Language as a Stack of Homeostatic Property-Cluster Kinds: From Phonemes to Constructions

Brett Reynolds [ID]*

Humber Polytechnic & University of Toronto
brett.reynolds@humber.ca

January 15, 2026

Abstract

This paper develops two complementary criteria – projectibility and homeostasis – for deciding when linguistic categories warrant treatment as homeostatic property-cluster (HPC) kinds. PROJECTIBILITY is tested directly: does a category support reliable out-of-sample inference? HOMEOSTASIS is inferred: are identifiable mechanisms plausibly responsible for maintaining the cluster, and do the data show signatures consistent with those mechanisms? I apply these criteria to three structural levels. At the phoneme level I use PHOIBLE inventories ($n \approx 2{,}700$ languages) to show family-wise concentration of inventory sizes and a scaling relation for the front-rounded vowel /y/. At the lexical level I evaluate 16 high-drift adjectives plus 16 matched controls, testing whether distributional neighbourhoods persist through semantic change; only a subset pass per-lexeme thresholds. At the constructional level I test a battery of 10 constructions (including designed failure cases) across UD English corpora; *or even* serves as a worked example, while the battery carries the evidential burden. The contribution is methodological: concrete, reproducible tests that keep kind-claims local

and evidence-driven. Phoneme-level evidence is strongest; word-level evidence is selective; construction-level evidence is provisional pending larger samples. Where both criteria are satisfied, treating a category as an HPC is empirically warranted; where they aren't, a more local or descriptive account is preferable.

Introduction

Language presents a familiar problem for cognitive science: the categories we rely on to speak and understand are stable enough to underwrite reliable inference but flexible enough to drift and diversify. Consider how the phoneme /r/ varies across English dialects but remains recognizable, or how the word *awful* drifted from 'awe-inspiring' to 'terrible' while keeping its identity. An attractive middle ground for understanding such categories – originating in Boyd's account of homeostatic property-cluster (HPC) kinds – offers a way to reconcile stability with change (Boyd 1991, 1999)[1].

HOMEOSTATIC PROPERTY-CLUSTER kinds emerge when identifiable forces keep characteristic features bundled – not through essence, but through contingent regularities. The category holds together as a family of properties maintained by mechanisms (biophysical, developmental, social) strongly enough to support induction. Think of biological species: robins share characteristic features – size, coloration, song patterns, nesting behaviour – not because of an essential "robin-ness" but because developmental programs, ecological pressures, and reproductive isolation keep these traits correlated generation after generation. Boyd's insight is that many scientific and social categories work similarly: they're probabilistic clusters stabilized by mechanisms that keep enough of the cluster together for inductive use.

This article proposes a general, testable program: many linguistic categories – phonemes, lexemes, and language-internal constructions – are HPC kinds. The claim is operational, not merely analogical. I test the claim with two simple diagnostics that any proposed linguistic category has to pass.

By LINGUISTIC CATEGORY I mean a TYPE maintained by identifiable mechanisms within a population (e.g., the phoneme /r/ in English, a lexeme such as *dog*, or a named construction such as *or even*). For present purposes I treat superordinate labels (PHONEME, WORD) as

---

[1]For the first explicit HPC formulation (applied to moral terms), see Boyd (1988: §3.8). For an explicit allowance that social mechanisms can underwrite homeostasis, see Boyd (2000).

taxonomic umbrellas; whether they themselves qualify as HPC kinds is an open question requiring different evidence and diagnostics. Here I target language-internal types, where the tests bite. Each case specifies its projection unit (token→token; language→language) and its stabilizers. How long a kind persists, and how widely it extends, are empirical questions about whether stabilizing mechanisms maintain covariance.

First, PROJECTIBILITY means that tokens have to support reliable out-of-sample inference about form, meaning, or distribution. Second, HOMEOSTASIS means that the property cluster can be tied to candidate stabilizers whose predicted signatures are consistent with the observed patterns.

For words, Miller (2021) develops this stance at the level of particular lexemes – *dog*, *run*, *egregious* – rejecting essence-based individuation in favour of mechanism-indexed clusters that are historically delimited and population-relative. On this view, such lexemes earn kind status because mechanisms sustain covarying properties: pronunciation, orthography, meaning, distribution. Miller identifies both cognitive mechanisms (consistent retrieval from mental lexicons) and social mechanisms (community enforcement of usage norms), though he acknowledges the partition depends on theoretical commitments. The lexeme *dog*, for instance, maintains its identity not through a platonic essence but because spelling conventions, pronunciation norms, semantic associations, and syntactic patterns travel together, stabilized by frequency of use, register and genre licensing, literacy education, and community norms.

At the phonological level, recent work by Ekström et al. (2025) argues that phonemes are culturally maintained cognitive tools (Heyes 2018) anchored in articulatory and auditory constraints. Two quantitative signatures exemplify the sort of measurable "homeostasis" HPC requires: family-wise ridgelines of inventory sizes (showing most languages cluster roughly between 20 and 50 segments across unrelated language families; see Figure 2), and a scaling curve in which the probability that a language includes /y/ rises with vowel-inventory size, while /i/ remains common even in small systems.

Both patterns are presented with explicit methodological detail and tied to biophysical and efficiency pressures, drawing on data from PHOIBLE, an openly available database of segmental phoneme inventories for roughly 2,700 languages worldwide (Moran, McCloy & Wright 2019), and a compact checklist of tool criteria summarizes the stabilizers involved (Ekström et al. 2025: Fig. 1 p. 4, Fig. 2 p. 7, Table 1 p. 14). These are precisely the ingredients HPC seeks: projectible distributions and identifiable stabilizing mechanisms.

