

Geometrías semánticas comparadas

El núcleo pan-indoeuropeo y lo propio de cada sistema (germánico, romance, eslavo)

Alejandro Toledo Martínez

Segunda parte del programa *la polisemia como dato*: si cada sistema lingüístico construye su propio espacio semántico histórico, ¿cuánto comparten y en qué difieren? Comparamos las redes de colexificación de tres sistemas indoeuropeos —germánico, romance y eslavo— en el dominio cognición-percepción. Emerge un núcleo pan-indoeuropeo pequeño y estable, pero cada sistema conserva su *propia geometría*: ni espacios idénticos (contra una semántica universal) ni inconmensurables. La comparación se hace *controlando por cobertura* y con la genealogía y la cognación como árbitro.

1 Los tres sistemas

Con el mismo pipeline (IDS + NorthEuraLex → Concepticon; cognación híbrida etimológica/esqueleto con procedencia declarada) construimos, por sistema, la red de colexificación del dominio —una lengua colexifica dos conceptos cuando una sola forma realiza ambos (François, 2008)—. La identidad de lengua es canónica (una lengua en IDS y NEL cuenta una vez).

sistema	lenguas	colexif.	cognación etim.
Germánico	18	123	33%
Romance	14	42	69%
Eslavo	15	28	30%

Table 1: Panorama. El germánico aporta más lenguas y más colexificaciones; el romance tiene la cognación etimológica más limpia (anclaje latino, 69 %); el eslavo alcanza 30 % tras romanizar el cirílico (ruso, búlgaro) para emparejarlo con la transliteración latina de IDS —sus mayores aportes etimológicos son justamente el ruso y el búlgaro—.

La asimetría de datos importa y se declara: el germánico, con más lenguas y etapas, produce redes mayores; las cifras absolutas no son directamente comparables, pero sí lo son las *coincidencias* concepto a concepto, que es lo que medimos.

2 El núcleo pan-indoeuropeo

Cruzando los tres sistemas —y contando una colexificación como ausente en un sistema *solo* si ambos conceptos están atestiguados en él (control de cobertura)—, 9 colexificaciones aparecen en los tres:

- BRIGHT ↔ CLEAR
- LEARN ↔ TEACH
- LIGHT (COLOR) ↔ BRIGHT
- LIGHT (COLOR) ↔ CLEAR
- SALTY ↔ BRACKISH

- SMELL (PERCEIVE) ↔ SMELL (STINK)
- SMELL (PERCEIVE) ↔ SNIFF
- SOUND OR NOISE ↔ TONE (MUSIC)
- THINK (REFLECT) ↔ THINK (BELIEVE)

Un núcleo compartido, no una semántica universal. Que percibir/heder un olor, reflexionar/creer, aprender/enseñar o claro/brillante colexifiquen en germánico, romance y eslavo indica una capa semántica común —sea por herencia indoeuropea, sea por tendencia cognitiva universal (Youn et al., 2016)—. Pero es un núcleo *pequeño* frente a lo propio de cada sistema (§4): la evidencia no respalda que las lenguas compartan una sola geometría semántica, sino un *esqueleto* común vestido de forma distinta en cada tradición.

3 Distancias semánticas entre sistemas

Medimos la distancia como 1 menos el Jaccard de las colexificaciones *elegibles en ambos* sistemas.

	Germánico	Romance	Eslavo
Germánico	–	0.14	0.17
Romance	0.14	–	0.21
Eslavo	0.17	0.21	–

Table 2: Solapamiento (índice de Jaccard) de las redes de colexificación por pares. Valores bajos: cada sistema tiene mucha colexificación propia. El par romance–eslavo es el más cercano (0,21).

El solapamiento es bajo en todos los pares (0,14–0,21): la mayor parte de la colexificación de cada sistema es *suya*. Y —con cautela por la asimetría de datos— el orden no reproduce limpiamente el árbol indoeuropeo (el par romance–eslavo resulta el más cercano): la *geometría semántica* de un sistema no es una simple función de su distancia filogenética, lo que abre la pregunta de H5.

