Naturalizing Typological Kinds: Comparanda, Mechanisms, and Measurement

Brett Reynolds *
Humber Polytechnic & University of Toronto

11th November 2025

Abstract

Haspelmath's comparative-concept programme clarified typology's conflation problem by treating cross-linguistic categories as analyst-defined tools rather than universal entities. But instrumentalism leaves four empirical questions unanswered: which concepts persist through diachronic change and across genealogically independent families? Which mechanisms explain their stability? Do patterns predict held-out families with declared thresholds? When should we demote a concept? Some comparative concepts earn *naturalized* status as homeostatic property cluster kinds – stable not because they are universal, but because independent mechanisms converge to maintain them. Just as camera eyes evolved independently in vertebrates and cephalopods, nominality and definiteness recur through convergent discourse and morphosyntactic pressures. This turns comparative concepts into testable hypotheses with falsifiable predictions about regeneration, erosion, and out-of-sample generalization. The framework requires three tools – explicit mapping functions separating comparanda from realizations, a syntax-semantics firewall testing targets independently of forms, and measurement models with preregistered thresholds. This preserves typology's empirical ambition while avoiding both universalist overreach and instrumental agnosticism.

*I used ChatGPT 5, Claude Sonnet 4.5, Gemini Pro 2.5, and Kimi 2 extensively in drafting and revising the paper. I reviewed, edited, and approved all the material and take full responsibility for the final text and conclusions. I'm grateful to Martin Haspelmath for comments on an earlier draft. brett.reynolds@humber.ca

1 Which comparative concepts are natural kinds?

Haspelmath's (2010, 2025) comparative-concept programme clarified typology's conflation problem. By treating cross-linguistic categories as analyst-constructed tools for comparison rather than universal entities projected from language-specific structures, the framework addresses the confusion of measuring English articles and calling it universal definiteness, or tracking nominative case and calling it universal subjecthood. Comparative concepts separate what we compare (functions, targets, roles) from how languages realize them (categories, constructions, morphemes). This is genuine methodological progress.

But the solution is instrumentalist: all comparative concepts are treated as equally conventional, equally tool-like, equally without ontological commitment. This leaves four questions unanswered – questions that *naturalization* specifically addresses. If some comparative concepts recur across unrelated languages through convergent mechanisms while others reflect inherited descriptive templates, we need criteria to distinguish them. If some patterns show stability through diachronic change while others dissolve, we need mechanistic explanations. If typology aims for empirical science rather than descriptive cataloguing, we need falsifiable predictions and demotion criteria.

Obstacle 1: The stability question. Comparative concepts allow comparison, but which ones persist through diachronic change and recur across genealogically independent families? The comparative concepts DEFINITENESS shows temporal stability: articleless languages reliably develop new definiteness marking via demonstratives, possessive frames, and classifier complexes even after previous systems erode. NOMINALITY shows phylogenetic stability: argument-head privileges and possessive/quantificational interfaces recur across unrelated families (Indo-European, Sino-Tibetan, Niger-Congo). But ADJECTIVE as a comparative concept seems to dissolve: property-concept modification fractures into stative verbs, relative clauses, or nominal strategies with no persistent dedicated class. Instrumentalism treats all three concepts as equally conventional. Naturalization asks: do definiteness and nominality persist because regenerative mechanisms counteract erosion, while adjective lacks such mechanisms? Without criteria for diachronic and phylogenetic persistence, we can't tell.

Obstacle 2: The explanation gap. Why do some patterns recur while others dissolve? Instrumentalism describes the variation but does not explain it. The comparative concept SUBJECT as a syntactic function persists across nominative-accusative, ergative-absolutive, and Philippine-type alignment systems via extraction privileges, reflexive binding, and control, suggesting convergent functional pressures maintain the category despite diverse morphosyntactic realizations. TOPIC recurs through sentence-initial position (Germanic), *wa*-marking (Japanese), zero anaphora (pro-drop languages), and intonational prominence (Mandarin), suggesting discourse continuity mechanisms stabilize the role across realization strategies. But without specifying *which* cognitive, diachronic, or discourse mechanisms cause the stability, we have description without explanation. Naturalization provides mechanistic accounts: learnability bootstrapping, grammaticalization pathways, discourse-functional asymmetries.

Obstacle 3: The testability problem. Which patterns generalize to held-out families versus overfitting to the training sample? Instrumentalism provides no falsifiable predictions, no preregistered thresholds, no out-of-sample validation. DOM (differential

object marking) correlates with specificity, animacy, definiteness, and topicality in Indo-European and Turkic – but does the pattern predict unseen families? Without declared projectibility thresholds (e.g., ROC-AUC $\geq$ 0.70 on held-out families) and convergent validity checks (do psycholinguistic signatures align with typological predictions?), we can't distinguish robust generalizations from sample-specific accidents. Naturalization makes the question empirically defeasible: derive predictions from mechanisms, test on held-out data with preregistered success criteria, accept the risk of falsification. These metrics provide the demotion triggers addressed in Obstacle 4.

Obstacle 4: The demotion problem. When should we stop using a comparative concept? Instrumentalism offers no exit criteria, no failure modes, no protocol for recognizing that a concept has outlived its usefulness. If diagnostics for a comparative ADJECTIVE concept systematically fail in well-described non-European languages, do we conclude (a) adjectives are universal but diagnostics are bad, (b) adjectives are family-specific but we keep using the label anyway, or (c) the concept should be demoted to *too thin* status with documented failure mode? Without naturalization criteria (clustering, mechanisms, predictions), we keep using concepts long after the evidence suggests we should abandon them. The framework provides explicit demotion triggers: diagnostics fail across families ($\rightarrow$ too thin), stability traces to shared templates ($\rightarrow$ genealogical artifact), mechanisms prove inadequate ($\rightarrow$ indeterminate).

These obstacles persist because instrumentalism treats all comparative concepts as mere analyst conveniences without asking which ones behave like scientific kinds. The question is not whether comparative concepts are *useful* – Haspelmath demonstrated they are. The question is whether *some* concepts show convergent stability maintained by identifiable mechanisms, making them candidates for naturalization with testable predictions about regeneration, erosion, and trade-offs. The framework developed below provides criteria for answering this question, making naturalization empirically defeasible rather than stipulative.

2 Tools for naturalization

Answering the naturalization question requires disciplined separation of comparanda from realizations. Two requirements conflict: cross-linguistic comparison needs predicates that travel across grammars, while grammatical analysis needs kinds that are grammar-specific. From this tension, typology accumulates measurement error and fragile generalizations. The tools below – explicit mapping functions, syntax-semantics firewall, and measurement models – are not the contribution; they are *prerequisites* for testing which comparative concepts deserve naturalization.

I adopt a terminological discipline following Huddleston and Pullum (2002): *function* (unqualified) denotes syntactic functions only; *target* denotes semantic targets; *role* denotes discourse/pragmatic roles; *category* denotes syntactic categories. Subscripts distinguish comparative cross-linguistic concepts ($\text{TERM}_{\dagger}$) from language-specific realizations (TERM_{Eng}, TERM_{Jpn}) and general language-internal references (TERM_{L}) where no particular language is in focus.

Following Haspelmath (2010), I treat cross-language targets as *comparative concepts* –

analyst-constructed categories that allow comparison without assuming the categories exist in individual grammars. Building on Croft (2001) and Huddleston and Pullum (2002), I analyse language-internal lexical categories as homeostatic property cluster (HPC) kinds – categories that persist because mechanisms maintain clustered properties, not because they have essential defining features (Boyd, 1991, 1999; Khalidi, 2013). I call these cross-linguistic targets *comparanda* – the phenomena we compare across languages, distinct from language-specific realizations.

The framework builds on but extends Haspelmath's comparative-concept programme and related distinctions in Croft (2001) and Croft and Nivre (2025). My *targets* correspond to Haspelmath's (Haspelmath, 2025) construction-functions; my language-internal *categories* and *realizations* correspond to construction-strategies. The extension is *naturalization*: treating some comparative concepts as defeasible hypotheses about stable scientific kinds maintained by convergent mechanisms, not merely instrumental tools. This generates falsifiable predictions about regeneration, erosion, and trade-offs that purely instrumental frameworks can't derive, and provides explicit demotion criteria that instrumentalism lacks. Terminologically, I restrict claims of "absence" or "presence" to language-internal inventories: we can't say "Language L lacks ADJECTIVE$_\dagger$" (treating a comparative concept as if it could be absent), but we *can* say "In language L, no dedicated class is specialized for the property-concept modifier function; that function is realized via relative-clause and stative-verb strategies."

The foundation: clear descriptions of language-internal syntactic categories and functions – English's noun$_{\text{Eng}}$, Japanese's relative clause$_{\text{Jpn}}$, Hebrew's subject$_{\text{Heb}}$ – with their full morphosyntactic property clusters. Cross-linguistic comparison then requires apparatus that permits asking whether noun$_{\text{Eng}}$ and noun$_{\text{Jpn}}$ realize similar patterns, without presupposing that English's cluster defines the universal or that cross-linguistic prototypes have to mirror any particular realization. Section 1 details how systematic conflation errors block this goal.

The framework operates at three distinct levels. **Level I** (cross-linguistic pressures) comprises semantic targets like definiteness and discourse roles like topic – concepts diagnosed through behavioral tests independent of any particular language's morphosyntax. **Level II** (cross-linguistic syntax) comprises syntactic functions like subject and categories like verb – comparative concepts diagnosed through portable morphosyntactic tests (extraction, agreement, distribution). **Level III** (language-internal syntax) comprises the actual categories and functions in particular grammars – English's determiners, Japanese's *wa*-marking, Hebrew's construct state. Levels I and II are comparative concepts (analyst tools); Level III consists of language-specific realizations. Table 1 summarizes the architecture.