The first case study revisits the phoneme level using the PHOIBLE database introduced above, but focuses on observable patterns rather than re-deriving articulatory physiology. I examine two signatures: how inventory sizes cluster by language family (visualized as ridgelines), and how the probability of finding /y/ in a language increases with the size of its vowel inventory. These patterns survive robustness checks like sampling one language per subfamily. The HPC reading is straightforward: the tight clustering within families demonstrates projectibility (new languages behave like their relatives), while the known mechanisms – quantal stability, dispersion, developmental tuning, and community norms – provide the homeostatic foundation.

The second case study turns to words. Building on Miller (2021), I evaluate 16 high-drift adjectives (e.g., *nice*, *sick*, *gay*, *awful*) plus 16 POS- and frequency-matched controls from the HistWords COHA embeddings. The question is whether contextual patterns – what words appear near them – remain coherent enough that a model trained on earlier decades can still identify the word in later ones. This demonstrates projectibility empirically for a subset; the rest fail per-lexeme thresholds, as the framework predicts. The homeostasis comes from multiple forces: entrenchment through repeated use and norm-guided conventions ensure that even as meanings shift, spelling, pronunciation, and core distributional patterns stay bundled together at usable levels.

The third case study treats English constructions as language-internal HPC kinds. The flagship *or even* construction – a scalar-additive cousin of the better-studied *let alone* (Fillmore, Kay & O'Connor 1988) – serves as a worked example: its cue bundle (anchor string,

syntactic parallelism, scalar/contrast cues) transfers across corpora and degrades predictably under ablation. But a single construction is too easy to tailor the analysis to. Accordingly, I test a battery of 10 constructions spanning four cue regimes, including two designed *brakes* cases (pooled resultative, register-local *X much?*) to force the framework to say *no* when it should. The stabilizers that maintain these patterns include frequency of use, cue redundancy, and normative enforcement through editorial practices.

Equally important are the failure cases – categories that don't qualify as HPC kinds. Some proposals are too thin: one-off items (like nonce coinages such as *bromance* before it caught on), speech errors that blend words together, and children's overregularizations (*goed* for 'went') lack the stabilizers that would make them predictable beyond their immediate context. Others are too fat: when typologists group together all "resultative" constructions across languages, or all "ditransitive" patterns, they're pooling structures maintained by largely distinct mechanisms in each language – different morphosyntactic resources, cue reliabilities, and normative regimes (Haspelmath 2010). With many mechanisms local to each language, the cross-linguistic umbrella groups multiple language-internal kinds rather than forming a single HPC kind. Finally, negative or complement classes (e.g., "all ungrammatical strings") are defined by what they're not, rather than by shared properties held together by identifiable mechanisms. The discipline here is mechanism-first, in line with the word-kinds program (Miller 2021).

The framework yields predictions and disconfirmers. A perturbation prediction: weakening a stabilizer (e.g., lowering frequency or impoverishing input) will reduce cluster covariance before norms re-stabilize – a pattern testable under register shifts or in learner corpora. A scaling prediction: rarer, articulatorily marked segments that lack quantal robustness (e.g., /y/ in the vowel space, which requires precise front-rounding coordination) become more probable as system size grows; analogously, low-frequency constructional variants should be more prevalent in larger constructicons. Both predictions borrow directly from the phoneme results (Ekström et al. 2025: Fig. 2 p. 7) and extend them to higher levels.

In sum, properly operationalized HPC naturalizes linguistic ontology for cognitive science. It tells us when a category is the right sort of thing to underwrite inference, what keeps it stable enough, and how it can change. Each tier has its stabilizers; each stabilizer leaves signatures; each signature supports prediction. The result isn't that everything is an HPC, but that many linguistic categories pass a disciplined test – and some don't.

1 Framework and diagnostics

The introduction outlined how linguistic categories might qualify as homeostatic property-cluster kinds. This section operationalizes that claim into executable diagnostics with specific thresholds and falsifiable predictions.

Claims about kindhood depend on identifiable mechanisms within populations: English /r/ is stabilized by English articulatory norms and community transmission, not by universal phonetic laws. How long a kind persists, and how widely it extends, are empirical questions about whether stabilizing mechanisms maintain covariance. Cross-linguistic categories like "resultative" often pool constructions maintained by completely different mechanisms in each language – these are useful classifications but not single kinds. The framework discovers kinds empirically rather than declaring them universally.

Before detailing the criteria, a note on units of analysis. Each tier targets a different object:

- **Phonemes:** system-level regimes (inventory bands, markedness scaling) across the world's languages.

- **Words:** lexeme-level kinds within a single population–time slice (20th-century written American English).

- **Constructions:** construction-level kinds within English UD corpora (2010s–2020s web and news registers).

The claims that follow are scoped to these units; generalization beyond them requires separate evidence.

