Polysemy as data

Genealogical semantic spaces: reconstructing Germanic polysemy from cognacy (cognition–perception pilot)

Alejandro Toledo Martínez

Concepticon projects every language onto a single, English-anchored conceptual axis, and in doing so flattens genealogy and treats polysemy as noise to be removed. This work reverses the arrow: each *linguistic system* builds its own historical semantic space, which we reconstruct backwards, from daughter languages to the proto, with polysemy as data. Pilot: the Germanic system, cognition and perception domain. The result is not a table of concepts but a dated semantic graph, where polysemy becomes *history* —inheritance, split, merger, shift— and is shielded against coincidence by a null model. This is not an improvement of Concepticon: it is another way of building semantic spaces, one that would survive its disappearance —Concepticon is used here only as an *interoperability coordinate*, not as the ontology.

Structural linguistics, from Saussure to Concepticon and semantic maps, abstracts away historical polysemy in order to *stabilise* the concept; meaning is treated as a point and its instability as noise. We reverse the move: we take that historical instability as the primary observable. The thesis is not that languages share a universal semantic geometry, but that *each linguistic system develops its own*, reconstructable from its genealogy.

1 Data and method

This is the sibling of the *coderivative codes* project: that one models the signifier (the consonant skeleton) per branch; this one, the signified (the conceptual field), with the same discipline —per system, backwards, with cognacy as arbiter and a null as floor—. Data: IDS (Key & Comrie, 2016) (comparable dictionary series, with diachronic depth: Gothic, Old and Middle English and High German, and moderns) and NorthEuraLex (Dellert et al., 2020) (seven moderns, with segments and a German gloss), linked to Concepticon (List et al., 2016). Domain: the *Cognition* and *Sense perception* fields (235 concepts). The dated ladder is reused from the Germanic system of the coderivative codes (Eastern/Western/Northern subbranches; Old/Middle/New stages).

Cognacy is derived in a *hybrid, provenance-declared* way: (i) etymological —Proto-Germanic etymon extracted from Wiktionary/Kaikki etymologies (Ylönen, 2022), the high-confidence tier—; (ii) by a consonant skeleton normalised to the Germanic correspondences, as a fallback only. Polysemy is operationalised as colexification (François, 2008): a language colexifies two concepts when a single form realises both. Language identity is *canonical*: a modern language present in both IDS and NEL counts as one, not two, so as not to inflate the counts.

2 The colexification network (synchrony)

Pooling the seven/eleven languages, 123 content concept–concept edges emerge. The most widespread reproduce well-known colexifications confirmed by the cross-linguistic database CLICS (Rzymski et al., 2020): smell-perceive/smell-stink, learn/teach, reflect/believe, light-in-colour/bright, light-in-weight/easy, heavy/difficult.

metric	value
Forms (non-loan)	2786
Cognate nodes	314
— etymological	113
Forms with etymological cognacy	438 (33%)
Colexification edges	171
— content-word	123

Table 1: Pilot audit. Etymological cognacy (33 % of forms in a node) is a minority but high-confidence tier: it does not over-merge (mean field 1.57 concepts vs. 2.52 for the skeleton fallback).

concept	concept	lg	status
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: Most widespread content colexifications (number of Germanic languages) and their diachronic status.

3 Polysemy as history (diachrony)

Each colexification is reconstructed over the tree by parsimony. If present in ≥2 of the three subbranches, it reconstructs to Proto-Germanic; 16 content colexifications do so (e.g. REFLECT↔BELIEVE, LEARN↔TEACH, CLEVER↔WISE). The bulk of the system's polysemy is therefore inherited, not a recent fashion.

diachronic status	n
Inherited to Proto-Germanic (≥2 subbranches)	16
Old, single subbranch	90
Modern, diffuse	0
Innovation / local	17

Table 3: Content colexifications by diachronic status.