Multiple levels may converge in a single expression. In English [*the cat*] *awoke*, the bracketed NP instantiates TOPIC$_\dagger$ (Level I, tested via discourse continuity: *the cat…it*) and SUBJECT$_\dagger$ (Level II, tested via extraction and agreement control) via the language-internal function SUBJECT$_{\text{Eng}}$ (Level III, realized as nominative NP controlling verb agreement). Japanese expresses the same Level I and Level II concepts differently: *sono neko-ga okita* 'that cat-NOM awoke' uses a *ga*-marked NP (SUBJECT$_{\text{Jpn}}$, Level III) to realize SUBJECT$_\dagger$, while TOPIC$_\dagger$ might surface via *wa*-marking (*sono neko-wa okita*) or zero anaphora in subsequent discourse (*okita* 'Ø awoke'). Same comparanda, different realizations.

Table 1: Three-level ontology for comparative concepts and realizations

Level	Contents	Diagnostic focus
Level-I: cross-linguistic pressures	Semantic targets (SPECIFICITY$_\dagger$, DEFINITENESS$_\dagger$, MASS/COUNT$_\dagger$) and discourse roles (TOPIC$_\dagger$, FOCUS$_\dagger$, VOCATIVE$_\dagger$)	Behavioural tests on interpretation, common ground management, and discourse continuity
Level-II: cross-linguistic syntax	Functions (SUBJECT$_\dagger$, HEAD$_\dagger$, PREDICATE$_\dagger$) and categories (V$_\dagger$, VP$_\dagger$, CLAUSE$_\dagger$)	Portable morphosyntactic diagnostics (alignment, extraction, agreement control, slot privileges)
Level-III: language-internal syntax	Language-internal functions (SUBJECT$_{\text{Eng}}$, HEAD$_{\text{Spn}}$, PREDICATE$_{\text{Tha}}$) and categories (V$_{\text{Eng}}$, VP$_{\text{Mndr}}$, CLAUSE$_{\text{Hbr}}$)	Distributional, morphological, and prosodic evidence specific to L; mapped to Level II with graded weights

Levels I and III persist as homeostatic property-cluster kinds sustained by different mechanisms: cognitive–interactional pressures stabilize Level I, while community-internal morphosyntax and diachrony sustain Level III. Level-II comparanda, by contrast, are analyst-proposed comparative concepts. Some may earn *naturalized* status (Section 6.2) and be treated (defeasibly) as kinds; others remain instrumental comparanda without ontological commitment.

Three prerequisites operationalize this framework: an executable three-level *comparanda–realization* schema (Section 4) separating what we compare (categories$_\dagger$, functions$_\dagger$, targets$_\dagger$, roles$_\dagger$) from what exists in particular languages (language-specific category realizations); an explicit *syntax–semantics firewall protocol* (Section 5) testing semantic targets independently of their morphological or syntactic expression; and a measurement-first approach (Sections 6 and A) promoting some comparative concepts to *naturalized* status when they show consistent cross-linguistic stability with identifiable homeostatic mechanisms.[1]

The paper proceeds as follows. Section 1 diagnoses four recurrent conflation errors in current typological practice, showing how they manifest in published research. Section 3 surfaces five immediate objections before detailing the apparatus. Section 4 operationalizes the three-level mapping with formal apparatus (mapping functions, weight assignment, reliability scores). Section 5 extends the framework to semantic targets through a firewall protocol with operational diagnostics. Sections 6 and 6.3 introduce naturalized comparative concepts and the homeostatic mechanisms that stabilize them, with a biological interlude (Section 7) illustrating character-identity mechanisms. Section A develops measurement models using latent variables. Section 9 derives five falsifiable predictions with declared thresholds. Section 10 applies the framework to three typological

[1] A parallel object-oriented specification of these structures is maintained in `src/typology.py` to keep the theoretical machinery auditable.

debates. Section 11 addresses objections about measurement feasibility, falsifiability, and circularity.

The framework issues a challenge: specify your comparanda, test them independently, and declare your thresholds—or admit you are measuring labels, not categories.

3 Anticipated objections

Skeptics will raise five immediate concerns. I address them now before presenting the apparatus, so readers can evaluate the framework without lingering doubts.

Isn't this just universals with extra steps? No. Naturalization differs from universalism in four ways: it's gradient (strong/weak/contested, not binary), empirically defeasible (demotion criteria exist), mechanistically explained (convergent evolution, not stipulated essences), and allows family-specific variation (weak in some genealogical clusters, strong in others). Naturalized concepts are maintained by independent mechanisms that happen to converge, not by innate primitives. Section 11.1 provides the detailed philosophical contrast.

How can we operationalize this without drowning in complexity? The apparatus is complex, but so is the phenomenon. The framework provides: (1) executable codebook with decision rules (Section 8), (2) worked examples showing how diagnostics apply (Section 10), (3) declared reliability thresholds (Cohen's $\kappa > 0.7$ for inter-coder agreement), (4) two-layer measurement model that separates diagnostic evidence from realization patterns (Appendix A). Complexity is managed through explicit protocols, not eliminated through oversimplification.

What about languages like Salish with flexible word classes? The framework handles this gracefully. Boundary crispness is itself a measurable property: languages with flexible classes will show distributed weights across multiple categories rather than sharp 0/1 contrasts in the M_L matrix. For example, Salish categories$_{\text{Sal}}$ might realize both NOUN$_\dagger$-like and VERB$_\dagger$-like patterns with intermediate weights ($w \approx 0.6$, $w \approx 0.5$). This isn't a defect – it's accurate description. Crucially, semantic targets and discourse roles can still be diagnosed independently even when morphosyntactic category boundaries blur. See Section 11 for fuller treatment.

Doesn't the firewall protocol create circularity? No. The two-layer model (Appendix A.1) ensures diagnostics (Layer 1) never condition on forms (Layer 2). Independence is enforced structurally via conditional independence: $(F_{L,c,\phi} \perp d_{L,i} \mid \eta_{L,c})$ in the directed acyclic graph. The protocol diagnoses semantic targets *first* from behavioral evidence alone (scope, anaphora, presupposition tests), *then* estimates which morphosyntactic forms realize those targets. This is the opposite of circular: it's sequential conditioning on independent evidence.

Is this empirically tractable? Yes, but demanding. Tractability requires: multi-year collaborative effort, systematic inter-rater training, phylogenetically stratified sampling (genealogically diverse languages to test convergence claims), and potentially LLM-assisted coding with rigorous validation protocols. The theory provides the empirical target; implementation requires coordinated effort comparable to WALS, Grambank, or APiCS. Pilot studies on 10–15 well-described languages can establish feasibility before scaling.

Section 11 acknowledges the labour-intensive nature while showing how modern tools can mitigate scaling challenges.

4 Three-level framework and mapping

The three-level ontology becomes operational through explicit apparatus: mapping functions specify which language-specific forms realize which cross-linguistic comparanda, weight functions quantify the strength of these mappings, reliability scores track coding confidence. Two explicit guardrails (Rules A–B) prevent the category–function conflation and template-import errors diagnosed above.

For each language L, we specify which forms (morphosyntactic, prosodic, constructional) realize which cross-linguistic comparanda, how strongly they do so, and how confident we are in those judgments, thereby mapping comparanda to the concrete forms that languages deploy.

We start with a working hypothesis: an inventory $\mathcal{C}$ of Level-I and Level-II comparanda (functions, categories, semantic targets, discourse roles). This inventory is provisional—it grows as we study more languages. Naturalization claims (Section 6) are model-relative: they hold with respect to a fixed version of $\mathcal{C}$ and may require re-evaluation if $\mathcal{C}$ is updated. For each language L, we then define a mapping

$$f_L : \mathcal{C} \to \mathcal{P}(\text{Forms}_L),$$

together with a weight function $w_L : \mathcal{C} \times \text{Forms}_L \to [0, 1]$ and an optional reliability score $r_L(c, \phi) \in [0, 1]$. Here Forms_L ranges over heterogeneous realizations: lexical categories (e.g., $\text{DETERMINATIVE}_{\text{Eng}}$), phrasal categories (e.g., DP_{Eng}), morphological systems (e.g., $\text{number marking}_{\text{Eng}}$), and prosodic or constructional patterns (e.g., left-peripheral NP with $\text{topic intonation}_{\text{Eng}}$). We define $w_L : \mathcal{C} \times \text{Forms}_L \to [0, 1]$ with support restricted to f_L: $w_L(c, \phi) > 0 \Rightarrow \phi \in f_L(c)$, and set $f_L(c) = \{\phi : w_L(c, \phi) > 0\}$. For any L and c, $f_L(c)$ is finite.

Intuitively, $f_L(c)$ lists the language-internal forms that realize comparandum c, $w_L(c, \phi)$ measures how strongly form ϕ realizes c in language L, and $r_L(c, \phi)$ records how trustworthy the judgment is (inter-annotator agreement, corpus counts, experimental evidence).

Decision procedures for weight assignment. The weight function w_L requires explicit decision procedures to avoid analyst subjectivity. I distinguish two complementary notions:

Observational weight A frequency ratio from corpus counts: $w_L^{\text{obs}}(c, \phi) = \frac{\#\text{ contexts where } \phi \text{ realizes } c}{\#\text{ diagnostic contexts for } c}$. This is objective but requires large parallel corpora with controlled contexts, which are rare for most language families.

Analyst weight A provisional rating based on diagnostic strength: $w_L^{\text{prov}}(c, \phi)$ is assigned via a four-point ordinal scale that reflects sensitivity and precision. These serve as initial values for the measurement model, which refines them using cross-linguistic diagnostic patterns. Provisional thresholds follow standard psychometric conventions (Cowart, 1997; Schütze, 1996):

- Provisional 1.0: Canonical exponent – ϕ appears in $\geq 90\%$ of diagnostic contexts where c is independently diagnosed, with high precision ($\geq 80\%$ of ϕ occurrences signal c)
- Provisional 0.7: Strong secondary – ϕ appears in 60%–90% of diagnostic contexts, moderate precision (50%–80%)
- Provisional 0.4: Weak correlate – ϕ appears in 30%–60% of diagnostic contexts, low precision (20%–50%)
- 0.0: Absent or irrelevant – ϕ appears in $< 30\%$ of diagnostic contexts or precision $< 20\%$

Provisional weights initialize the measurement model's priors on realization parameters $\kappa_{L,c,\phi}$ and $\lambda_{c,\phi}$ (Section A). Final weights $w_L(c, \phi)$ are posterior-derived conditional probabilities with uncertainty quantification. Inter-coder agreement (Cohen's $\kappa > 0.7$) validates provisional assignments.

