

La polisemia como dato

Espacios semánticos genealógicos: reconstruir la polisemia germánica desde la cognación
(piloto cognición–percepción)

Alejandro Toledo Martínez

Concepticon proyecta todas las lenguas sobre un eje conceptual único anclado al inglés, y al hacerlo aplanar la genealogía y trata la polisemia como ruido a eliminar. Este trabajo invierte la flecha: cada *sistema lingüístico* construye su propio espacio semántico histórico, y lo reconstruimos hacia atrás, de las lenguas hijas al proto, con la polisemia como dato. Piloto: el sistema germánico, dominio cognición y percepción. El resultado no es una tabla de conceptos sino un grafo semántico fechado, donde la polisemia se vuelve *historia* —herencia, escisión, fusión, desplazamiento— y se blindar con un modelo nulo contra la coincidencia. No es una mejora de Concepticon: es otra forma de construir espacios semánticos, que sobreviviría a su desaparición —Concepticon se usa aquí solo como *coordinada de interoperabilidad*, no como la ontología.

La lingüística estructural, de Saussure a Concepticon y los mapas semánticos, abstrae la polisemia histórica para *estabilizar* el concepto; el significado se trata como un punto y su inestabilidad, como ruido. Aquí invertimos el gesto: tomamos esa inestabilidad histórica como el observable primario. La tesis no es que las lenguas compartan una geometría semántica universal, sino que *cada sistema lingüístico desarrolla la suya*, reconstruible desde su genealogía.

1 Datos y método

Es el proyecto hermano de los *códigos coderivativos*: aquel modela el significante (el esqueleto consonántico) por rama; éste, el significado (el campo conceptual), con la misma disciplina —por sistemas, hacia atrás, con la cognación como árbitro y un nulo como suelo—. Datos: IDS ([Key & Comrie, 2016](#)) (serie de diccionarios comparables, con profundidad diacrónica: gótico, inglés y alto alemán antiguos y medios, y modernas) y NorthEuraLex ([Dellert et al., 2020](#)) (siete modernas, con segmentos y glosa alemana), alineados a Concepticon ([List et al., 2016](#)). Dominio: los campos *Cognition* y *Sense perception* (235 conceptos). Escalera fechada reutilizada del sistema germánico de códigos coderivativos (subramas oriental/occidental/septentrional; etapas antigua/media/nueva).

La cognación se deriva de forma *híbrida y con procedencia declarada*: (i) etimológica —étimo protogermánico extraído de la etimología de Wiktionary/Kaikki ([Ylönen, 2022](#)), la de alta confianza—; (ii) por esqueleto consonántico normalizado a las correspondencias germánicas, solo como respaldo. La polisemia se operacionaliza como colexificación ([François, 2008](#)): una lengua colexifica dos conceptos cuando una sola forma realiza ambos. La identidad de lengua es *canónica*: una lengua moderna presente en IDS y en NEL cuenta como una, no dos, para no inflar los recuentos.

2 La red de colexificación (sincronía)

Reunidas las siete/once lenguas, emergen 123 aristas concepto–concepto de contenido. Las más extendidas reproducen colexificaciones bien conocidas y confirmadas por la base translingüística CLICS ([Rzymiski et al., 2020](#)): percibir/heder un olor, aprender/enseñar, reflexionar/ creer, claro-de-color/brillante, ligero-de-peso/fácil, pesado/difícil.

métrica	valor
Formas (no préstamo)	2786
Nodos cognados	314
— etimológicos	113
Formas con cognacia etimológica	438 (33%)
Aristas de colexificación	171
— de contenido	123

Table 1: Auditoría del piloto. La cognacia etimológica (33 % de las formas en nodo) es minoritaria pero de alta confianza: no sobre-fusiona (campo medio 1,57 conceptos frente a 2,52 del respaldo por esqueleto).