The two criteria – projectibility and homeostasis – require different evidence at each linguistic level:

Boyd treats projectibility as following from homeostatic maintenance: if mechanisms hold a cluster together, the cluster will support induction. I revise this relationship, treating them as independent criteria for operational clarity and falsifiability. This isn't mere operationalization – it's a methodological commitment. Crucially, the two criteria are not operationalized symmetrically. PROJECTIBILITY is *tested* directly via out-of-sample prediction (cross-validated AUC, held-out F1, cross-corpus transfer). HOMEOSTASIS is *inferred*: I name candidate stabilizers (articulatory constraints, frequency entrenchment, normative enforcement), derive their predicted signatures (scaling curves, perturbation sensitivity, cue covariance), and check whether the data are consistent with those predictions. The inference is defeasible – the signatures could arise from other sources – but it's grounded in mechanism–signature pairings rather than asserted by narrative alone.

The relationship between mechanism and projection is asymmetric in a specific way. Figure 1 shows the structure.
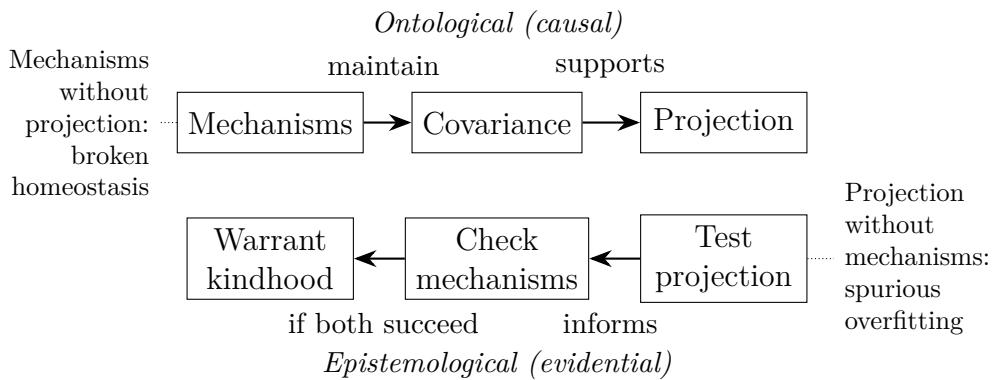

Figure 1: The causal arrow runs from mechanisms to projection; the warrant arrow runs the reverse. Both diagnostics have to succeed independently.

1. Projectibility: Can we predict new data from old? If phonemes form genuine kinds,

then knowing a language's family should constrain expectations about its inventory size. If words are kinds despite semantic drift, then patterns learned from earlier decades should help identify the same word later. If constructions are kinds, then cue patterns from one corpus should work in another. By contrast, nonce words or one-off blends would fail this test – they lack the regularities that enable prediction. This isn't philosophical hand-waving – it means concrete metrics like cross-validated prediction accuracy and held-out F1 scores.

2. Homeostasis: What keeps the patterns stable? The term "homeostasis" captures how mechanisms actively maintain property clusters despite perturbations, analogous to biological self-regulation. Projectibility is an *epistemic* success criterion (out-of-sample prediction); homeostasis is the *ontological* claim that mechanisms maintain the relevant property cluster. For each category, I name specific mechanisms and look for their signatures. Phonemes are stabilized by quantal regions that create articulatory "sweet spots" (Stevens 1989), dispersion that maintains distinctiveness (Liljencrants & Lindblom 1972), perceptual magnets that pull varied pronunciations toward prototypes (Kuhl et al. 1991), and community norms that transmit systems across generations. These mechanisms predict specific patterns: languages should converge on similar inventory sizes (not scatter randomly), and marked vowels like /y/ should appear mainly in larger systems (not randomly).

A category might project without identifiable mechanisms (overfitting to spurious patterns) or exhibit putative mechanisms without projecting (broken homeostasis). Both diagnostics have to succeed independently for kindhood to be warranted.

The protocol is straightforward: identify clustered properties, name stabilizers, test prediction, verify signatures. Boyd's HPC framework treats projectibility as supporting inductive inference: from known properties to novel properties, from observed instances to unobserved cases. The cross-validated prediction tests I employ – held-out accuracy, F1 scores, cross-corpus transfer – operationalize this criterion: if a category supports reliable

out-of-sample prediction, it licenses the kind of inductive generalization Boyd has in mind. Where both tests succeed, the category qualifies as an HPC kind over whatever period and extent mechanisms maintain covariance. Where they fail – no predictive power or no credible mechanisms – the category might be useful for description but doesn't constitute a kind in this technical sense. The failure modes are detailed in

S

refsec:failures: categories that are too thin, too fat, or merely negative complement classes.

Recent work provides models for what success looks like. Phonemes show remarkably consistent inventory sizes across language families and systematic scaling patterns for marked segments – exactly what we'd expect from the mechanisms described above (Ekström et al. 2025). Words maintain enough coherence through meaning change that their identifying features – spelling, pronunciation, distribution – stay bundled at usable levels (Miller 2021).

Negative categories don't qualify: "all ungrammatical strings" or "exceptions to rule R" are defined by what they lack, not by properties held together by mechanisms. The approach remains neutral about grammatical frameworks – it asks what patterns travel and why, not how to represent them theoretically.

2   Methods: tests, scope, and robustness

The ontological commitment is robust – these categories are natural kinds maintained by mechanisms – but the epistemology is disciplined: the tests have to be executable and falsifiable. This section outlines the conceptual approach and primary success criteria; complete statistical specifications appear in Appendix A.