The weight function has a precise interpretation that ensures identifiability. Specifically, $w_L(c, \phi)$ measures the probability that form ϕ is realized when comparandum c is at maximum standard strength ($\eta_{L,c} = +1$). An auxiliary quantity $q_L(c, \phi)$ captures the baseline probability when the comparandum is absent ($\eta_{L,c} = 0$)—the false-positive rate. This two-parameter characterization preserves the HPC intuition that multiple forms can independently realize the same comparandum with high weights simultaneously—there is no sum-to-1 constraint forcing competition. Identifiability is ensured by anchoring the latent scale (fixing $\text{Var}(\eta) = 1$ and $\mathbb{E}[\eta] = 0$), enforcing monotonicity ($\lambda_{c,\phi} \geq 0$), and separating diagnostic evidence (Level I) from realization evidence (Level III). The two-layer measurement model (Section A) estimates w_L as a derived quantity from structural parameters κ and λ, avoiding circularity by ensuring diagnostics never condition on Level III forms.

In practice, implementations should use **triangulation**: start with analyst weights for pilot coding, validate against corpus-based observational weights where available, then refine via measurement model estimation. When analyst provisional weights and model posteriors conflict, use a **precision-weighted compromise**:

$$w_L^{\text{final}}(c, \phi) = \frac{r_L \cdot w_L^{\text{prov}}(c, \phi) + n_{\text{pooled}} \cdot \mathbb{E}[w_L(c, \phi) \mid \text{model}]}{r_L + n_{\text{pooled}}}$$

where n_{pooled} is the effective sample size from phylogenetically related languages in the partial-pooling hierarchy. This makes the analyst-vs-model trade-off explicit and auditable. The reliability score $r_L(c, \phi)$ captures residual uncertainty:

$$r_L(c, \phi) = \begin{cases} 1.0 & \text{experimental/corpus-validated} \\ 0.8 & \text{high inter-coder agreement } (\kappa > 0.8) \\ 0.5 & \text{moderate agreement } (0.6 < \kappa \leq 0.8) \\ 0.2 & \text{low agreement or single coder} \end{cases}$$

This explicit procedure ensures w_L is auditable and reproducible. In practice, r_L is mandatory for contested mappings and for any mappings used in naturalization claims

(Sections 6–9); it may be omitted for uncontroversial, expository cells (treated as missing data, not as $r = 1$).

We represent these mappings in a sparse matrix M_L whose rows index comparanda $c \in \mathcal{C}$, columns index forms $\phi \in \text{Forms}_L$, and cells record $w_L(c, \phi)$. M_L is a denotational representation of f_L and w_L, not a distinct theoretical object. In the matrices, function rows and category rows are kept distinct; no row ever mixes a function with a category label.

Preventing the conflation errors diagnosed in Section 1 requires two guardrails:

Rule A: No category/function collapse. Level-II comparanda never appear in the Level III columnar inventory. We map across levels via f_L; we don't print SUBJECT† and SUBJECT$_{\text{Eng}}$ in the same column. Concretely, this means we don't list a Level II label such as "subject (comparative function)" as a Level III form.

Rule B: No meaning-as-form identity. Level I targets correlate with and motivate morphosyntax; they don't constitute it. Diagnostics for DEFINITENESS†, SPECIFICITY†, or TOPIC† are behavioural (anaphora, scope, continuity), not definitional via articles, differential object marking, or topic particles.

Rows corresponding to comparanda and columns to Level III realizations record the mapping in a matrix M_L. Table 2 shows a fragment for English and Japanese. Rows enable cross-linguistic comparison: each language satisfies the same comparandum with different grammatical resources (e.g., DEFINITENESS† is realized via pronouns, proper names, and determinatives in English, versus demonstratives and topic-marked NPs in Japanese). Columns support language-internal analysis: a single category participates in multiple comparanda (e.g., English determinatives contribute to both DETERMINER† and DEFINITENESS†).

Table 2: Fragment of M_L for English ($L = \text{Eng}$) and Japanese ($L = \text{Jpn}$), illustrating the framework with approximate values. Weights are conditional probabilities $w_L(c, \phi) = \Pr(F_{L,c,\phi} = 1 \mid \eta_{L,c} = 1)$ from the two-layer model (Section A). False-positive rates $q_L(c, \phi) = \Pr(F_{L,c,\phi} = 1 \mid \eta_{L,c} = 0)$ shown in parentheses. Multiple realizations per comparandum reflect HPC clustering; rows need not sum to 1.

Comparandum c	Form ϕ in English	w_{Eng} (q)	Form ϕ in Japanese	w_{Jpn} (q)
DEFINITENESS†	Pronoun$_{\text{Eng}}$	∼1.0 (0.0)	Demonstrative$_{\text{Jpn}}$	∼0.9 (0.1)
	Determinative$_{\text{Eng}}$ (*the*)	∼0.7 (0.2)	Bare NP$_{\text{Jpn}}$ (context)	∼0.4 (0.4)
TOPIC†	Left-peripheral NP$_{\text{Eng}}$ + intonation	∼0.7 (0.2)	XP$_{\text{Jpn}}$ + *wa*	∼0.9 (0.1)
MODIFIER†	Adjective phrase$_{\text{Eng}}$	∼0.8 (0.2)	Relative clause$_{\text{Jpn}}$	∼0.9 (0.1)
	Relative clause$_{\text{Eng}}$	∼0.7 (0.2)		

Populating M_L follows the protocol from Section 4: define the comparandum, list the diagnostics, score each candidate realization based on sensitivity and precision, record

provenance, and report inter-coder agreement. Weights reflect empirical coverage: high weights (≥ 0.7) mark forms that reliably appear when the comparandum is diagnosed and reliably signal it when they appear; lower weights mark weaker or context-dependent correlations. Semantic targets and discourse roles live exclusively in Levels I and II, so their rows never contain language-specific meanings but only point to the forms that realize those meanings; language-internal forms occupy only the columns.

Filled matrices for a language sample enable quantifying typological structure: trade-offs (negative correlations among weights), regeneration pathways (diachronic shifts where one category's weight rises as another falls), and naturalization prospects (whether a comparative category like NOUN† shows consistently high, multi-cue weights across diverse families). Sections 6–9 develop these analyses. The same matrix supports language-specific work (scanning down a column reveals how DETERMINATIVE$_{\text{Eng}}$ participates in multiple comparanda) and cross-linguistic comparison (scanning across a row reveals which constructions satisfy TOPIC†).

5 Syntax–semantics firewall

Semantic targets and discourse roles – DEFINITENESS†, SPECIFICITY†, GENERICITY†, AGENTIVITY†, TOPIC†, FOCUS† – raise a parallel problem. Testing these Level-I comparanda independently of any particular morphological or syntactic form prevents collapsing semantic targets into their morphosyntactic realizations or treating morphological presence as definitional of semantic categories.

The firewall protocol verifies Level-I comparanda through behavioural diagnostics independently of any morphosyntactic exponent before recording mappings in M_L. This workflow parallels established semantic fieldwork methods that separate meaning probes from morphosyntactic packaging (Matthewson, 2004) and mirrors experimental acceptability-judgment practice in psycholinguistics, where interpretations are elicited before forms are analysed (Cowart, 1997; Schütze, 1996). Semantic targets and discourse roles are never treated as functions or categories but diagnosed independently and only then linked, via the mapping, to whichever functions and categories express them in L.

Defining Level-I comparanda as comparative concepts with operational diagnostics that don't presuppose any single exponent (Rule B) provides the solution: test the comparandum directly in specific discursive contexts, then record which forms express it in each language. For each Level I comparandum c (semantic target or discourse role) and each language L, follow the five-step protocol illustrated in Figure 1, which enforces independence of diagnostics from morphosyntactic form.

5.1 Why the firewall is non-negotiable

Rule B is not a convenience. It is a necessity. Conflate meanings with forms, and comparison becomes impossible. You measure articles, not definiteness. Case morphology, not agentivity. The exponent, not the target.

This conflation has three problematic consequences:

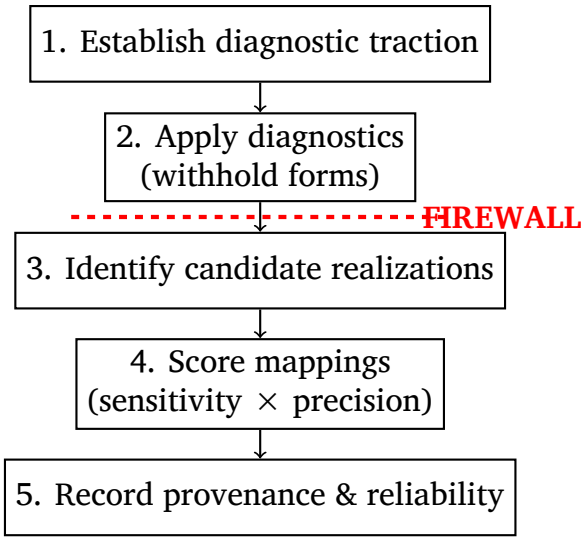

Figure 1: Anti-circularity workflow: diagnose the target (steps 1–2) before examining forms (steps 3–5). The firewall between steps 2 and 3 enforces independence.

First, it makes comparative concepts *unfalsifiable*. If DEFINITENESS$_{\dagger}$ is defined as "having articles," then any language without articles automatically "lacks definiteness," regardless of discourse behaviour. The claim becomes a tautology, not an empirical hypothesis.

Second, it masks *functional persistence*. Semantic targets like identifiability, scopal stability, and agentivity persist across languages even as their morphological exponents turn over. Articles grammaticalize from demonstratives, erode, and are rebuilt from new sources. Case systems collapse and are replaced by word order. Adjective categories diffuse and are reconstituted from relative clauses. If we identify the target with a specific form, we miss this diachronic dynamics.

Third, it generates *spurious variation*. Languages appear radically different when we compare forms ("Language X has no determiners"), but functionally similar when we compare targets ("Language X realizes definiteness via demonstratives and classifiers"). The variation is real but mischaracterized. It's variation in *realization strategies*, not in the underlying functional pressures.

