Comparative semantic geometries

The pan-Indo-European core and what each system keeps of its own (Germanic, Romance, Slavic)

Alejandro Toledo Martínez

Second part of the *polysemy as data* programme: if each linguistic system builds its own historical semantic space, how much do they share and where do they differ? We compare the colexification networks of three Indo-European systems —Germanic, Romance and Slavic— in the cognition–perception domain. A small, stable pan-Indo-European core emerges, yet each system keeps its *own geometry*: neither identical spaces (against a universal semantics) nor incommensurable ones. The comparison is coverage-controlled, with genealogy and cognacy as arbiter.

1 The three systems

With the same pipeline (IDS + NorthEuraLex → Concepticon; hybrid etymological/skeleton cognacy with declared provenance) we build, per system, the domain's colexification network —a language colexifies two concepts when a single form realises both (François, 2008)—. Language identity is canonical (a language in both IDS and NEL counts once).

system	languages	colexif.	etym. cognacy
Germanic	18	123	33%
Romance	14	42	69%
Slavic	15	28	30%

Table 1: Overview. Germanic contributes more languages and more colexifications; Romance has the cleanest etymological cognacy (Latin anchoring, 69 %); Slavic reaches 30 % after romanising the Cyrillic (Russian, Bulgarian) to match IDS's Latin transliteration —its largest etymological contributors are in fact Russian and Bulgarian—.

Data asymmetry matters and is declared: Germanic, with more languages and stages, produces larger networks; the absolute counts are not directly comparable, but the concept-by-concept *matches* are —which is what we measure.

2 The pan-Indo-European core

Crossing the three systems —and counting a colexification as absent in a system *only* if both concepts are attested there (coverage control)—, 9 colexifications appear in all three:

- 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)

A shared core, not a universal semantics. That smell-perceive/stink, reflect/believe, learn/teach or clear/bright colexify in Germanic, Romance and Slavic points to a common semantic layer —whether by Indo-European inheritance or by universal cognitive tendency (Youn et al., 2016)—. But it is a *small* core against what each system keeps of its own (§4): the evidence does not support one shared semantic geometry, but a common *skeleton* dressed differently in each tradition.

3 Semantic distances between systems

We measure distance as 1 minus the Jaccard of the colexifications *eligible in both* systems.

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

Table 2: Overlap (Jaccard index) of the pairwise colexification networks. Low values: each system has much colexification of its own. The Romance–Slavic pair is the closest (0.21).

Overlap is low across all pairs (0.14–0.21): most of each system's colexification is *its own*. And —with caution owing to data asymmetry— the ordering does not cleanly reproduce the Indo-European tree (the Romance–Slavic pair comes out closest): a system's *semantic geometry* is not a simple function of its phylogenetic distance, which opens the H5 question.

4 What each system keeps of its own

system	n	own examples
Germanic	90	INTENTION ↔ MIND; STINKING ↔ DIFFICULT; CLEVER ↔ TEACH; THINK (REFLECT) ↔ UNDERSTAND
Romance	19	SALTY ↔ DIRTY; LOUD ↔ CLEAR; FEEL ↔ SMELL (PERCEIVE); WET ↔ DAMP
Slavic	3	HOT ↔ BITTER; DOUBT ↔ SUSPECT; SHOW ↔ SEEM

Table 3: Colexifications present in a single system (while eligible in all three). Each system carves the space its own way: Germanic fuses perception and cognition (SEE / KNOW); Romance, sensory nuances (SOFT / SWEET, WET / DAMP).

Moreover, many colexifications appear in two of the three systems (e.g. HEAVY / DIFFICULT, LIGHT / EASY, HEAR / UNDERSTAND, i.e. perception→cognition), an intermediate ring between the core and the idiosyncratic.

5 The inheritance gradient

Why does the core exist? Reconstructing each Germanic colexification over its tree and asking whether it recurs in the sister branches, the inherited ones (reconstructed to Proto-Germanic) reappear in Romance or Slavic 88 % of the time, versus 18 % for the shallow ones (subbranch-internal or innovations). The pan-Indo-European core is mostly *deep inherited polysemy*. Honest caveat: recurrence conflates common inheritance with universal tendencies of semantic change (Wilkins, 1996); the magnitude separates deep from shallow but does not isolate the mechanism.

6 Limits

(i) Data asymmetry: Germanic has more languages/stages, hence more colexifications —absolute counts favour the best-documented system—. (ii) Slavic without an old stage: 30 % etymological cognacy (after romanising Cyrillic Russian/Bulgarian), but no ancient attestation available, so it has split and merger (**mysl* 'mind', **zvǫk* 'sound') but no shift. (iii) Inheritance vs universality: recurrence does not distinguish the two mechanisms; directionality or a null over unrelated branches is needed. (iv) Concepticon: the Concept-Set comparison inherits its segmentation of the semantic space —an independent definition of the field is the programme's pending step—.

Synthesis. Three Indo-European systems share a *core* of 9 colexifications (perception, cognition, light, temperature), mostly inherited (88 % of the inherited recur, vs 18 % of the shallow); yet their networks overlap only 0.14–0.21, and each carves the space its own way. Neither a universal semantics nor incommensurable islands: a shared skeleton with distinct geometries.

7 Hypotheses

H5 (topology $\propto$ genealogical distance). Semantic-graph similarity should decrease with phylogenetic distance; the preliminary finding (Romance–Slavic the closest pair) suggests *not* in a simple way —a possible contact or typology effect— and deserves a test controlled for coverage and size. H6 (directionality as de-confounder). Restricting the comparison to *directed* shifts (perception→cognition, concrete→abstract; Traugott & Dasher, 2002) would separate inheritance from universal tendency: an inherited colexification should also share the *direction* of change. In a domain with a concreteness gradient (body→space→time) that directionality is already confirmed (32 of 42 etymological shifts, $p = 0.001$).

Reproducibility

`src/newcon/compare_systems.py` reads `newcon.db`, `newcon.romance.db`, `newcon.slavic.db` (built by the `NC_SYSTEM`-parameterised pipeline) and writes this document's tables to `docs/paper-newcon-compare/`. Data: IDS, NorthEuraLex, Concepticon; etymologies from Wiktionary via Kaikki.

References

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

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

Youn, H. et al. (2016). On the universal structure of human lexical semantics. *PNAS* 113(7), 1766–1771.