2.1   Scope and populations

Kindhood claims depend on identifiable mechanisms within populations, with temporal and spatial extent determined empirically by covariance. Phoneme results apply to the PHOIBLE

2.0 snapshot as integrated with Glottolog/WALS for genealogy and area. Word results apply to contemporary written English in large, time-stamped corpora spanning the 20th–21st centuries. Construction results apply to Universal Dependencies (UD)-annotated English (news/web genres) in the 2010s–2020s. The program isn't a universal ontology of linguistic kinds; it's a method for deciding when a proposed category earns kindhood by passing predictive and mechanism-indexed tests.

2.2   Executable diagnostics

The two core diagnostics – projectibility and homeostasis – require different evidence at each linguistic level, but the logic remains constant: can patterns learned from one dataset predict another, and do identifiable mechanisms maintain those patterns?

Primary evaluation metrics include area under the receiver operating characteristic curve (ROC–AUC), area under the precision-recall curve (PR–AUC), F1 score (harmonic mean of precision and recall), and cross-validation (CV) performance. Uncertainty is quantified via bootstrap confidence intervals (CIs).

PROJECTIBILITY tests whether categories support out-of-sample prediction. For phonemes, I ask whether vowel inventory size (computed *excluding* the target segment to avoid circularity) predicts the presence of marked segments like /y/ – a front rounded vowel requiring precise articulatory coordination. Success requires a positive relationship (odds-ratio > 1 with 95% confidence), reasonable discrimination (10-fold cross-validated AUC ≥ 0.70), and monotonicity (verified by trend test at $\alpha = 0.01$). The common vowel /i/ serves as a control.

For words, I test whether distributional patterns persist through semantic drift. Using high-drift adjectives from HistWords COHA (Hamilton, Leskovec & Jurafsky 2016) (top decile of average cosine displacement among non-stop, non-proper adjectives, with documented drift terms forced in when available), I train a prototype classifier on early decades and test recognition in later ones. Success requires macro-F1 ≥ 0.35, substantial neighbour overlap (≥ 0.30), and temporally local errors (nearest-decade MAE $\leq$ one decade). Matched

controls with minimal drift and similar POS/frequency establish baselines.

For constructions, I test cross-corpus transfer using *or even.* A model trained on cues from one corpus (UD English GUM: Genre and Use of Meaning) must identify the construction in another (UD English EWT: English Web Treebank) with precision-recall area under curve (PR–AUC) ≥ 0.70.[2] The cue bundle – the anchor string, syntactic parallelism, and scalar/contrast cues – should degrade when components are removed, with the largest drops expected when parallelism is ablated.

These thresholds follow conventions in machine learning and corpus linguistics but remain somewhat arbitrary. What matters is declaring them in advance, calibrating them to distinguish structure from noise, and checking robustness (Appendix A). The discipline is prospective, not post-hoc.

HOMEOSTASIS requires identifying mechanisms and their signatures. Phonological systems are stabilized by articulatory attractors (quantal regions), dispersion pressure, and community transmission. These predict inventory sizes clustering between 20–50 segments across families (band test and bootstrap CIs defined in Appendix A; outputs in `out/ridgeline_band_metrics.cs` and marked segments scaling with system size. Words are stabilized by orthographic conventions, frequency-based entrenchment, and editorial norms, predicting the cohesion patterns above. Constructions rely on cue redundancy and normative enforcement, predicting cross-corpus stability and ablation sensitivity.

2.3   Managing non-independence

Languages inherit features from ancestors and borrow from neighbours. I address this through three approaches: modelling family relationships via fixed effects (dummy codes), adding geographical (macro-area) effects, and testing robustness through lineage pruning (one language per lower-level lineage when available, else one per Glottocode within family). Cross-validated AUC uses GroupKFold by family to prevent leakage. Key patterns have to

---

[2]PR–AUC captures how well the model separates true instances from false positives across different decision thresholds.

persist across all three specifications to count as genuine rather than genealogical artifacts.

2.4 Data and reproducibility

PHOIBLE provides cross-linguistic phoneme inventories; large timestamped corpora supply word distributions; UD treebanks offer syntactically parsed constructions. Preprocessing follows documented standards (detailed in Appendix A). All analyses are reproducible through archived code at https://github.com/BrettRey/Constructions-as-HPCs, with each figure mapped to its generating script. Primary outcomes are declared in advance: the /y/ scaling effect and its discrimination accuracy for phonemes; cohesion and classification performance for words; cross-corpus transfer and ablation effects for constructions.

3 Case A – Phonemes: a positive HPC

The phoneme tier is the cleanest place to test the claim that linguistic categories are HPC kinds. Inventories are comparable across languages, there's independent theory about plausible stabilizers, and open resources allow a fully reproducible analysis. I use PHOIBLE 2.0 (Moran, McCloy & Wright 2019) – introduced above as a cross-linguistic inventory database – to derive two complementary signatures: a family-wise concentration of total inventory sizes and a scaling relation linking the probability of /y/ to vowel-inventory size. On the projectibility side, the question is whether unseen languages behave like held-out members of their families or like random draws from a diffuse space. On the homeostasis side, the question is whether known mechanisms – articulatory–auditory attractors, dispersion, learning dynamics, and sociocultural norms – are a credible basis for the observed covariance (Stevens 1989, Liljencrants & Lindblom 1972, Lindblom 1990, Ekström et al. 2025).