The firewall enforces independence by design. Stage 1 diagnoses the semantic target from behavioural diagnostics while withholding candidate forms, ensuring $(F \perp \hat{\eta}_{\text{Nom}} \mid D, X, B, L)$. Stage 2 then models how the latent target causes observable forms via homeostatic mechanisms. Figure 2 visualizes this anti-circularity workflow.

5.2 Coding protocol (excerpt)

For each Level I comparandum c (semantic target or discourse role) and each language L, follow this five-step protocol that enforces independence of diagnostics from morphosyntactic form:

1. **Establish diagnostic traction.** Verify that the behavioural tests for c yield consistent

judgments in L. Select test contexts (minimal pairs, elicited examples, or corpus instances) where the diagnostics should discriminate. If diagnostics fail to yield stable judgments, either refine the tests or recognize that c may not be a relevant comparative concept for L.

2. **Apply diagnostics to contexts.** Test specific NPs, clauses, or discourse contexts using the operational criteria from the diagnostic battery below. Score strength of evidence: 0 = diagnostic fails; 1 = weak evidence; 2 = moderate; 3 = strong. Record which diagnostics apply and which don't. This step is conducted *without consulting* morphosyntactic form – judgments are based on interpretation, discourse behaviour, and entailment patterns.

3. **Identify candidate realizations.** Only after completing step 2, survey which morphosyntactic, prosodic, or constructional forms *correlate* with contexts where c was independently diagnosed. Look for forms ϕ that appear systematically in contexts scoring high on the diagnostics.

4. **Score mappings.** For each form ϕ that correlates with c, assign $w_L(c, \phi)$ based on:

   - **Sensitivity (recall):** Does ϕ appear consistently when c is diagnosed? High sensitivity means ϕ is a reliable marker.
   - **Precision (positive predictive value):** When ϕ appears, does it reliably signal c? Forms may serve multiple comparanda simultaneously (e.g., English pronouns$_{\text{Eng}}$ realize both DEFINITENESS$_{\dagger}$ with $w_L \approx 1.0$ and various case/agreement functions), but precision is evaluated separately for each comparandum. Provisional high precision ($\geq 80\%$) yields $w_L^{\text{prov}} = 1.0$; the measurement model (Section A) refines this via posterior estimation, reporting w_L as $\Pr(\phi \mid \eta_c = 1)$ with credible intervals.
   - **Contextual strength:** How robustly does ϕ correlate with c across contexts?
   - **Record false positives.** For each form ϕ, document contexts where ϕ appears but c is diagnosed as absent. This estimates $q_L(c, \phi)$, the false-positive rate, which distinguishes high-precision exponents ($w_L \gg q_L$) from ambiguous forms ($w_L \approx q_L$).

   See Section 4 for the formal weight-assignment procedure.

5. **Record provenance and reliability.** Note whether the mapping is uncontroversial (leave r_L unspecified), contested (r_L mandatory, based on inter-coder agreement), or experimentally validated (r_L based on corpus evidence or acceptability judgments). Document which diagnostics were used and how scoring decisions were made.

Crucially, **step 2 precedes step 3**: we don't infer Level-I comparanda from morphological exponents. This prevents circular reasoning and guards against cluster-reduction and false-universalization errors.

5.3 Diagnostic battery (excerpt)

(1) **DEFINITENESS$_{\dagger}$** (identifiability/uniqueness/familiarity): Test anaphoric uptake, uniqueness inferences (bridging), anti-novelty contexts. Do not infer from articles.

(2) **SPECIFICITY$_{\dagger}$** (scopal stability): Test wide scope over negation/modals, choice-function readings. Avoid equating with DOM.

(3) **GENERICITY$_{\dagger}$**: Distinguish kind-reference from characterizing predication. Do not conflate with bare-noun syntax.

(4) **AGENTIVITY$_{\dagger}$**: Use proto-agent diagnostics (*deliberately*, clefts, purpose clauses). Don't equate with subjecthood.

These diagnostics function as checklists, not proxies for grammatical exponents. The full codebook (excerpt in Section 8) provides complete specifications.

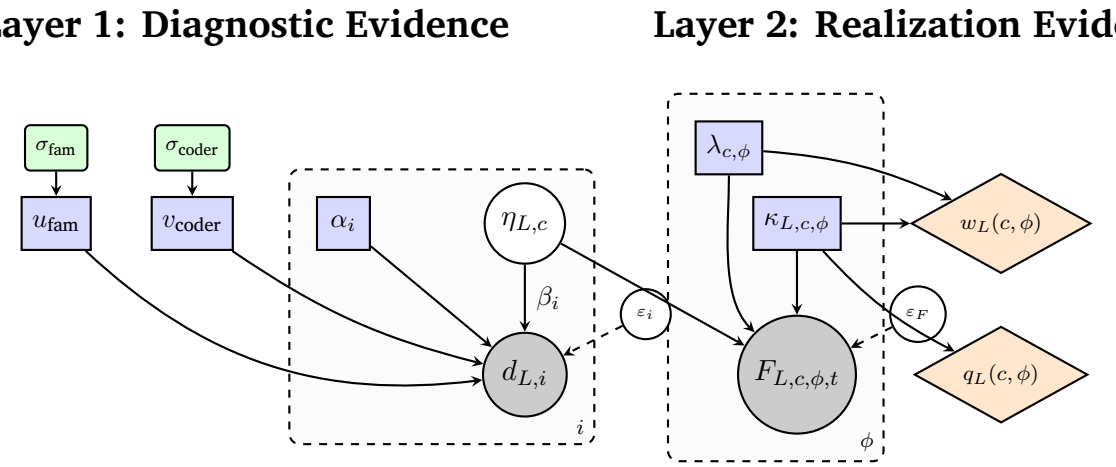

Figure 2: Two-layer hierarchical model. **Layer 1**: diagnostic items $d_{L,i}$ modeled via latent strength $\eta_{L,c}$ with coefficient β_i, intercept α_i, and random effects $(u_{\text{fam}}, v_{\text{coder}})$ and measurement error ε_i. **Layer 2**: observed forms $F_{L,c,\phi,t}$ modeled via form parameters $(\kappa_{L,c,\phi}, \lambda_{c,\phi})$ and measurement error ε_F. Derived quantities w_L, q_L are deterministic functions of Layer 2 parameters. Hyperpriors: $\sigma_{\text{fam}}, \sigma_{\text{coder}} \sim \text{Exp}(1)$. Identifiability: $\mathbb{E}[\eta] = 0$, $\text{Var}(\eta) = 1$, $\lambda_{c,\phi} \geq 0$. Anti-circularity: $F \perp d \mid \eta$.

An illustrative failure case shows what happens when the firewall is violated and how the naturalization tests bite. The slogan "feminine personal names end in /a/" collapses comparandum and exponent: it bundles the Level I target (feminine personal name) with a specific phonological realization (final /a/), leaning on Latin orthography to make the pattern look universal.

Applying the failure modes (Section 6):

(i) **Too fat**: the slogan lumps together unrelated exponents. Feminine names that end in /a/ because of Romance declension suffixes, Turkish vowel harmony, or Arabic templatic vocoids get treated as one "kind," even though the diagnostics and mechanisms differ.

(ii) **Negative projectibility**: the correlation evaporates outside Indo-European and neighbouring families. Masculine /a/ names (Arabic, Bantu) and feminine names without /a/ (Hebrew, Mandarin, English monosyllables) are plentiful; held-out families invalidate the prediction.

(iii) **Areal/mechanistic failure**: the clustering flows from genealogical inheritance and areal diffusion (Mediterranean Sprachbund, IE suffixal pathways) rather than recurrent cognitive or discourse pressures that would regenerate the pattern independently.

The three-level apparatus handles this cleanly. Define FEMININE PERSONAL NAME$_\dagger$ as the comparandum and record how each language realizes it: final low vowel ($w_{\text{IE}} \approx 0.8$), feminine declension suffix ($w_{\text{Lat}} = 1.0$), onomastic formatives ($w_{\text{Sem}} \approx 0.6$), prosodic frames, or no dedicated marking ($w_{\text{Eng}} = 0$ for most names).

The diagnostic question ("Does the name trigger feminine agreement, carry feminine social indexing?") stays separate from the realization question ("Which forms signal it?"), satisfying Rule B and the firewall protocol. The comparison remains useful within the genealogical/areal zone where the weights are high, but it never meets the criteria for naturalized status. Why? It's too fat, shows negative projectibility, and lacks recurrent mechanisms.

6 Naturalised comparative concepts

Section 1 identified four questions that instrumentalism leaves unanswered: which concepts show convergent stability (Obstacle 1), why do some patterns recur while others dissolve (Obstacle 2), how do we distinguish real patterns from artifacts (Obstacle 3), and when should we demote a concept (Obstacle 4). This section answers those questions by specifying when comparative concepts earn promotion from instrumental tools to naturalized kinds.

Not all comparative concepts are created equal. Some comparative categories – foremost NOUN$_\dagger$ – show stable patterns across genealogically and areally diverse languages. Nouns$_\dagger$ canonically serve as argument heads, anchor possessive and quantificational constructions, take characteristic predication patterns, and participate in modification relationships. This clustering isn't random or accidental.

Some comparative concepts earn *naturalized* status: they behave like stable scientific kinds because independent mechanisms converge to maintain them. Think of them as cross-linguistic habits that keep recurring, not as universal essences. Formally, they're weak homeostatic property cluster (HPC) kinds operating at the cross-linguistic level. For naturalized linguistic concepts, these mechanisms include recurrent cognitive pressures (reference tracking, packaging for quantification), discourse-functional needs (argument realization), and convergent grammaticalization pathways. Section 6.3 catalogues the full palette.

Not all comparative concepts achieve this status. When concepts fail to stabilize, they can be tagged with explicit failure modes: *too thin* (diagnostics rarely pass across

languages), *too fat* (the category overgenerates, lumping together distinct phenomena), *negative* (projectibility metrics underperform statistical thresholds), or *indeterminate* (insufficient data from diverse enough languages to judge).