With etymological cognacy, polysemy decomposes into the four operations of semantic change (Wilkins, 1996; Traugott & Dasher, 2002):

The *glad* case. The system detects, on its own, a textbook shift: the Germanic etymon **glad-* colexifies BRIGHT/CLEAR/LIGHT in the old languages, and in the moderns shifts to SMOOTH (and, outside the domain, to English *glad*). Same skeleton, one semantic history. Concepticon would erase it; here it is the datum.

type	etymon	field / change
Split	**bi-*	BETRAY, TOUCH
Split	**dol*	BLUNT, MAD, STUPID
Split	**glad*	BRIGHT, CLEAR, LIGHT (COLOR), SMOOTH
Split	**hlūd*	LISTEN, LOUD, SOUND OR NOISE
Split	**hug*	MIND, REMEMBER, THINK (REFLECT)
Stable polysemy	**derk*	DARK, OBSCURE
Stable polysemy	**dunkal*	DARK, OBSCURE
Stable polysemy	**frōd*	CLEVER, WISE
Stable polysemy	**galaub*	BELIEVE, THINK (BELIEVE)
Shift	**dol*	STUPID → BLUNT, MAD
Shift	**glad*	BRIGHT, CLEAR, LIGHT (COLOR) → SMOOTH
Shift	**hug*	MIND, THINK (REFLECT) → REMEMBER
Shift	**mann-*	PUPIL (STUDENT) → MANNER

Table 4: Operations of semantic change read from real etyma. A *split* differentiates an etymon across daughters; *stable polysemy* keeps one form with several senses; a *shift* moves the field from the old reflex to the new. A merger (not tabled) is one concept served by several etyma: KNOW ← **wit / *kunn / *knā / *kann*.

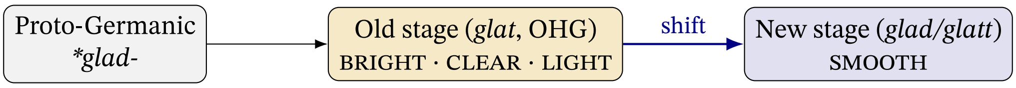

Figure 1: Diachronic reconstruction of a colexification. The etymon **glad-* realises a 'luminous' field at the old stage and shifts to SMOOTH in the modern one (and, outside the domain, to English *glad*). The same consonant skeleton carries a dated semantic trajectory.

4 Field conservation (semantic decay)

Is an etymon's field conserved over time, as its code is? Formally: an etymon's *field* at a stage is the set of Concept Sets its reflexes realise there; its conservation is the Jaccard index $J = |C_{\text{old}} \cap C_{\text{new}}|/|C_{\text{old}} \cup C_{\text{new}}|$ between the old and new fields ($J = 1$ full conservation, $J = 0$ full shift). Of the 63 etyma spanning old to new, the mean Jaccard is $\bar{J} = 0.73$; with threshold $J \geq 0.5$, 51 (81 %) conserve the field, 5 do so partially and 7 shift. The semantic field is thus mostly *stable* —a slow decay, analogous to that of the consonant skeleton in the coderivative codes and to the lexical stability of Leipzig-Jakarta-type lists (Tadmor, 2009)—; the few shifts are the interesting cases (**glad*, **dol* STUPID→BLUNT/MAD). *Stated limit:* this definition of the field inherits Concepticon's segmentation; a fully independent definition —a semantic space of the system's own— is the programme's next step.

5 The ③′ null (against apophenia)

Is this network chance? A permutation test shuffles, within each language, which concepts each form realises (preserving how many forms colexify and each concept's frequency) and recounts the colexifications recurring in ≥3 languages. Observed: 14; null: essentially zero (0.1 ± 0.3); $z = 49.4$. Under chance almost no colexification recurs across three languages: shared polysemy is inherited structure, not coincidence.

The z value is large because the null is nearly *degenerate*: under permutation almost no colexification recurs across three or more languages. It is robust to the threshold k: requiring more languages, the observed count decays gradually while the null stays at zero.

threshold k (langs)	observed	null	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: Sensitivity of the ③′ null to the threshold k (minimum number of languages). The observed count decays smoothly; the null stays ≈ 0, so from $k \geq 4$ the null never reaches recurrence and z is unbounded. The signal does not rely on a lucky threshold.