For each language I count distinct consonant and vowel segments, excluding tones and prosodic units; when PHOIBLE lists multiple inventories for the same language (different sources may disagree on details), I keep the largest and document the choice. Families follow

Glottolog. Figure 2 plots kernel density ridgelines of total inventory sizes by family (families with fewer than ten languages are omitted; ordering is by family median). The picture is strikingly regular: medians cluster in a narrow band roughly between 20 and 50 segments, with thin tails beyond that range. Because densities are estimated independently by family, the shared band isn't an artifact of pooling; it's a cross-family regularity that enables inventory-level projection (from sampled languages in a family to unseen relatives within that family). The pattern matches typological summaries and aligns with the cognitive-tool review (Ekström et al. 2025: Fig. 1). (Band metrics in `out/ridgeline_band_metrics.csv`; Appendix A.)

The vowel /y/ – a front rounded vowel as in French *tu* – requires precise coordination of lip rounding with forward tongue position, making it articulatorily "marked" relative to cardinal vowels like /i/ that sit in quantal regions. Where the ridgelines demonstrated inventory-level projection, Figure 3 demonstrates segment-level projection: it fits a logistic model predicting /y/-presence from vowel-inventory size *excluding /y/* (and excluding /i/ for the control curve), with fixed effects for language family and macro-area, and 10-fold cross-validation grouped by family; a dashed comparison curve for /i/ serves as a control.

The /y/ curve rises monotonically with system size, whereas /i/ is common across the range and essentially flat. The slope for /y/ survives grouped CV by family; exclusion of small families ($n < 10$); and a permissive front-rounded coding ('y'/'ʏ') – see `out/y_model_metrics.csv` and `out/y_model_sensitivity.csv`. Discrimination is above chance (ROC–AUC ≈ 0.70; meaning the model correctly differentiates inventories with /y/ from those without in $\sim 70\%$ of cases), so the relation has predictive force rather than being a descriptive accident.

These two results satisfy the projectibility diagnostic in different ways: the ridgelines constrain where inventories fall by family, and the /y/ model supports a scaling inference about specific segments as systems expand. They also have a plausible HOMEOSTATIC basis. Quantal regions make some vowels (notably /i a u/) robust to articulatory imprecision

14

Figure 2: Phoneme inventory sizes by language family (PHOIBLE 2.0). Densities are shown for families with $n \geq 10$ languages, ordered by family median. Most families cluster in the 20–50 range (shaded), consistent with a homeostatic regime under articulatory–perceptual and dispersion constraints. Data: PHOIBLE 2.0; analysis code and exact processing steps are provided in the companion repository.

(Stevens 1989), dispersion spreads categories for distinctiveness (Liljencrants & Lindblom 1972, Lindblom 1990), and learning and control bind cues within categories (prototype attraction, audio–motor calibration), while local performance standards and literacy practices stabilize inventories across generations. The cognitive-tool synthesis documents these mechanisms and their empirical signatures (Ekström et al. 2025: Fig. 1, Fig. 2, Table 1). Read together, they explain why families share a stability band and why /y/ appears mainly in larger vowel systems.

Obvious worries are addressable. Counting conventions can inflate tails; excluding tones
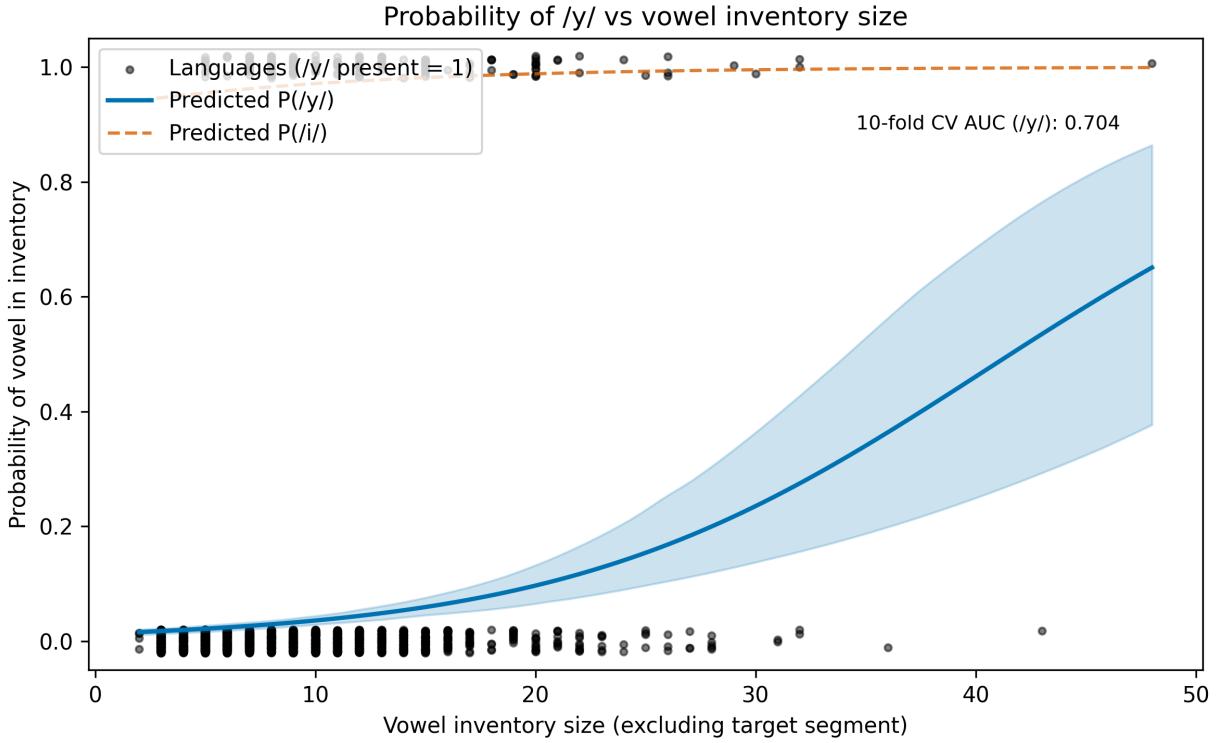
Figure 3: Presence of /y/ as a function of vowel-inventory size. Solid line: logistic fit with 95% confidence interval (CI) ribbons; dashed line: comparison curve for /i/. Points are languages jittered vertically at 0/1 for presence/absence. Model includes fixed effects for language family and macro-area; 10-fold cross-validated ROC–AUC ≈ 0.70. The increasing probability for /y/ with larger systems matches the scaling prediction; /i/ shows a high baseline with a weak slope.