concepto	concepto	lg	estatus
THINK (REFLECT)	THINK (BELIEVE)	9	PGmc-heredada
SMELL (PERCEIVE)	SMELL (STINK)	9	PGmc-heredada
LEARN	TEACH	7	PGmc-heredada
LIGHT (COLOR)	BRIGHT	7	PGmc-heredada
OBSCURE	DARK	7	PGmc-heredada
LIGHT (WEIGHT)	EASY	6	PGmc-heredada
THINK (BELIEVE)	BELIEVE	6	PGmc-heredada
CLEVER	WISE	5	PGmc-heredada
BRIGHT	CLEAR	4	PGmc-heredada
WARM	HOT	4	PGmc-heredada
HEAVY	DIFFICULT	4	antigua-subrama
SEE	LOOK	3	PGmc-heredada

Table 2: Colexificaciones de contenido más extendidas (nº de lenguas germánicas) y su estatus diacrónico.

3 La polisemia como historia (diacronía)

Cada colexificación se reconstruye sobre el árbol por parsimonia. Si está presente en ≥ 2 de las tres subramas, se reconstruye al proto-germánico; 16 colexificaciones de contenido lo son (p.ej. REFLEXIONAR ↔ CREER, APRENDER ↔ ENSEÑAR, LISTO ↔ SABIO). La polisemia mayoritaria del sistema es, pues, heredada, no una moda reciente.

estatus diacrónico	n
Heredada al proto-germánico (≥ 2 subramas)	16
Antigua, una subrama	90
Moderna difusa	0
Innovación / local	17

Table 3: Colexificaciones de contenido por estatus diacrónico.

Con la cognacia etimológica, la polisemia se descompone en las cuatro operaciones del cambio semántico (Wilkins, 1996; Traugott & Dasher, 2002):

El caso *glad*. El sistema detecta solo un desplazamiento de manual: el étimo germánico **glad-* colexifica en las lenguas antiguas BRILLANTE/CLARO/LUZ, y en las modernas se desplaza a LISO (y, fuera del dominio, al inglés *glad* ‘contento’). El mismo esqueleto, una historia semántica. Concepticon lo borraría; aquí es el dato.

tipo	étimo	campo / cambio
Escisión	<i>*bi-</i>	BETRAY, TOUCH
Escisión	<i>*dol</i>	BLUNT, MAD, STUPID
Escisión	<i>*glad</i>	BRIGHT, CLEAR, LIGHT (COLOR), SMOOTH
Escisión	<i>*hlūd</i>	LISTEN, LOUD, SOUND OR NOISE
Escisión	<i>*hug</i>	MIND, REMEMBER, THINK (REFLECT)
Polisemia estable	<i>*derk</i>	DARK, OBSCURE
Polisemia estable	<i>*dunkal</i>	DARK, OBSCURE
Polisemia estable	<i>*frōd</i>	CLEVER, WISE
Polisemia estable	<i>*galaub</i>	BELIEVE, THINK (BELIEVE)
Desplazamiento	<i>*dol</i>	STUPID → BLUNT, MAD
Desplazamiento	<i>*glad</i>	BRIGHT, CLEAR, LIGHT (COLOR) → SMOOTH
Desplazamiento	<i>*hug</i>	MIND, THINK (REFLECT) → REMEMBER
Desplazamiento	<i>*mann-</i>	PUPIL (STUDENT) → MANNER

Table 4: Operaciones del cambio semántico leídas de étimos reales. La *escisión* diferencia un étimo entre hijas; la *polisemia estable* mantiene una forma con varios sentidos; el *desplazamiento* mueve el campo del reflejo antiguo al nuevo. La fusión (no en la tabla) es un concepto servido por varios étimos: SABER ← **wit* / **kunn* / **knā* / **kann*.

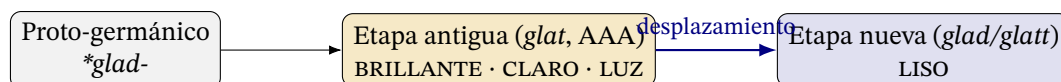


Figure 1: Reconstrucción diacrónica de una colexificación. El étimo **glad-* realiza un campo ‘luminoso’ en la etapa antigua y se desplaza a LISO en la moderna (y, fuera del dominio, al inglés *glad* ‘contento’). El mismo esqueleto consonántico soporta una trayectoria semántica fechada.

