

Las formas del código endolingüístico

Cinco ejes, y por qué ninguna afirmación sobre «el código» es interpretable sin declararlos

Alejandro Toledo Martínez

Qué clase de documento es éste. No es especulativo y no propone tesis nuevas. Ordena. El concepto de código endolingüístico creció —de palabra, de raíz, de red, con subclases, con procedencia, con estados de pérdida— y hoy la frase «el código de esta palabra es $\Lambda \cdot \Xi$ » puede querer decir cinco cosas distintas. Este texto pone los ejes uno al lado del otro, dice qué regla gobierna cada uno, y deja una lista de verificación. Cada cifra citada sale de una medición sobre el corpus, con su fuente.

1. Por qué hace falta

En 2025 el código era un objeto simple de enunciar: *una combinación acotada y ordenada de dos o tres clases OAS, extraída de la raíz normalizada de una palabra por vía de su pronunciación*. Sigue siendo cierto, y sigue siendo insuficiente — porque la práctica destapó que la misma definición admite lecturas que no dan el mismo objeto.

Tres ejemplos reales, los tres de este corpus:

- Lat. *rīvus* es $\Lambda \cdot \Phi \cdot \Sigma$ leído entero, y Λ en su descendiente esp. *río*. ¿Cuál es «el código»?
- Ing. *ignite* es $X \cdot \Xi \cdot \Theta$, pero el latín *ignis* del que viene es $\Xi \cdot \Xi \cdot \Sigma$: la velar se había nasalizado ([*ɪŋ.nis*]) y el inglés la restauró leyendo la letra. ¿Cuál es «el código» de esa red?
- Ing. *tooth* es $\Theta \cdot \Theta$, pero su red de coderivados es mayoritariamente $\Theta \cdot \Xi \cdot \Theta$. ¿Cuál se enseña?

Ninguna de las tres preguntas tiene respuesta hasta declarar en qué eje se está hablando. De ahí este documento.

2. Lo que un código es, y lo que no

La definición larga, que no cambia:

Un código endolingüístico es una combinación acotada y ordenada de dos o tres clases OAS —ellas mismas particiones discretas, ancladas articulatoriamente y sin significado, indexadas al sistema psíquico bajo estudio—, extraída de la raíz normalizada de una palabra por vía de su pronunciación, no de su ortografía. Su conjunto de clases no ordenado es su núcleo. El orden porta la cualidad. Un código no significa nada: orienta un campo de posibilidad semántica, observable sólo a través de las redes de coderivados que lo instancian.

Y las negaciones, que en la práctica sirven más:

No es una cadena de consonantes · no se lee de la ortografía · no es un morfema ni una etimología · no es un significado · no es la suma de los perfiles de sus clases · no es un conjunto (el orden cuenta) · no es de una palabra: es de una red.

3. Eje I · El nivel de abstracción — la escalera

El mismo material se puede leer a tres alturas, y son las tres que la disciplina ya tenía:

nivel	qué es	ejemplo	especificidad	cobertura
esqueleto	las consonantes canónicas reales	d · n · t	máxima	mínima
código	las clases ordenadas	θ · ∃ · θ	media	media
núcleo	las clases sin orden	{θ, ∃}	mínima	máxima

Medido sobre 19.623 redes con ≥ 3 lenguas: la cobertura media del valor modal es 56,4 % en el esqueleto, 65,1 % en el código y 73,7 % en el núcleo. (*network_code_profile*.)

La cobertura sube al subir y la especificidad baja. No hay una altura correcta: hay una por pregunta. Y toda afirmación debe decir a cuál se hizo.

Por debajo del esqueleto está el cuallo —la sensación consonántica originante, continua y pre-simbólica— y por encima del código está la molécula, el sistema enlazado de códigos que forma una palabra compuesta. Ninguno de los dos es un peldaño de esta escalera: son los suelos.

4. Eje II · La extensión leída — palabra o raíz

Un mismo objeto da códigos distintos según cuánto de la palabra se lea.

	código de RAÍZ	código de PALABRA OÍDA
contiene	la raíz sin afijos	lo que llega al oído, afijos incluidos
sirve para	afirmaciones comparativas y de descendencia	afirmaciones sobre legibilidad psíquica
el afijo es	ruido específico de cada lengua	parte del dato: el oyente no segmenta antes de resonar
se cruzan	nunca dentro de una misma prueba	

Y el afijo no es simplemente ruido, por dos razones y la segunda es la fuerte: el oyente recibe la palabra entera; y así surgen los temas de las raíces — una ampliación se pega y deja de ser afijo, de modo que la frontera raíz/afijo no es ontológica sino diacrónica.

Estado del dato: de las formas del corpus, 539.456 tienen **core_skeleton** y en el 100 % de las redes donde está calculado el código de raíz difiere del de palabra. La elección nunca es inocua.