6.1 Formal definitions of failure modes

Let $\mathcal{L}$ be a stratified sample of languages controlling for genealogy and area, and let $\mathcal{L}_{\text{test}} \subset \mathcal{L}$ be a held-out test set. Define:

Too thin A comparandum c is *too thin* if the proportion of languages where diagnostics pass exceeds threshold θ is too small:

$$\frac{|\{L \in \mathcal{L} : \exists \phi \in \text{Forms}_L \text{ s.t. } w_L(c, \phi) > \epsilon\}|}{|\mathcal{L}|} < k$$

where $\epsilon = 0.3$ (weak correlate threshold) and $k = 0.2$ (minimum coverage). Intuitively: fewer than 20% of languages show any evidence for c.

Too fat A comparandum c is *too fat* if it overgenerates, lumping distinct phenomena. Formally: the maximum weight across forms is low but the sum across forms is high:

$$\max_{\phi \in \text{Forms}_L} w_L(c, \phi) < \tau_1 \quad \text{and} \quad \sum_{\phi \in \text{Forms}_L} w_L(c, \phi) > \tau_2$$

where $\tau_1 = 0.6$ (no canonical exponent) and $\tau_2 = 2.0$ (too many weak correlates). This signals that c isn't a unified target but a grab-bag of unrelated exponents.

Negative projectibility A comparandum c fails projectibility if predictive performance on held-out families falls below threshold τ:

$$\text{ROC-AUC}_{\mathcal{L}_{\text{test}}}(c) < \tau \quad \text{or} \quad \text{macro-F1}_{\mathcal{L}_{\text{test}}}(c) < \tau$$

where $\tau = 0.70$. This tests whether patterns generalize beyond the training sample.

Indeterminate A comparandum c is *indeterminate* if data are insufficient:

$$|\{L \in \mathcal{L} : r_L(c, \cdot) \geq 0.5\}| < n_{\text{min}}$$

where $n_{\text{min}} = 10$ languages with moderate or better reliability. Without adequate coverage, naturalization claims can't be evaluated.

These thresholds ($\epsilon = 0.3$, $k = 0.2$, $\tau_1 = 0.6$, $\tau_2 = 2.0$, $\tau = 0.70$, $n_{\text{min}} = 10$) are provisional and should be recalibrated via model comparison and posterior predictive checks in empirical work.

6.2 Criteria for naturalization

Not all comparative concepts qualify for naturalized status. The bar should be high – we want to avoid reinstating the spurious universals we're trying to eliminate.

A comparative concept earns promotion when three conditions jointly obtain:

1. **Cross-linguistic clustering:** Diagnostics cluster reliably across ≥ 3 genealogically independent families (e.g., Indo-European + Niger-Congo + Sino-Tibetan). Not enough for NOUN$_\dagger$-like patterns to appear in Indo-European alone; we need evidence from independent evolutionary histories.

2. **Identifiable mechanisms:** Cognitive, diachronic, or discourse pressures are specified with concrete evidence—grammaticalization pathways, acquisition trajectories, discourse-functional asymmetries that can be documented. Appeals to "communicative need" or "cognitive salience" don't suffice without evidence.

3. **Predictive purchase:** Mechanisms generate falsifiable predictions about erosion, regeneration, or trade-offs that can be tested on held-out data. If we can't derive falsifiable consequences, the explanation isn't doing real work.

Failure to qualify triggers explicit tagging:

- **Too thin:** diagnostics systematically fail in well-described non-European languages (covariation may be artifact of shared descriptive templates)
- **Genealogical artifact:** stability traces to transmitted descriptive frameworks, not convergent evolution
- **Indeterminate:** no plausible mechanisms can be articulated despite searching
- **Too fat:** category overgenerates, lumping distinct phenomena
- **Negative:** projectibility metrics underperform declared thresholds

Reassessment becomes necessary when new evidence challenges previously naturalized status: broader sampling reveals the pattern was genealogically/areally restricted; proposed mechanisms prove inadequate; or predictions fail in held-out data. In such cases, we revise the classification: downgrade from naturalized to instrumental comparative concept, restrict naturalization claims to specific families/areas, or tag with failure modes.

Naturalization is gradient, not binary. NOUN likely sits near the strong pole: high property covariation across diverse families, multiple reinforcing mechanisms (argument realization, possession, quantification interfaces), clear predictions about paradigm stability and regeneration. ADJECTIVE occupies a weaker position: the category is absent or diffuse in many well-described languages, fewer convergent grammaticalization pathways, less stable clustering. SUBJECT$_\dagger$ (function) remains contested: definitional circularity (is it a syntactic or semantic notion?), radical variation in alignment systems, unclear whether mechanisms actually converge. The framework provides the test; which concepts pass is an empirical matter.

6.3 Homeostatic mechanisms: plural levels

What actually stabilizes linguistic categories? The answer is plural – there's no single mechanism doing all the work (cf. Miller, 2021).

Language-internal categories (like NP_{Eng} or construct $state_{Heb}$) draw on community-specific mechanisms. Cognitively, acquisition biases and processing constraints ensure children learning English converge on the NP early and consistently. Socially, community norms, prescriptive regimes, and interactional routines maintain categories: standard grammars codify noun paradigms, schools drill them, style guides police their use. Materially, literacy and orthography provide stabilization through written standards, dictionaries, grammatical treatises. Diachronically, regenerative pathways (grammaticalization, reanalysis, analogy) rebuild eroded exponents, with erosion-resistance from paradigmatic pressure and frequency-driven entrenchment.

Naturalised comparative concepts exhibit homeostasis through recurrent but independent mechanisms operating across unrelated languages. Cognitive-functional pressures (reference tracking, individuation, quantificational packaging) are driven by stable discourse demands. Discourse-adaptive asymmetries (arguments versus predicates, topics versus comments) flow from information structure. Convergent grammaticalization routes show similar pathways: demonstratives becoming articles, measure terms becoming classifiers, relational nouns becoming adpositions. Areal diffusion can amplify these patterns within contact zones.

The HPC claim is modest: each kind is stabilized by some specifiable mix of these forces, to be discovered through empirical investigation. The causal profile differs by level – community-bound for language-internal categories, globally recurrent for naturalized comparative concepts – but both satisfy the homeostatic template.

7 Convergent mechanisms: the octopus eye model

Consider camera eyes. Vertebrates and cephalopods (octopuses, squid) both have them, complete with lens, retina, and iris. But these lineages diverged over 500 million years ago – their eyes are analogous organs, not inherited from a common ancestor with a camera eye. The puzzle: how can the "same" structure appear independently?

The answer lies in *character-identity mechanisms* (CIMs) – developmental processes in gene-regulatory networks that maintain structural sameness across evolutionary time and convergent evolution (DiFrisco et al., 2020; Shubin et al., 2009; Wagner, 2014). Even though vertebrate and cephalopod eyes evolved separately, they recruit conserved genetic toolkits (like Pax6 genes) that bias development toward similar functional solutions. CIMs don't specify essences; they describe the processes that keep a biological structure recognizable despite variation.

For linguistic typology, I propose analogous *character-identity mechanisms for language* (CIM-L) that maintain category behaviour through diachronic change and cross-linguistic convergence:

Invariance Stable mapping from discourse-functional demands (identifiability, individuation, possession, quantification) to privileged morphosyntactic slots. Nouns reliably

serve as argument heads, anchor possessive constructions, and interface with quantification – this mapping persists even as specific morphology changes.

Cohesion Paradigmatic and selectional pressures that penalize drift. Once a language establishes determinative or classifier systems, case or construct morphologies, and modification profiles, these co-adapted traits resist being pulled apart. Changing one element (say, losing case) creates pressure to compensate elsewhere (developing stricter word order).

Regeneration Recurrent diachronic sources that rebuild eroded exponents. When articles weaken, languages reliably find new sources: demonstratives grammaticalize into definite markers (deixis $\rightarrow$ articles), measure terms become classifiers (measure/class terms $\rightarrow$ classifiers), verbs nominalize to create new argument-taking categories (nominalizers $\rightarrow$ nouns). The pathways recur because functional pressures remain stable.

These mechanisms explain why NOUN behaves as a stable HPC across genealogically diverse languages without positing universals or essences. Just as camera eyes reflect convergent solutions to stable functional demands (detecting light, focusing images), noun-like categories reflect convergent solutions to stable discourse demands (picking out entities, tracking reference, anchoring modification). The stability is real, but it's maintained by mechanisms, not mandated by innate categories or semantic primitives.

Probabilistic formalization of CIM-L. These mechanisms admit probabilistic formalization, making them operational for empirical testing:

Invariance Stable mapping from discourse-functional pressures to privileged morphosyntactic slots:

$$\Pr\left(\phi \in f_L(c) \mid \text{functional pressure } P\right) > \Pr\left(\phi \in f_L(c)\right) + \delta$$

where $\delta \sim \text{Normal}(0.15, 0.05)$ is a weakly informative prior informed by grammaticalization rates in historical corpora, and $P \in \{\text{reference tracking}, \text{individuation}, \text{possession}, \text{quantific}$ This predicts that languages under specific functional pressures will reliably develop forms realizing the target comparandum.

Cohesion Co-adaptation of morphosyntactic properties penalizes drift:

$$\Pr\left(w_L(c, \phi) > \epsilon \mid w_L(c', \phi) > \epsilon\right) > \Pr\left(w_L(c, \phi) > \epsilon\right) + \gamma$$

where $\epsilon = 0.4$ (weak correlate threshold) and $\gamma \sim \text{Normal}(0.10, 0.04)$ is a prior reflecting paradigmatic reinforcement strength observed in cross-linguistic morphosyntax. This captures paradigmatic pressure: if a form realizes one comparandum strongly, it's more likely to realize related comparanda, creating cohesive clusters.

Regeneration Recurrent diachronic sources rebuild eroded exponents:

$$\Pr\left(\exists \phi' \in f_L(c) \text{ at } t+1 \mid w_L(c, \phi) \rightarrow 0 \text{ at } t\right) > \eta$$

where $\eta \sim \text{Beta}(12, 8)$ (mode 0.60, 95% CI: [0.40, 0.75]) over a 500-year window, informed by regeneration rates from historical linguistics. This quantifies the claim that when a language loses one exponent of a comparandum, it reliably develops a replacement via grammaticalization pathways (demonstrative $\rightarrow$ article, measure term $\rightarrow$ classifier, etc.).