4 Conservación del campo (decaimiento semántico)

¿Se conserva el campo de un étimo con el tiempo, como se conserva su código? Formalicemos: el *campo* de un étimo en una etapa es el conjunto de Concept Sets que sus reflejos realizan en ella; su conservación es el índice de Jaccard $J = |C_{\text{ant}} \cap C_{\text{nue}}| / |C_{\text{ant}} \cup C_{\text{nue}}|$ entre el campo antiguo y el nuevo ($J = 1$ conservación total, $J = 0$ desplazamiento total). De los 63 étimos que abarcan de la etapa antigua a la nueva, el Jaccard medio es $\bar{J} = 0.73$; con umbral $J \geq 0.5$, 51 (81 %) conservan el campo, 5 lo hacen parcialmente y 7 se desplazan. El campo semántico es, así, mayoritariamente *estable* —un decaimiento lento, análogo al del esqueleto consonántico en los códigos coderivativos y a la estabilidad léxica de las listas de tipo Leipzig-Jakarta (Tadmor, 2009)—; los pocos desplazamientos son los casos interesantes (**glad*, **dol* ESTÚPIDO → ROMO/LOCO). *Límite declarado*: esta definición del campo hereda la segmentación de Concepticon; una definición plenamente independiente —un espacio semántico propio del sistema— es el siguiente paso del programa.

5 El nulo ③’ (contra la apofenia)

¿Es esta red azar? Un test de permutación baraja, dentro de cada lengua, qué conceptos realiza cada forma (conservando cuántas formas colexifican y la frecuencia de cada concepto) y recuenta las colexificaciones que recurren en ≥ 3 lenguas. Observadas: 14; nulo: prácticamente cero (0.1 ± 0.3); $z = 49.4$. Bajo azar casi ninguna colexificación se repite en tres lenguas: la polisemia compartida es estructura heredada, no coincidencia.

El valor de z es enorme porque el nulo es casi *degenerado*: bajo permutación prácticamente ninguna colexificación recurre en tres o más lenguas. Es robusto al umbral k : al exigir más lenguas, lo observado decae de forma gradual mientras el nulo permanece en cero.

umbral k (lenguas)	observado	nulo	z
≥ 2	27	5.58	8.7
≥ 3	14	0.07	49.4
≥ 4	11	0.0	> 200
≥ 5	8	0.0	> 200

Table 5: Sensibilidad del nulo ③' al umbral k (nº mínimo de lenguas). Lo observado decrece suavemente; el nulo se mantiene ≈ 0 , de modo que a partir de $k \geq 4$ el nulo nunca alcanza recurrencia y z es ilimitado. La señal no depende de un umbral afortunado.

Cinco resultados del piloto. (1) La red de colexificación germánica reproduce y fecha las colexificaciones que CLICS registra en plano. (2) Su parte mayoritaria es *heredada*, reconstruible al proto-germánico. (3) La cognacia etimológica descompone la polisemia en escisión, fusión y desplazamiento sobre étimos reales. (4) El campo semántico se conserva ($\bar{J} = 0.73$; 81 % con $J \geq 0.5$) en el tiempo: decaimiento lento. (5) El nulo ③' ($z = 49.4$, robusto al umbral) descarta el azar. La polisemia deja de ser ruido: es el objeto.

6 Prueba de H1: ¿la polisemia heredada recurre en las ramas hermanas?

Con las mismas herramientas construimos la red de colexificación de dos sistemas hermanos (romance: latín más seis modernas; eslavo: antiguo eslavo eclesiástico más nueve) y preguntamos si las colexificaciones germánicas reaparecen en ellos, contando la ausencia *solo* cuando ambos conceptos están atestiguados en la rama hermana (control de cobertura).

grupo germánico	en romance	en eslavo	en ≥ 1 hermana
Heredada (proto-germánico)	7/16 (44%)	12/16 (75%)	14/16 (88%)
Intermedia (una subrama)	6/90 (7%)	4/90 (4%)	9/90 (10%)
Local / innovación	7/17 (41%)	6/17 (35%)	10/17 (59%)

Table 6: Recurrencia de las colexificaciones germánicas en romance y eslavo, por su estatus diacrónico.