Y hay un tercer estado que no es ninguno de los dos: la raíz desnuda — la palabra que es su raíz, donde el código de palabra ya es el correcto. Medido: 592.527 formas están certificadas así, y reconocerlas llevó la población utilizable de la prueba de emergencia de 3.232 a 70.260 redes.

5. Eje III · El portador — la forma o la red

Una palabra no tiene código. Pertenece a una red coderivada, y el código es lo que se conserva en ella. *Teil* es $\Theta \cdot \Lambda$ porque está en la red *Teil-deel-del-dails*. Un patrón de consonantes sin red no es un código: es un patrón de consonantes.

De ahí que el objeto medible no sea un valor sino un perfil de red: el valor modal en cada nivel, su cobertura, y cuántos valores distintos conviven.

Ejemplo trabajado — TOOTH, red colapsada, 138 lenguas:

nivel	modal	cobertura	valores distintos
esqueleto	d · n · t	21,0 %	39
código	θ · Ξ · θ	34,8 %	24
núcleo	{θ, Ξ}	58,7 %	12

El inglés *tooth* (θ · θ) es el segundo, con 17,4 %.

Dos unidades de conteo, y una tercera que no existe. Se cuenta por lengua —a cuántas lenguas le sirve— y por rama —si el modal cruza ramas o está concentrado—. Por familia no se puede: medido, ninguna red coderivada cruza dos familias (63.321 en una sola, cero en más de una). Es definicional, no un umbral: un conjunto cognado es intra-familiar por construcción. La unidad «familia» de A-07 gobierna afirmaciones tipológicas, que son otra cosa.

Y las redes hay que colapsarlas antes de medir nada. El mismo etimón aparece partido en varios conjuntos: 157.327 conjuntos cognados se reducen a 69.416 redes. Sin ese colapso, una prueba puede salir significativa por contar el mismo etimón varias veces — ya ocurrió ($p = 0,0155 \rightarrow 0,133$).

6. Eje IV · La procedencia — de dónde le vino el código a la red

Atributo de primera clase, con cuatro estados:

procedencia	qué significa	testigo
heredado	el código descende de la raíz	esl. *bēlō [?] *bhelH-
sobrevenido	el código surgió durante la historia de la lengua	lat. <i>bellus</i> [?] *duenelos, donde la B-L no estaba
prestado	el código llegó con la palabra	esp. <i>bello</i> [?] occitano <i>bel</i>
restituido	una clase que la lengua hablada había perdido vuelve por lectura de la escritura	ing. <i>ignite</i> X·Ξ·θ frente a lat. <i>ignis</i> Ξ·Ξ·Σ

La cuarta es nueva, y su firma hubo que corregirla al computarla — vale la pena dejar dicho por qué. La primera versión decía «el préstamo tiene más clases que su étimo». Falla con su propio testigo: *ignite* X·Ξ·θ frente a *ignis* Ξ·Ξ·Σ tiene las mismas tres clases. La restitución no es un alargamiento: es una sustitución — vuelve la X donde el latín hablado había puesto Ξ ([ˈiɲ.nɪs]).

Firma correcta, y exige TRES generaciones: el hijo recupera una clase que el ABUELO tenía y el PADRE había perdido. Es el patrón clásico de *coincidir con el antepasado contra el intermedio*.

Computada (code_provenance, 183.250 cadenas de tres generaciones con código): 3.369 formas restituidas, de las cuales 21 siguen el patrón exacto del *gn* latino. Y la firma encuentra el fenómeno por su cuenta: ing. *magnes* — abuelo M·X·Ξ·Σ → lat. *magnes* [ˈmaɲ.ne:s] M·Ξ·Ξ·Σ → ing. M·X·Ξ·Σ. Instancia independiente de *ignis*, hallada sin buscarla.

Y su recall está limitado por el dato: *ignite* no cae en la firma, porque su cadena concreta no tiene abuelo con código. La firma es específica pero incompleta, y así se declara.

«Convergente» no entra en el nombre de ninguna categoría. La convergencia es la hipótesis, no un rasgo del objeto; meterla en el nombre hornea la conclusión en el vocabulario.

7. Eje V · El estado de cada posición — presente, oculto, desestructurado, ausente

Un código puede acortarse por tres razones distintas que el modelo escribía igual. Dos son sucesos de la lengua; una es un suceso nuestro.

escritura	estado	¿en la palabra?	¿en el código?	dónde ocurrió
Φ	presente	sí	sí	—
$[\Phi]$	oculto — el instrumento no lo muestra	sí	sí, invisible	en nosotros
(Φ)	desestructurado — pasó a la capa cuálica	sí	no	en la lengua
$\emptyset$	ausente — ya no está en la forma	no	no	en la lengua

Ejemplos: esp. *río* = $\Lambda \cdot (\Phi)$ — la *w* de *rīvus*, de cara a la vocal, se disolvió en ella; al. *rein* $[\text{?}]$ *hrainiz = $\emptyset \cdot \Lambda \cdot \Xi$ — la *h* se perdió sin tránsito; y todo /j/ del corpus está [**oculto**], porque la celda J no tiene clase OAS asignada: 320.845 segmentos invisibles al código, que es ceguera nuestra y no silencio del objeto.