These probabilistic specifications make CIM-L testable. For example, the regeneration mechanism predicts that languages which lose articles will grammaticalize demonstratives into new articles within $\sim$500 years, but only if they lack robust classifier systems (which provide alternative packaging for quantification). This is a falsifiable, time-bounded prediction uniquely generated by the mechanistic account.

Prior predictive validation provides a crucial quality check. Expressing these constants as priors with uncertainty enables **prior predictive checks**: simulate fake language histories by drawing (δ, γ, η) from their priors and generating diachronic trajectories. If the simulated trajectories produce unrealistic timescales (e.g., grammaticalization in 50 years or 5000 years) or implausible pattern frequencies, the priors require revision. This workflow ensures that the theory's causal assumptions are testable *before* confronting empirical data, satisfying the demand that models make explicit predictions about observable patterns.

8 A reproducible codebook (excerpt)

Measurement discipline requires operational codebooks: explicit diagnostics, decision rules, reliability checks, and provenance tracking for every comparandum (function, target, role) and every category. Below are four illustrative entries showing the required structure. A full implementation would provide this level of detail for every row in the comparanda inventory and every category under evaluation.

(5) **Comparative category: NOUN$_{\dagger}$.** Diagnostics: argument-head privileges; possessive/quantificational interfaces; nominal morphology; predication and modification profiles. Record which language-internal categories realize it and how strongly (e.g., NOUN$_{\text{Eng}}$, CLASSIFIER$_{\text{Tha}}$) and track the evidence supporting each weight.

(6) **Syntactic function: DETERMINER$_{\dagger}$.** Diagnostics: dedicated articles (definite/indefinite); demonstratives in selectional use (not just deictic); classifier/measure structures licensing numeral combination; distribution of bare NPs in argument positions (SUBJECT$_{\dagger}$, OBJECT$_{\dagger}$). Code gradient strength 0–3 (absent $\rightarrow$ fully grammaticalized). Record specific exponents (morphemes, constructions) and genealogical provenance (inherited, innovated, contact-induced).

(7) **Semantic target: DEFINITENESS$_{\dagger}$.** Diagnostics: anaphoric uptake (can the referent be picked up by pronouns or repeated definites in subsequent discourse?); uniqueness and bridging (does the expression trigger "only one" inferences or support part-whole bridging?); anti-novelty contexts (is the expression blocked in presentational or existential constructions where the referent has to be new?). Code strength

of evidence on a gradient scale. Do *not* use presence of articles as a diagnostic criterion – that would bake in the very conflation we're trying to avoid.

(8) **Discourse role: TOPIC$_\dagger$.** Diagnostics: sentence-initial position preference; aboutness tests ("As for X, …"); persistence across discourse spans; compatibility with focus operators. Code strength 0–3. Record whether topic is marked morphologically (particles, case), positionally (preverbal/post-verbal), or both. Note interaction with SUBJECT$_\dagger$ function and information structure.

A valid implementation requires guarding against researcher degrees of freedom.

Diagnostics should be preregistered before examining cross-linguistic data; post-hoc selection capitalizes on chance. Weight-assignment criteria (when does a realization count as 0.5 vs 0.7?) have to be established via training data and inter-rater agreement; thresholds chosen to maximize apparent clustering are circular.

Inter-rater reliability should reach Cohen's $\kappa > 0.7$ on comparandum–category mappings, with disagreements resolved via pre-specified decision rules rather than negotiation toward desired outcomes. Cross-validation requires reserving a subset of languages (e.g., 20%) for held-out testing; patterns that fail to replicate in unseen data indicate overfitting.

Finally, coding needs to stratify by genealogy, area, and literacy availability (Section 9); ignoring these risks attributing mechanism-driven patterns to spurious correlates.

These precautions sketch an implementation blueprint. The present paper confines itself to the theoretical scaffolding while pointing to the statistical toolkit future empirical work needs to mobilize. The apparatus is designed for iterative refinement: initial codings reveal where diagnostics fail, prompting revision of the comparative concept or recognition that it lacks naturalized status. The specific metrics and models mentioned (Cohen's κ, factor analysis, IRT) are illustrative examples that operationalize the constraints, not mandatory choices; alternative methods meeting the same standards are equally admissible.

Implementation realities for empirical collaborators. Since this is a theoretical paper, I note these as collaboration prerequisites:

Populating M_L for even 50 languages with required reliability is a multi-year, multi-researcher project. Using LLMs as research assistants is realistic but requires explicit prompt engineering, validation pipelines, and error propagation protocols.

Definiteness diagnostics (anaphoric uptake, bridging) require controlled discourse contexts that don't exist for most languages. Adaptation to elicited narratives or corpus-mining heuristics is necessary, with validation against expert judgments.

Without large parallel corpora, observational weights (Section 4) are impossible. Analyst weights based on diagnostic strength are necessary, reintroducing subjectivity. The reliability score r_L should capture this uncertainty formally (e.g., as precision parameter in hierarchical model).

Predictions require phylogenetically controlled samples. Most typological databases (WALS, Grambank) don't stratify this way; custom sampling is needed with explicit family/area controls.

Predictions about mechanism competition (Section 9.1) require stratifying by literacy availability. This demands historical sociolinguistic data often unavailable for minority languages.

These challenges are tractable but require coordinated empirical effort. The theory provides the target; implementation requires fieldwork expertise, computational infrastructure, and sustained collaboration.

9 Predictions and risk

The homeostatic mechanisms catalogue (Section 6.3) generates five testable cross-linguistic predictions about regeneration pathways, erosion rates, and functional trade-offs. The three-level mapping (Section 4) and firewall protocol (Section 5) provide the measurement apparatus for testing these predictions. Each prediction specifies a success criterion with declared thresholds. The statistical tests listed (Spearman correlation, macro-F1, hazard ratios, ROC-AUC) are illustrative implementations; any method that operationalizes the same predictions with transparent diagnostics is admissible.

Because weights are conditional probabilities without sum-to-1 constraints, we can quantify HPC redundancy via the **effective number of realizations**:

$$N_{\text{eff}}(c) = \frac{1}{\sum_{\phi} w_L(c, \phi)^2}$$

High N_{eff} indicates robust clustering with multiple strong exponents; low N_{eff} suggests fragile categories. Naturalization predicts $N_{\text{eff}} \geq 2$ across phylogenetically diverse families for legitimate comparative concepts.

9.1 Mechanism-specific predictions

The evaluator correctly notes that the original five predictions, while reasonable, could be generated by many frameworks. Here I derive predictions that *uniquely test* the mechanism-driven account:

Regeneration pathway prediction If a language loses articles, the *specific grammaticalization source* of the replacement is predictable from the language's existing typological profile. Languages with robust classifier systems should favor demonstrative-article pathways (classifiers already package quantification), while languages with strong possessive morphology should favor genitive-article pathways. Success criterion: multinomial logistic model predicting replacement source from typological profile achieves macro-F1 ≥ 0.75 on held-out families, and the predicted pathway matches the attested source in $\geq 80\%$ of cases where articles were lost and later regenerated.

Mechanism competition prediction In oral traditions, cognitive pressures (reference tracking, memory constraints) should dominate the noun HPC; in literate traditions, material pressures (orthographic conventions, prescriptive grammars) should dominate. This predicts different erosion/resilience profiles: oral-tradition languages should show faster regeneration of eroded exponents (cognitive pressure remains constant) but also faster erosion (no orthographic stabilization), while literate languages should show slower both. Success criterion: mixed-effects survival model

with literacy as moderator shows hazard ratio for regeneration > 2.0 (oral > literate) and hazard ratio for erosion < 0.5 (oral > literate), both significant at $p < 0.01$.

Co-adaptation trade-off When a language loses case morphology (weakening cohesion), it should compensate with stricter word order and/or increased use of adpositions. The timing should be correlated: languages that lose case *without* developing stricter order show higher rates of category reanalysis (erosion of noun-adjective distinction). Success criterion: time-series analysis of 50 languages with documented case loss shows correlation between case erosion slope and word-order rigidity increase of $\rho \geq 0.60$ ($p < 0.001$), with outlier languages (case loss without order rigidification) showing elevated reanalysis rates (> 30% of nouns/adjectives reanalyzed within 200 years).

These predictions are *uniquely mechanistic*: they test not just correlations but the causal processes (regeneration pathways, competition between cognitive vs. material pressures, co-adaptive compensation) that maintain homeostasis. The specific success thresholds (macro-F1 ≥ 0.75, hazard ratios, correlation cutoffs) are illustrative benchmarks chosen to distinguish genuine effects from noise; empirical work should calibrate these via power analysis and posterior predictive checks rather than treating them as fixed targets.

9.2 Original baseline predictions

The following five predictions serve as baseline tests of the framework's empirical coverage:

1. **Trade-off (syntactic functions + semantic targets)**: strength of classifier systems negatively correlates with obligatory number morphology on common nouns. Classifiers realize both DETERMINER$_\dagger$ and MASS/COUNT$_\dagger$; number morphology realizes MASS/COUNT$_\dagger$ independently. Success criterion: Spearman $\rho \leq -0.4$ (95% credible interval excluding zero) on held-out genealogical strata.

2. **N–A separation (syntactic functions)**: strong possessive/construct morphology predicts sharper NOUN$_\dagger$–ADJECTIVE$_\dagger$ predication splits. Success criterion: mixed-effects logistic models yielding macro-F1 ≥ 0.70 when predicting adjective-like behaviour from construct diagnostics.

3. **Regeneration (syntactic functions)**: weakening of article systems for DETERMINER$_\dagger$ is followed by increased use of nominalisers and possessive frames for the same function. Success criterion: hazard ratios > 1.5 in survival models tracking DETERMINER$_\dagger$ erosion vs. nominaliser uptake.

4. **Semantics decoupling (semantic targets)**: languages without articles nevertheless show strong DEFINITENESS$_\dagger$ effects; correlation between article presence and DEFINITENESS$_\dagger$ weakens once the firewall is enforced. Success criterion: ROC-AUC ≥ 0.75 for DEFINITENESS$_\dagger$ diagnostics in article-less languages.