and prosodic units eliminates that source. Genealogical non-independence can spuriously sharpen slopes; modelling family and area effects with grouped cross-validation leaves the inference intact. Orthographic noise can misclassify vowels; the repository records Unicode normalization and diacritic handling. None of these considerations undermines the main point: phoneme inventories exhibit projectible structure underwritten by concrete stabilizers, and so qualify as HPC kinds at the population–time scales analyzed here.

4   Case B – Words: mixed evidence under drift

If homeostatic property–cluster kinds are to do real work beyond phonology, they ought to earn their keep where categories are visibly historical. Words are the hard case and the natural next step. The aim here isn't to freeze a lexeme at a moment in time, but to show that a word can drift semantically while preserving enough covariation among its properties for inductive use. On the projectibility side, the question is whether held-out decades are predictable from earlier usage; on the homeostasis side, the question is whether there are plausible stabilizers that would make such predictability non-accidental. Miller's mechanism-first treatment of word-kinds sets the bar: kindhood, if it applies, has to be earned a posteriori by sustained covariation among orthography, phonology, meaning, and distribution in a particular population and time slice (Miller 2021).

I illustrate with drift-prone adjectives from the HistWords COHA lemma embeddings (Hamilton, Leskovec & Jurafsky 2016) and report a multi-lexeme evaluation in Appendix A. Targets are the top decile of average cosine displacement among non-stop, non-proper adjectives (with documented drift terms like *nice, sick, gay, awful* forced in when available) plus 16 POS- and frequency-matched controls. The operationalization is deliberately spartan. By "distributional neighbourhood" I mean the set of words that typically appear within a fixed window (here, five tokens) of the target word in large corpora. These neighbourhoods are represented as vectors in semantic space, where proximity captures co-occurrence patterns. Representations are decade-binned skip-gram negative sampling (SGNS) embeddings (window = 5, dim = 300) already aligned across decades; tokens are lemmatized and lower-cased. To check that results are not an alignment artifact, I also run an alignment-free variant that uses decade-internal neighbour sets and Jaccard similarity for prototype matching (Appendix A). No sense inventory is imposed; the question isn't which senses exist, but whether the family of properties that travel with the word remains coherent enough to support inference.

Two simple checks supply the answer. First, a cohesion check: some drift adjectives retain

organized neighbourhoods even as their centres move. For *nice*, nearest neighbours shift from {*pretty, lovely, pleasant*} in the 1900s to {*cute, wonderful, really*} by the 2000s; for *sick*, from {*ill, tired, hungry*} to {*hurt, drunk, upset*}. These trajectories meet the overlap threshold (top-50 overlap ≥ 0.30), but the broader high-drift set averages lower cohesion (mean overlap 0.20 vs. controls 0.64; Appendix A), consistent with the framework's expectation that not all drifted words remain kind-like. Distributional neighbourhoods serve as proxies for patterns of use. When sense distinctions matter, these distributional patterns can be cross-checked against explicit semantic criteria (e.g., dictionary senses, human sense annotation).

Second, a held-out prediction: a prototype classifier trained on early decades (1900–1940) and tested on later decades (1950–2000) recovers word identity well above a shuffled-label baseline (macro-F1 = 0.84 vs. 0.03; indicating that the model recovers word identity with high reliability, far exceeding random guessing). An alignment-free neighbour-set classifier still performs well (macro-F1 = 0.74 vs. 0.02) and yields mean target accuracy 0.58 (Appendix A). Mean per-lexeme accuracy is 0.71 for targets (95% CI [0.58, 0.83]) and 1.00 for controls. Temporal locality, measured by the nearest-decade mean absolute error, averages 1.71 decades for targets and 1.03 for controls; only *nice* and *sick* fall at or below the one-decade threshold (Appendix A). Combined with the cohesion cutoff, these are the only targets that satisfy both per-lexeme criteria in this run.

For the subset that passes, these signatures meet the projectibility diagnostic in the only sense that matters for an historical object: past usage fixes expectations that carry forward. The relevant test is whether the distributional neighbourhood resists dissolution when the mean location of a word's use shifts; that resistance is precisely what the HPC picture predicts when stabilizers preserve enough shared properties as a word moves. Where the diagnostics fail, the framework tells us to withhold kindhood for that lexeme–time slice rather than forcing a positive verdict. Those stabilizers are not mysterious.