Las heredadas (reconstruidas al proto-germánico) recurren en una rama hermana el 88 % de las veces; las superficiales (internas de una subrama o innovaciones locales), solo el 18 %. El gradiente confirma H1: la polisemia profunda es indoeuropea, no germánica. *Matiz honesto*: la recurrencia de las innovaciones locales (59 %) es mayor de lo esperado —la recurrencia cross-rama mezcla herencia común con tendencias universales del cambio semántico (LIGHT/EASY, HEAVY/DIFFICULT son metáforas que cualquier rama innova por su cuenta; Wilkins, 1996)—: la recurrencia sola no separa herencia de universal, aunque la magnitud (88 frente a 18 %) sí separa lo profundo de lo superficial.

7 Hipótesis derivadas

H1 (probada arriba). Las colexificaciones heredadas recurren en las ramas hermanas mucho más que las superficiales (88 vs 18 %); separar herencia de tendencia universal exige un control adicional (p.ej. direccionalidad, o el nulo sobre ramas no emparentadas). H2 (probada aparte, en el dominio cuerpo-espacio-tiempo). Los desplazamientos son *unidireccionales*: corriendo el mismo pipeline sobre un dominio con gradiente de concreción real (cuerpo $\rightarrow$ espacio $\rightarrow$ tiempo), 32 de 42 desplazamientos etimológicos limpios van en la dirección esperada ($p = 0.001$; Traugott & Dasher, 2002). En cognición-percepción, en cambio, el eje es demasiado plano para probarlo (casi no hay conceptos

concretos): la elección de dominio decide qué hipótesis es comprobable. H3 (puente con el significativo). La vida media del *campo* de un cognado debería correlacionar con la del *código* consonántico del mismo cognado —uniendo las dos caras del signo con una sola llave etimológica—. H4 (metalenguaje). Anclar el nodo al étimo y no a la glosa inglesa reduce el sesgo del inglés como metalenguaje; comprobable comparando la segmentación de campos bajo glosa inglesa frente a alemana (NEL) sobre los mismos datos. H5 (topología $\propto$ distancia). La topología del grafo semántico debería conservarse en proporción inversa a la distancia genealógica: cuanto más alejadas dos ramas, más degradado su grafo compartido —medible como una función decreciente de la similitud de grafos frente a la profundidad del MRCA—.

Este es el piloto de un dominio y un sistema; el diseño está pensado para escalar a miles de conceptos y a los demás sistemas lingüísticos (véase la bitácora del proyecto).

Reproducibilidad

Pipeline `src/newcon/build_s01..s5.py` $\rightarrow$ `data/db/newcon.db` (tablas `form`, `concept`, `stage_map`, `cognate_node`, `colexification`, `semantic_change`). Fragmentos y cifras generados por `build_s5.py` en `docs/paper-newcon/`. Datos primarios: IDS, NorthEuraLex, Concepticon; etimologías de Wiktionary vía Kaikki. Cognacia híbrida con procedencia declarada (`method=etym|skel`).

References

- Dellert, J. et al. (2020). NorthEuraLex: a wide-coverage lexical database of Northern Eurasia. *Language Resources and Evaluation* 54, 273–301.
- François, A. (2008). Semantic maps and the typology of colexification. In *From Polysemy to Semantic Change*, 163–215. Benjamins.
- Key, M.R. & Comrie, B. (eds.) (2016). *The Intercontinental Dictionary Series*. MPI-EVA, Leipzig.
- List, J.-M., Cysouw, M. & Forkel, R. (2016). Concepticon: a resource for the linking of concept lists. *Proc. LREC*, 2393–2400.
- Rzymiski, C. et al. (2020). The Database of Cross-Linguistic Colexifications (CLICS³). *Scientific Data* 7, 13.
- Tadmor, U. (2009). Loanwords in the world’s languages: findings and results. In *Loanwords in the World’s Languages*, 55–75. De Gruyter.
- 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.
- Ylönen, T. (2022). Wiktextextract: Wiktionary as machine-readable structured data. *Proc. LREC*, 1317–1325.