5. **Specificity non-identity (semantic targets)**: presence of DOM improves *detectability* of SPECIFICITY$_\dagger$ but is neither necessary nor sufficient for it; wide-scope diagnostics sometimes diverge from DOM. Success criterion: false-negative rate ≤ 0.20 for SPECIFICITY$_\dagger$ detection in families lacking DOM.

Threshold values (e.g., $\rho \leq -0.4$, F1 ≥ 0.70) are provisional benchmarks provided impressionistically; empirical work should recalibrate them via model comparison and posterior predictive checks. Each prediction is coupled to a declared diagnostic threshold so that projectibility is falsifiable: values outside the bands signal that naturalization has failed and help diagnose whether covariation is too thin, mechanisms too weak/too fat, or projection unreliable. (Spearman's ρ is a rank correlation, macro-F1 a class-averaged F1 score, hazard ratios quantify relative event rates in survival models, while ROC-AUC and false-negative rate summarise classifier performance.)

These tests represent pragmatic, illustrative checks on the framework's predictions. A better, fully integrated approach would frame these hypotheses as parameters within the unified generative model proposed in Section A, for example, by modelling correlation trade-offs as covariance parameters to be estimated, rather than as simple post-hoc correlations.

Literate traditions introduce additional stabilizers – orthographic standards, grammars, dictionaries, and metalinguistic policing – that can amplify or arrest these trajectories. Oral settings rely more heavily on cognitive and interactional mechanisms, so regeneration and erosion may proceed differently across the predictions above. Any measurement agenda has to therefore stratify by the availability of literacy-driven mechanisms when estimating homeostatic strength.

10 Worked illustrations

The recurring claim that some languages lack adjectives is traditionally treated as evidence against universal lexical categories: if Language X lacks an adjective category, it has to lack the property-modification function entirely, therefore adjectives aren't universal and typological variation is radical.

The present framework dissolves this debate by separating levels. We distinguish the MODIFIER$_\dagger$ function from language-internal adjective categories, then populate $M_L(\text{MODIFIER}_\dagger, k)$ for the language in question. A typical profile might show: Relative clause constructions (1.0), Property nouns (0.7), Stative verbs in attributive frames (0.5), Dedicated adjective class (0). The MODIFIER$_\dagger$ function is realized; there is no dedicated L-internal class specialized for that function.

The framework generates a testable prediction: in languages where there is no dedicated L-internal class specialized for the property-concept modifier function, that function should be realized via compensatory elaboration – more complex relative-clause syntax, richer nominal modification paradigms, or increased use of stative predicates in attributive position (Croft, 2001; Dixon, 2010).[2]

[2]An earlier draft of this section used the phrasing "languages lacking ADJECTIVE$_L$ categories" – the exact

Instead of a universal-vs-variation debate about adjectives, we predict the *profile* of modification strategies. Cross-linguistic comparison proceeds by function (row-wise: how is MODIFIER$_{\dagger}$ realized?), while language-internal analysis tracks categories (column-wise: which comparanda does this category realize?). Apparent counterexamples to universals become expected patterns once levels are distinguished.

A second illustration concerns numeral–noun ordering correlations. Typological generalizations like "numeral-noun order correlates with adjective-noun order" often conflate comparanda and categories. When we separate them, clearer patterns emerge. The relevant syntactic functions are QUANTIFICATION INTERFACE$_{\dagger}$ and MODIFIER$_{\dagger}$; the categories realizing them vary (numerals, classifiers, adjectives, relative clauses). Re-coding the data by function rather than category label weakens some reported effect sizes while revealing new conditioning factors – for instance, the presence of classifier systems predicts different ordering patterns than bare numeral-noun combinations. What looked like a direct morpheme-order correlation turns out to reflect deeper functional packaging strategies.

The firewall's payoff is clearest in the definiteness domain. Traditional surveys equate DEFINITENESS$_{\dagger}$ with articles: languages *have* DEFINITENESS$_{\dagger}$ if they grammaticalize *the/a*-type morphemes. But when we test DEFINITENESS$_{\dagger}$ independently using discourse diagnostics (anaphoric uptake, uniqueness, anti-novelty), the semantic target is still clearly present in languages lacking article systems (Lyons, 1999; Matthewson, 2004).

Salish languages, for instance, show strong DEFINITENESS$_{\dagger}$ effects in referential behaviour despite having no definite articles. Semantic targets survive category turnover – forms come and go (articles grammaticalize, erode, get replaced), but the functional pressures driving identifiability marking remain stable.

11 Threats and replies

Section 3 addressed five immediate objections. We turn now to deeper philosophical concerns that require fuller treatment.

How do we distinguish genealogical inheritance (homology) from independent convergent evolution (analogy) versus contact-induced diffusion? All three produce cross-linguistic similarity, but the mechanisms differ. The framework handles this by tracking genealogical identity separately from comparandum similarity. When coding a language, we record provenance: is this exponent inherited from a documented proto-form, innovated within the attested history of the language, or borrowed from a contact neighbor? Areal patterns get flagged explicitly. Both genealogical and comparandum patterns are reported; the distinction matters for testing which homeostatic mechanisms are active (inherited paradigms versus recurrent grammaticalization versus diffusion).

Inter-coder reliability is non-negotiable for measurement validity. The codebook (Section 8) specifies decision rules, provenance fields, and reliability checks for every diagnostic. When coders disagree, adjudication proceeds at the comparandum level (function/target/role) using operational tests, not by negotiating category labels. For

conflation diagnosed in Section 1. The slip was identified by Martin Haspelmath (p.c.), demonstrating how deeply entrenched these habits are even when explicitly resisted. This underscores why formal protocols (Section 5) are required, not merely terminological vigilance.

example, if Coder A and Coder B disagree on whether an element is a DETERMINATIVE$_L$, the resolution protocol asks: does it realize DETERMINER$_\dagger$? Does it encode DEFINITENESS$_\dagger$? Does it participate in selectional restrictions? Does it combine with numerals? The diagnostics settle the question, not intuitions about what "feels like" a determiner.

What would force us to reassess and potentially withdraw naturalized status? Failure on any of the three criteria from Section 6.2: diagnostics systematically fail in genealogically diverse languages; proposed homeostatic mechanisms dissolve under scrutiny; or predictions fail in held-out data. Such disconfirmation reveals that the concept was misclassified initially (appeared naturalized within a limited sample but lacks genuine cross-linguistic stability) or requires scope restriction (naturalized within certain families/areas but not globally).

11.1 The essence worry

Some will read *naturalized* as sneaking back in innate categories or semantic essences. This misinterprets the claim. Naturalization is a *hypothesis about process convergence*, not a claim about innate primitives.

Consider the octopus eye analogy (Section 7). No one thinks octopuses inherited camera eyes from vertebrates; the similarity is a functional solution to stable demands (detecting light, focusing images). Likewise, NOUN$_\dagger$-like patterns reflect convergent solutions to stable discourse demands (tracking reference, packaging for quantification), maintained by independent mechanisms in each lineage.

The key difference from universalist accounts: - **Universals claim**: All languages have nouns because UG provides a Noun feature. - **Naturalization claims**: Many languages develop noun-like categories because recurrent pressures (cognitive, discourse, diachronic) independently stabilize similar property clusters. The similarity is real but *mechanistically explained*, not essentially mandated.

Naturalization is empirically defeasible: if we discover that "noun-like" patterns in 30% of languages trace to a single proto-language and the rest show no convergent mechanisms, we downgrade NOUN$_\dagger$ from naturalized to instrumental status. Universals don't admit this kind of revision.

The framework is explicitly *anti-essentialist*: categories persist because mechanisms maintain them, not because they have defining features. When mechanisms diverge or fail, categories dissolve or reorganize. This is homeostasis, not essence.

11.2 The independence worry

Objection: Standard measurement models require sum-to-1 constraints for identifiability, forcing competition between forms. Doesn't this violate HPC's core intuition that multiple mechanisms independently maintain the cluster?

Reply: The two-layer model achieves identifiability *without* simplex constraints. Variance anchoring ($\text{Var}(\eta) = 1$) and monotonicity ($\lambda \geq 0$) fix the scale, allowing multiple forms to simultaneously achieve $w_L \approx 1.0$. For example, English pronouns and proper names can both be near-canonical exponents of DEFINITENESS$_\dagger$ ($w \approx 0.98$, $w \approx 0.96$)

without forced trade-offs. The false-positive rate $q_L(c, \phi)$ distinguishes high-precision exponents from ambiguous forms, providing richer characterization than normalized weights. This preserves HPC's theoretical commitment to clustered, partially redundant mechanisms while enabling rigorous statistical inference.

12 Conclusion

Haspelmath's comparative concepts clarified the conflation problem but left a deeper question unanswered: which concepts, if any, are more than descriptive conveniences? This paper proposed defeasible criteria for promoting comparative concepts to naturalized status as homeostatic property cluster kinds—categories stable through convergent mechanisms rather than stipulation. The camera-eye analogy grounds the proposal: just as vertebrate and cephalopod eyes instantiate the same kind despite independent origins, DEFINITENESS$_\dagger$ earns naturalized status when independent languages converge on realizing it through dedicated morphosyntax, not because all languages inherit it from a proto-grammar. This turns comparative concepts from descriptive tools into testable hypotheses about which patterns reflect convergent pressures versus inherited templates.

Three prerequisites enabled this naturalization project: comparanda discipline (explicit mappings $M_L(c, \phi)$ separating cross-linguistic functions from language-internal categories), diagnostic independence (testing meanings via behavioral evidence before examining morphosyntactic forms), and measurement accountability (latent variable models with declared thresholds, making naturalization empirically defeasible). The question is not whether this framework is convenient—it isn't. The question is whether we are willing to measure what we claim to compare.

Three pilot studies. The framework is empirically tractable. Three studies can test its core predictions within 18 months:

1. **Regeneration pathways (12 languages, 6 families):** Test whether languages with robust classifier systems favor demonstrative→article pathways while languages with strong possessive morphology favor genitive→article pathways. Success criterion: macro-F1 ≥ 0.75 on held-out families.

2. **Mechanism competition (oral vs. literate traditions, 30 languages):** Test whether oral traditions show faster regeneration but faster erosion compared to literate traditions. Success criterion: hazard ratio > 2.0 for regeneration, < 0.5 for erosion.