Orthographic standardization constrains spelling through educational institutions and publishing practices, implemented cognitively via explicit instruction and error correction

that builds orthographic representations resistant to variation. Frequency-based entrenchment makes high-frequency items resistant to perturbation through sheer repetition in memory: each token strengthens the form-meaning link and automates retrieval (Bybee & Hopper 2001). Editorial norms are enforced through copy-editing workflows that flag nonstandard usage, implemented via conformity bias (speakers align with prestigious variants) and reputation monitoring (fear of correction motivates norm-following). Register licensing operates through genre conventions that sanction certain words in certain contexts, implemented via associative learning linking words to situational contexts. These are not mere labels for correlation – they are social practices with cognitive effects, operating through domain-general mechanisms like associative memory, conformity, and error-driven learning.

4.1   Objections and responses

There are obvious objections, and they can be separated from one another.

One is methodological: distributional neighbourhoods are proxies, not senses. That's correct, but it isn't a defect here. The claim under test is that the relevant family of properties stays bundled tightly enough for prediction; distributional stability is an appropriate read-out of that bundling, and the failure mode – a collapse in cohesion and in held-out performance – is clear.

A second objection distinguishes homeostasis from inertia: maybe distributional stability reflects momentum (frequency begets frequency) rather than active maintenance. This is testable. Homeostasis predicts that weakening a stabilizer (e.g., reducing editorial oversight in informal registers, removing spelling instruction) should degrade coherence before re-stabilizing at a new equilibrium, whereas inertia predicts monotonic decay. The contrast can be probed by comparing registers or by tracking divergence between prescriptive and colloquial usage, but the present data do not adjudicate it.

A third objection is genealogical: some drifts are punctuated, driven by contact or fashion, and so prediction should break. This is a fair disconfirmatory case; it's also consistent with

the framework. If a lexeme's covariance collapses or becomes unmoored from any credible stabilizer, we should withhold kindhood for that population–time slice rather than force an HPC verdict.

A fourth objection appeals to polysemy: if a family of uses fractionates, does the umbrella still count as one kind? Here again the framework is conservative. Population-relative equilibria are what matter. If distinct usage communities stabilize distinct covariances – two clusters with their own stabilizers – the correct description is two local kinds, not an analyst's disjunctive lump.

The positive case, then, is modest but informative. Some words can change and still remain HPC kinds because the mechanisms that matter for use – orthography and phonology, frequency and register, collocational habits and editorial standards – are strong enough to bind their properties across time. In this evaluation the overall classifier clears the macro-F1 threshold, but only a subset of targets (notably *nice* and *sick*) meet the per-lexeme cohesion and temporal-locality cutoffs; many high-drift adjectives fail at least one criterion. What the figures show isn't that a word keeps its older meaning, but that for some lexemes the distributional profile remains predictively structured as it moves. That's exactly the sense in which a linguistic category earns kindhood by HPC lights: it projects because it's homeostatically maintained. The contrast class is equally clear. One-off coinages that fail to diffuse, campaign-season blends, or transient vogue terms often show neither cohesion nor predictive grip when tracked across time; nothing stabilizes them. They are legitimate objects of study, but they aren't kinds in the relevant sense.

This tier, then, completes the bridge from phonemes to higher structure. In phonology, the stabilizers are largely biophysical and perceptual and yield stability bands and scaling relations. In the lexicon, the stabilizers are mainly sociocultural and distributional and yield cohesion under drift and recoverable neighbourhoods. The ontology is the same: a posteriori kinds stabilized by mechanisms we can name, whose signatures we can see.

5   Case C – Constructions: *or even* as a positive HPC

Constructions – conventionalized pairings of form and meaning that go beyond compositional rules – offer a different challenge for the HPC framework than single segments or words. Where phonemes cluster in articulatory space and words maintain distributional neighbourhoods, constructions rely on multiple converging cues that speakers recognize as a gestalt. The English *or even* construction provides a practical test case: it is a scalar-additive pattern closely related to *let alone*, but it is substantially more frequent in the UD corpora, enabling cross-corpus evaluation.

Consider the contrast in (1):

(1)   a.   *I can't afford coffee, or even dinner.*

Here the second item is presented as a stronger, more extreme, or less likely alternative on a contextually relevant scale. This scalar-additive relationship – where Y ranks higher than X on some contextually relevant scale – defines the construction's core meaning. But how do speakers recognize this pattern reliably? And what keeps its formal and semantic properties bundled together across different texts and registers?

I profile three observable cues that work together:

- **String anchor**: The phrase *or even* itself

- **Syntactic parallelism**: The contrasted elements (X and Y) typically match in grammatical category (both nouns, both verb phrases, etc.)

- **Scalar/contrast cues**: markers like negation or degree terms (*not, no, only, just, more, less*) and list punctuation (comma/semicolon) that frame Y as a stronger alternative

What is this construction cognitively? I treat *or even* as a stored form-meaning pairing: a schema specifying (a) the anchor string, (b) a syntactic template [X, *or even* Y] with parallelism expectations, and (c) scalar semantics (Y more extreme than X). The schema

abstracts over exemplar instances. This is compatible with usage-based Construction Grammar (Goldberg 1995, 2006), Sign-Based Construction Grammar (Boas & Sag 2012), HPSG (Pollard & Sag 1994), or any framework treating constructions as stored units rather than purely derived structures. The commitment is to storage + form-meaning pairing; the specific theory of how acquisition and processing implement this is an open question. For concreteness, I assume cue weights emerge from frequency and reliability (usage-based mechanisms), but constraint-based or parametric implementations could maintain the cluster via different stabilizers.

