE (FOR SEWING) · ternario · 41 lenguas · enriquecimiento 564×

Agta dɔgum	Atta da:gum	Bikol dagom	Butuanon dagom
Dumagat di'gum	Hanunóo dagum	Ilokano dagum	Isneg da:gum
Kagayanen dágəm	Kallahan dagum	Kankanay 'dagom	Manobo dagom
Manobo dagum	Mansaka dagɛm	Pangasinan dagum	Surigaonon dagom

d·l·n|PATH OR ROAD · ternario · 38 lenguas · enriquecimiento 208× · emergente d·3·n

Agutyanen dalan	Atta dalan	Batangan dalan	Buol dolan
Cuyonon dalán	Gaddang dallan	Ibanag dalan	Iraya dalan
Kallahan dalan	Maloh dalan	Masbatenyo dalan	Pangasinan dalan
Ratagnon dalan	Sangir daleŋ	Tadyawan dalan	Tiruray dolon

Cualidades

Y aquí el anexo se cierra con su propio contraejemplo, que es lo más útil que contiene.

Los enriquecimientos de esta sección son los más bajos de todo el anexo: **m·t|black** 8×, **n·n|yellow** 10×, **m·t|right** 16×, **m·r|left** 26×. La excepción es **m·m·s|dry** con 118×.

La causa está identificada y es sistemática. Muchos adjetivos austronesios llevan el prefijo estativo *ma-*, que pone una *m* inicial en todos ellos. El código no puede distinguir ese prefijo de una *m* radical (§5.4), así que los agrupa por lo que tienen en común —el prefijo— en vez de por su raíz. El resultado son códigos que reúnen muchas lenguas y dicen poco, y el enriquecimiento lo detecta.

Léase esta sección como el aviso de dónde el método rinde peor: en conceptos cuya morfología introduce material compartido que no es la raíz. Es el argumento más fuerte a favor de resolver la segmentación.

m·t|right · binario · 71 lenguas · enriquecimiento 16×

Akei matua-	Aore matua-	Ctl-E: Tawet: Nav matua	Futuna-Aniwa matau
Kokota mautu	Lehali moto	Lonwolwol mætu	Maori matau
Mota matua	Narango matua	Pukapuka mātau	Simbo matua
Tangoa mātua-	Tikopia matau	Uliveo matu	Vurës moto

m·r|left · binario · 68 lenguas · enriquecimiento 26× · emergente m·3

'Are'are mauri	Atchin: Orap mʷir	Buma mɔrɔ	Eton mour
Kwamera mour	Lelepa maur	Malo maŋau	Nahavaq meir
Neverver mʷɛr	Piamatsina marau	Ririo mɔr	Sisingga məri
Sye mor	Tesmbol mæɛr	Tutuba marau	Vano mʷaure

m·m·s|dry · ternario · 46 lenguas · enriquecimiento 118× · emergente m·m·2

Atchin: Orap -memes	Baetora mamasa	Burmbar -mamas	Kove mamasa
Löyöp mamás	Malo mamasa	Mbwenelang -memes	Mutu mamás
Nisvai -mimis	Numbami -mamasa	Patpatar mamasa	Rotuman mamasa
Sio mamɔsa	Talise mamača	Tesmbol -memes	Wab mamos

m·t|black · binario · 39 lenguas · enriquecimiento 8×

Ambae maeto	Angavae -mεtε	Antanosy meity	Bara mety
Betsileo maity	Ctl: Nalemba: Nas maito	Maewo maeto	N: Leonda maeto
Nese -mot	Ninde -mite	Nombotkote i-mεt	Piamatsina maeto
Tape -mit	Tesmbol -miεt	V'ënen Taut i-mit	W: Walaha maeto

n·n|yellow · binario · 36 lenguas · enriquecimiento 10×

Avok -ianjan	Blablanga ηɔɲε	Burmbar -ɲaɲ	Hiw ɲaɲo
Lehali neɲoi	Leleghia ηɔɲε	Manam aɲoaɲo	Mengen 'ɲaɲo
Mwerlap aɲaɲ	Nauna aɲoan	Nggao ηɔɲε	Nisvai -ianjan
Port Sandwich -ɲaɲ	Tesmbol -ianjan	Uliveo -ianjan	Uripiv: Durpif ε-ɲaɲo