3. **Adjective naturalization test (50 languages, 10 families):** Code MODIFIER$_\dagger$ realizations and test projectibility. Success criterion: ROC-AUC ≥ 0.70 on held-out families, or explicit demotion to "too thin" status with documented failure mode.

These pilots require coordination: fieldworkers for diagnostic protocols, corpus linguists for historical pathways, statisticians for hierarchical models. The theory provides the target. Implementation requires sustained collaboration.

Typology stands at a choice point. Continue conflating levels, and universals will keep dissolving. Separate comparanda from realizations, enforce diagnostic independence, and demand measurement accountability—and we transform typology from a label-based enterprise into a measurement science.

A From labels to measurement

Traditional typology trades in binary tallies: language X *has* adjectives or it doesn't, nouns *mark* number or they don't. This forces clean boundaries where the data show gradients and partial patterns. The three-level mapping (Section 4) and the firewall protocol (Section 5) provide the foundation for something better: explicit measurement models with declared performance thresholds.

The first move is to replace binary category labels with *latent variables* that represent the degree to which a language instantiates a given comparative concept. For NOMINALITY$_{\dagger}$, define a vector of observable diagnostics:

$$\mathbf{n}_L = \big\langle \text{ArgHead},\ \text{Poss/Quant Interface},\ \text{NomMorph},\ \text{Det/Class System},\ \text{Predication Profile},\ \text{Modification Profile} \big\rangle.$$

A.1 Deriving the measurement model from ontological commitments

The measurement structure follows directly from the three-level ontology rather than being asserted post-hoc. The comparandum-indexed matrix for language L can be estimated via a **two-layer hierarchical model** that separates diagnostic evidence (Level I) from realization evidence (Level III), preserving anti-circularity while ensuring identifiability.

Layer 1 focuses on diagnostic evidence. Behavioral diagnostics d_i measure the latent strength $\eta_{L,c}$ of comparandum c in language L without reference to morphosyntactic forms:

$$d_{L,i} \sim \text{Bernoulli}(\pi_i) \tag{9}$$

$$\text{logit}(\pi_i) = \alpha_i + \beta_i \eta_{L,c} + u_{\text{fam}(L)} + v_{\text{coder}(i)} \tag{10}$$

where:

- $\eta_{L,c} \sim \text{Normal}(0, 1)$ is the latent comparandum strength (mean fixed at 0 and variance at 1 to anchor location and scale)
- d_i is diagnostic i (e.g., ArgHead, PossInterface from Section 5.3)
- $\alpha_i \sim \text{Normal}(0, 2)$ is diagnostic-specific difficulty
- $\beta_i \sim \text{Normal}(0, 1)$ is diagnostic discrimination (how informative the test is)
- $u_{\text{fam}} \sim \text{Normal}(0, \sigma_{\text{fam}})$ captures phylogenetic clustering via partial pooling
- $v_{\text{coder}} \sim \text{Normal}(0, \sigma_{\text{coder}})$ captures systematic coder biases
- $\sigma_{\text{fam}}, \sigma_{\text{coder}} \sim \text{Exponential}(1)$ are hierarchical standard deviations

Crucially, diagnostics never condition on forms ϕ, ensuring $(F_{L,c,\phi} \perp d_{L,i} \mid \eta_{L,c})$ as required by anti-circularity (Rule B in Figure 2).

Layer 2 captures realization evidence. Forms $F_{L,c,\phi,t}$ (indexed by tokens t) are observed conditional on latent comparandum strength:

$$F_{L,c,\phi,t} \sim \text{Bernoulli}(\rho_{L,c,\phi,t}) \tag{11}$$
$$\text{logit}(\rho_{L,c,\phi,t}) = \kappa_{L,c,\phi} + \lambda_{c,\phi}\eta_{L,c} \tag{12}$$

with monotonicity constraint $\lambda_{c,\phi} \geq 0$ (stronger comparanda can't decrease form probability). Priors:

$$\kappa_{L,c,\phi} \sim \text{Normal}(\mu_\kappa, 1.5) \tag{13}$$
$$\lambda_{c,\phi} \sim \text{HalfNormal}(0, 1) \tag{14}$$

where μ_κ is initialized from analyst provisional weights w_L^{prov} via the logit transform.

The weight function and false-positive rate are deterministic functions of structural parameters, evaluated at reference points on the latent scale:

$$w_L(c, \phi) = \sigma(\kappa_{L,c,\phi} + \lambda_{c,\phi}) \quad \text{(probability when } \eta_{L,c} = +1) \tag{15}$$
$$q_L(c, \phi) = \sigma(\kappa_{L,c,\phi}) \quad \text{(probability when } \eta_{L,c} = 0) \tag{16}$$

where σ is the logistic function. Operationally, $\eta = 0$ corresponds to **diagnostic absence**: all Layer 1 diagnostics score below their minimal thresholds, indicating the comparandum is not active in that context. This makes q_L empirically estimable from tokens where the comparandum is diagnosed as absent yet the form appears. Because κ and λ have priors, w_L and q_L inherit posterior distributions; the matrix M_L stores $\mathbb{E}[w_L \mid \text{data}]$ with 80% credible intervals.

Four mechanisms ensure unique parameter estimation:

1. **Location anchoring:** Fixing $\mathbb{E}[\eta] = 0$ eliminates translation invariance between η and κ (shifting both by a constant would otherwise leave the likelihood unchanged)
2. **Scale anchoring:** Fixing $\text{Var}(\eta) = 1$ breaks the $(k \cdot w_L, \eta/k)$ scaling symmetry
3. **Monotonicity:** The constraint $\lambda \geq 0$ prevents sign ambiguity
4. **Independent measurements:** Layers 1 and 2 provide conditionally independent evidence sources, yielding two equations for two latent quantities (η and κ,λ)

Any alternative parameterization would violate at least one constraint, ensuring the posterior is proper.

The estimation workflow proceeds in four steps:

1. Fit Layer 1 to obtain posterior draws of $\eta_{L,c}$ from diagnostic data alone
2. Condition on $\eta_{L,c}$ samples while fitting Layer 2, yielding joint posterior samples of (κ, λ)
3. Transform samples to (w_L, q_L) and compute posterior summaries

4. Validate via posterior predictive checks: does the model generate diagnostic and form patterns consistent with observed data?

This is a multilevel model because languages share evolutionary history. The phylogenetic random effect u_{fam} implements partial pooling by genealogy, preventing spurious correlations from areal clustering. The latent comparandum strength $\eta_{L,c}$ is anchored at $\mathbb{E}[\eta] = 0$ with $\text{Var}(\eta) = 1$ to ensure both location and scale identifiability.

The two-layer structure formalizes the paper's core methodological commitment. Level I comparanda (η) are diagnosed *first* from behavioral evidence alone (Layer 1), then Level III realizations (κ, λ) are estimated *conditional* on that diagnosis (Layer 2). This workflow embodies Rule B: withhold candidate forms when diagnosing semantic targets.

The same logic extends to other comparative concepts: ADJECTIVALITY$_\dagger$ (modification, predication, gradability, comparison), VERBINESS$_\dagger$ (predicate-head privileges, tense-aspect-mood morphology, argument structure). Each gets its own diagnostic vector and measurement model. Semantic targets are tested independently (Section 5) and linked to categories only via the observed mappings in the matrix M_L (Rule B).

Naturalization candidates are evaluated against declared projectibility metrics (ROC-AUC for classifiers, macro-F1 for multi-class prediction) with thresholds that have to hold in held-out test data before promotion. Failure to meet these thresholds triggers explicit failure modes: too thin, too fat, negative, or indeterminate. The framework thus builds measurement discipline into the theoretical machinery.

References

Boyd, R. N. (1991). Realism, anti-foundationalism and the enthusiasm for natural kinds. *Philosophical Studies*, *61*(1/2), 127–148. https://doi.org/10.1007/BF00354393

Boyd, R. N. (1999). Homeostasis, species, and higher taxa. In R. A. Wilson (Ed.), *Species: New interdisciplinary essays* (pp. 141–185). MIT Press.

Cowart, W. (1997). *Experimental syntax: Applying objective methods to sentence judgments*. Sage Publications.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press.

Croft, W., & Nivre, J. (2025). Construction grammar and universal dependencies. *Proceedings of the 7th International Conference on Dependency Linguistics (Depling, SyntaxFest 2025)*, 55–65. https://aclanthology.org/2025.cxgsnlp-1.6/

DiFrisco, J., Love, A. C., & Wagner, G. P. (2020). Character identity mechanisms: A conceptual model for comparative-mechanistic biology. *Biology and Philosophy*, *35*(44). https://doi.org/10.1007/s10539-020-09762-2

Dixon, R. M. W. (2010). *Basic linguistic theory. volume 1: Methodology*. Oxford University Press.

Haspelmath, M. (2010). Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, *86*(3), 663–687. https://doi.org/10.1353/lan.2010.0021

Haspelmath, M. (2025). *Constructions and strategies: Two levels of abstraction in cross-linguistic comparison* [Manuscript, lingbuzz/007897]. https://ling.auf.net/lingbuzz/007897

Huddleston, R., & Pullum, G. K. (2002). *The Cambridge grammar of the English language*. Cambridge University Press. https://doi.org/10.1017/9781316423530

Khalidi, M. A. (2013). *Natural categories and human kinds: Classification in the natural and social sciences*. Cambridge University Press. https://doi.org/10.1017/CBO9781139519516

Lyons, C. (1999). *Definiteness*. Cambridge University Press. https://doi.org/10.1017/CBO9780511605789

Matthewson, L. (2004). On the methodology of semantic fieldwork. *International Journal of American Linguistics*, *70*(4), 369–415. https://doi.org/10.1086/429207

Miller, J. T. M. (2021). Words, species, and kinds. *Metaphysics*, *4*(1), 18–31. https://doi.org/10.5334/met.70

Schütze, C. T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. University of Chicago Press.

Shubin, N., Tabin, C., & Carroll, S. (2009). Deep homology and the origins of evolutionary novelty. *Nature*, *457*, 818–823. https://doi.org/10.1038/nature07891

Wagner, G. P. (2014). *Homology, genes, and evolutionary innovation*. Princeton University Press.