8. Refinamiento transversal · las subclases

Dentro de una clase puede haber una distinción que la lengua sí transmite y que la clase aplana. Decidido y medido para Λ :

$$\Lambda \supset \{\Lambda_r, \Lambda_l\}$$

Luna y *runa* se tratan distintas, conectadas en el padre. Siguen siendo siete clases: una subclase no crea clase — vuelve nombrable, bajo su padre, una distinción que el esqueleto ya guardaba.

Medido: la subclase se conserva en el 94 % de las redes de coderivados (0,983 frente a un nulo de 0,733). Y su 6 % de excepciones no es ruido: es un catálogo de 1.245 casos del cambio $L \leftrightarrow R$, que se ordena solo en fusión indoiraniana, heteroclisia heredada y alternancia de sufijo.

Y la regla de método que sale de ahí: partir una clase sólo puede bajar o igualar la conservación del código padre — nunca subirla. El test que corresponde no es «¿mejora la conservación?» sino «¿se conserva la subclase?».

9. La rejilla completa

Una afirmación sobre un código queda determinada por cinco coordenadas:

nivel	×	extensión	×	portador	×	procedencia	×	estado
esqueleto		palabra		forma		heredado		presente
código		raíz		red		sobrevenido		oculto
núcleo		(raíz desnuda)				prestado		desestructurado
						restituido		ausente

Y la mayoría de los desacuerdos sobre códigos son desacuerdos de coordenada, no de hecho. *Tooth* es $\Theta \cdot \Theta$ y su red es $\Theta \cdot \Xi \cdot \Theta$: las dos son verdaderas, en portadores distintos.

10. Las reglas que gobiernan cada eje

eje	regla
nivel	C-10' el código es una escalera; se declara el peldaño
nivel	C-25 el código es binario o ternario, porque las raíces lo son; 4+ clases es una palabra con afijos
nivel	C-21 el orden es portante: con 69.244 pares, el revuelto está plano — un código no es un conjunto
extensión	C-10 nunca se cruzan los dos códigos dentro de una misma prueba
portador	C-19 el código es propiedad de una red, no de una palabra
portador	C-22 colapsar redes duplicadas antes de emparejar; sin ese control ya salió un falso positivo
portador	C-20 el nulo se sortea entre objetos de la misma clase: si se comparan redes, se sortean redes
procedencia	C-18 cuatro estados; «convergente» no entra en el nombre
estado	PE-01...04 toda posición perdida declara su estado y dónde ocurrió la pérdida
todos	C-15 un código de dos clases lo llevan 513 palabras: compartirlo no es evidencia
todos	C-26 pero C-15 pesa la evidencia, no define el objeto: filtrar a ≥ 4 clases excluye los códigos reales

C-15 y C-25 conviven así, y es la arquitectura que la teoría siempre tuvo: el código es binario o ternario y por eso no identifica por sí solo. Quien identifica son la red y el campo.

11. Lista de verificación

Cualquier lectura de código debería poder contestar:

1. ¿En qué nivel — esqueleto, código o núcleo?
2. ¿Sobre palabra o raíz? Y si es palabra, ¿es raíz desnuda certificada?
3. ¿El portador es una forma o una red? Si es red, ¿se colapsaron los duplicados?
4. ¿Cuál es la cobertura del valor modal, y en qué unidad — lenguas o ramas?
5. ¿Qué procedencia tiene el código en esa red?
6. ¿Alguna posición está oculta, desestructurada o ausente? ¿La pérdida fue de la lengua o del instrumento?
7. ¿Cuántas palabras llevan ese valor — es decir, cuánto pesa compartirlo?

Las siete son contestables hoy con el corpus. La séptima no está precomputada pero se calcula en 115 ms por valor; falta un índice sobre `skeleton.code`.

12. Lo que queda abierto

- El óptimo didáctico. Se tiene la cobertura (eje I $\times$ III) y se puede calcular la especificidad (C-15); falta cruzarlas para elegir peldaño con criterio y no a ojo.
- La deduplicación de lenguas. 120 glottocodes tienen varios `lect_id` y 1.387 `lects` no tienen glottocode: la cobertura por lengua está inflada en una cantidad que no se ha medido.

- La celda J. Sin clase OAS, todo /j/ es invisible. Es el hueco más grande del inventario.
- Y la pregunta de fondo, que ningún eje contesta: los cinco ejes dicen cómo se lee un código. Ninguno dice qué detecta. Esa sigue siendo la pregunta de la disciplina: lo que permanece a través de la genealogía no es el código, sino una continuidad estructural que el código nos deja observar.