4 Lo propio de cada sistema

sistema	n	ejemplos propios
Germánico	90	INTENTION ↔ MIND; STINKING ↔ DIFFICULT; CLEVER ↔ TEACH; THINK (REFLECT) ↔ UNDERSTAND
Romance	19	SALTY ↔ DIRTY; LOUD ↔ CLEAR; FEEL ↔ SMELL (PERCEIVE); WET ↔ DAMP
Eslavo	3	HOT ↔ BITTER; DOUBT ↔ SUSPECT; SHOW ↔ SEEM

Table 3: Colexificaciones presentes en un solo sistema (siendo elegibles en los tres). Cada sistema recorta el espacio a su manera: el germánico funde percepción y cognición (SEE ↔ KNOW); el romance, matices sensoriales (SOFT ↔ SWEET, WET ↔ DAMP).

Además, muchas colexificaciones aparecen en dos de los tres sistemas (p.ej. HEAVY / DIFFICULT, LIGHT / EASY, HEAR / UNDERSTAND, es decir percepción → cognición), un anillo intermedio entre el núcleo y lo idiosincrásico.

5 El gradiente de herencia

¿Por qué existe el núcleo? Reconstruyendo cada colexificación germánica sobre su árbol y preguntando si recurre en las ramas hermanas, las heredadas (reconstruidas al proto-germánico) reaparecen en romance o eslavo el 88 % de las veces, frente al 18 % de las superficiales (internas de una subrama o innovaciones). El núcleo pan-indoeuropeo es, en su mayoría, *polisemia profunda heredada*. Matiz honesto: la recurrencia mezcla herencia común con tendencias universales del cambio semántico (Wilkins, 1996); la magnitud separa lo profundo de lo superficial, pero no aísla el mecanismo.

6 Límites

(i) Asimetría de datos: el germánico tiene más lenguas/etapas, luego más colexificaciones —los recuentos absolutos favorecen al sistema mejor documentado—. (ii) Eslavo sin etapa antigua: 30 % de cognacia etimológica (tras romanizar el cirílico ruso/búlgaro), pero sin atestiguación antigua disponible, luego hay escisión y fusión (**mysl* ‘mente’, **zvok* ‘sonido’) pero no desplazamiento. (iii) Herencia vs universalidad: la recurrencia no distingue ambos mecanismos; hace falta la direccionalidad o un nulo sobre ramas no emparentadas. (iv) Concepticon: la comparación por Concept Sets hereda su segmentación del espacio semántico —una definición independiente del campo es el paso pendiente del programa—.

Síntesis. Tres sistemas indoeuropeos comparten un *núcleo* de 9 colexificaciones (percepción, cognición, luz, temperatura), mayoritariamente heredado (88 % de las heredadas recurre, vs 18 % de las superficiales); pero sus redes solo solapan 0,14–0,21, y cada uno recorta el espacio a su modo. Ni una semántica universal, ni islas inconmensurables: un esqueleto compartido con geometrías propias.

7 Hipótesis

H5 (topología $\propto$ distancia genealógica). La similitud de los grafos semánticos debería decrecer con la distancia filogenética; el hallazgo preliminar (romance–eslavo el par más cercano) sugiere que *no* de forma simple —posible efecto del contacto o de la tipología—, y merece un test controlado por cobertura y tamaño. H6 (direccionalidad como desconfusor). Restringir la comparación a desplazamientos *dirigidos* (percepción→cognición, concreto→abstracto; Traugott & Dasher, 2002) separaría herencia de tendencia universal: una colexificación heredada debe además compartir la *dirección* del cambio. En un dominio con gradiente de concreción (cuerpo→espacio→tiempo) esa direccionalidad ya se confirma (32 de 42 desplazamientos etimológicos, $p = 0,001$).

Reproducibilidad

`src/newcon/compare_systems.py` lee `newcon.db`, `newcon.romance.db`, `newcon.slavic.db` (construidas por el pipeline parametrizado `NC_SYSTEM`) y escribe las tablas de este documento en `docs/paper-newcon-compare/`. Datos: IDS, NorthEuraLex, Concepticon; etimologías de Wiktionary vía Kaikki.

References

François, A. (2008). Semantic maps and the typology of colexification. In *From Polysemy to Semantic Change*, 163–215. Benjamins.

- Rzyski, C. et al. (2020). The Database of Cross-Linguistic Colexifications (CLICS³). *Scientific Data* 7, 13.
- Traugott, E.C. & Dasher, R.B. (2002). *Regularity in Semantic Change*. Cambridge University Press.
- Wilkins, D.P. (1996). Natural tendencies of semantic change and the search for cognates. In *The Comparative Method Re-viewed*, 264–304. Oxford University Press.
- Youn, H. et al. (2016). On the universal structure of human lexical semantics. *PNAS* 113(7), 1766–1771.