Five pilot results. (1) The Germanic colexification network reproduces and dates the colexifications CLICS records flat. (2) Its bulk is *inherited*, reconstructable to Proto-Germanic. (3) Etymological cognacy decomposes polysemy into split, merger and shift over real etyma. (4) The semantic field is conserved ($\bar{J} = 0.73$; 81 % at $J \geq 0.5$) over time: slow decay. (5) The ③′ null ($z = 49.4$, threshold-robust) rules out chance. Polysemy stops being noise: it is the object.

6 Testing H1: does inherited polysemy recur in the sister branches?

With the same tools we build the colexification network of two sibling systems —Romance (Latin plus six moderns) and Slavic (Old Church Slavonic plus nine)— and ask whether the Germanic colexifications reappear in them, counting absence *only* where both concepts are attested in the sister branch (coverage control).

Germanic group	in Romance	in Slavic	in ≥1 sister
Inherited (Proto-Germanic)	7/16 (44%)	12/16 (75%)	14/16 (88%)
Intermediate (one subbranch)	6/90 (7%)	4/90 (4%)	9/90 (10%)
Local / innovation	7/17 (41%)	6/17 (35%)	10/17 (59%)

Table 6: Recurrence of the Germanic colexifications in Romance and Slavic, by diachronic status.

The inherited ones (reconstructed to Proto-Germanic) recur in a sister branch 88 % of the time; the shallow ones (subbranch-internal or local innovations), only 18 %. The gradient confirms H1: deep polysemy is Indo-European, not Germanic. *Honest caveat:* the recurrence of local innovations (59 %) is higher than expected —cross-branch recurrence conflates common inheritance with universal tendencies of semantic change (LIGHT/EASY, HEAVY/DIFFICULT are metaphors any branch may innovate on its own; Wilkins, 1996)—: recurrence alone does not separate inheritance from universality, though the magnitude (88 vs 18 %) does separate the deep from the shallow.

7 Derived hypotheses

H1 (tested above). Inherited colexifications recur in the sister branches far more than shallow ones (88 vs 18 %); separating inheritance from universal tendency needs a further control (e.g. directionality, or the null over unrelated branches). H2 (tested separately, in the body–space–time domain). Shifts are *unidirectional*: running the same pipeline on a domain with a real concreteness gradient (body $\rightarrow$ space $\rightarrow$ time), 32 of 42 clean etymological shifts run in the expected direction ($p = 0.001$; Traugott & Dasher, 2002). In cognition–perception, by contrast, the axis is too flat to test (almost no concrete concepts): the choice of domain decides which hypothesis is testable. H3 (bridge to the signifier). The half-life of a cognate's *field* should correlate with that of its consonant *code* —joining the two faces of the sign through a single etymological key—. H4 (metalanguage). Anchoring the node to the etymon rather than the English gloss reduces the English-as-metalanguage bias; testable by comparing field segmentation under English vs. German (NEL) glosses on the same data. H5 (topology $\propto$ distance). The semantic graph's topology should be conserved in inverse proportion to genealogical distance: the more distant two branches, the more degraded their shared graph —measurable as a decreasing function of graph similarity against MRCA depth—.

This is the pilot of one domain and one system; the design is built to scale to thousands of concepts and to the other linguistic systems (see the project logbook).

Reproducibility

Pipeline `src/newcon/build_s01..s5.py` → `data/db/newcon.db` (tables `form`, `concept`, `stage_map`, `cognate_node`, `colexification`, `semantic_change`). Fragments and figures generated by `build_s5.py` in `docs/paper-newcon/`. Primary data: IDS, NorthEuraLex, Concepticon; etymologies from Wiktionary via Kaikki. Hybrid cognacy with declared provenance (`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.

Rzymski, C. et al. (2020). The Database of Cross-Linguistic Colexifications (CLICS^3). *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 Reviewed*, 264–304. Oxford University Press.

Ylönen, T. (2022). Wiktextract: Wiktionary as machine-readable structured data. *Proc. LREC*, 1317–1325.