To test whether this cue bundle qualifies as homeostatic, I use two independently annotated corpora from the Universal Dependencies (UD) project: English GUM (Zeldes 2017) – containing diverse genres from academic writing to online reviews – and English EWT (Silveira et al. 2014) – built from web text and email. These corpora provide syntactic parses that identify grammatical categories and dependencies, enabling automatic extraction of the parallelism and licensing features beyond simple string matching.

The projectibility test asks: can patterns learned in one corpus predict instances in another? I train a minimal classifier on the three-cue bundle using data from GUM, then evaluate its ability to identify true *or even* constructions in the held-out EWT corpus (and vice versa). This cross-corpus design is crucial – if the construction were merely a frozen idiom or a corpus-specific quirk, the patterns wouldn't transfer. Table 1 reports the discrimination performance using PR–AUC. Because evaluation is restricted to anchor-present candidates, the anchor-only baseline is uninformative (PR–AUC equals prevalence), so gains reflect genuine cue structure.

The full three-cue model achieves PR–AUC in the 0.83–0.89 range in both transfer directions (indicating robust precision and recall; GUM→EWT means training on GUM, testing on EWT). This exceeds the prevalence baseline and indicates robust cross-corpus generalization. Removing parallelism produces the largest drops ($\Delta \approx$0.27–0.30; a substantial loss of predictive power), while removing licensing yields smaller but consistent losses ($\Delta \approx$0.06–

Table 1: Cross-corpus evaluation for *or even.* Full model uses anchor+parallelism+scalar cues; ablations drop one cue. Anchor-present candidates; class prevalence shown in the note.

Direction	Model	PR–AUC	Δ
GUM→EWT	Full bundle	0.886	–
	Drop parallelism	0.612	–0.274
	Drop licensing	0.822	–0.063
EWT→GUM	Full bundle	0.829	–
	Drop parallelism	0.527	–0.302
	Drop licensing	0.650	–0.179

Note: Metrics are computed on anchor-present candidate sets; full precision/recall/F1 values are available in supplementary materials.
Anchor-present counts: GUM $n = 12$ (pos=5), EWT $n = 18$ (pos=10). These results are illustrative; the confirmatory analysis expands to additional UD English corpora and a larger anchor sweep.

0.18). The string anchor alone isn't sufficient; the construction needs its supporting cast of cues.

Figure 4 decomposes the cue distributions in each corpus. Despite different genres and collection methods, both corpora show similar profiles: verbs and nouns dominate the *Y* position, and scalar/contrast cues appear at comparable rates. This cross-corpus stability – maintained without explicit coordination between the corpus creators – suggests genuine linguistic regularities rather than annotation artifacts.

What mechanisms maintain this stability? Three are plausible and testable:

1. **Frequency and entrenchment**: The construction appears often enough in the available corpora that speakers internalize its pattern through repeated exposure

2. **Cue redundancy**: Multiple signals converge – even if parallelism fails in a rushed email, the anchor and licensing context still signal the construction

3. **Normative pressure**: Editorial practices and style guides reinforce the canonical pattern, especially in formal registers; odd, unlicensed uses like *I bought coffee or even dinner* in neutral contexts would likely be corrected in editing

These mechanisms operate at different timescales but interact to co-stabilize the con-
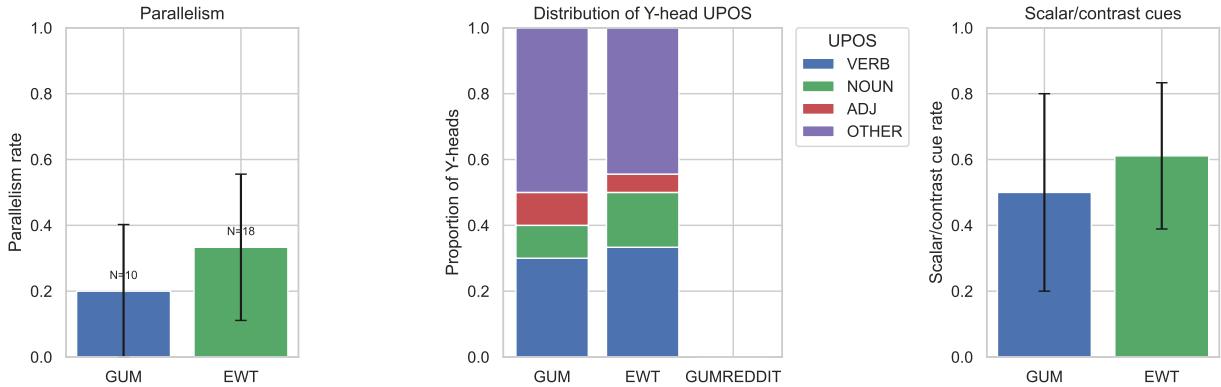
Figure 4: Cue profile for the *or even* construction in UD English GUM and EWT. Left to right: proportion of tokens with syntactic *parallelism* (matching Universal Part-of-Speech (UPOS) match between contrasted heads for the heads of *X* and *Y*), distribution of *Y*–head UPOS, and prevalence of scalar/contrast cues within five tokens to the left. Error bars are bootstrap 95% intervals (2,000 resamples). Token counts and exact estimates are reported in the repository tables.

struction. Frequency and entrenchment work rapidly (milliseconds to weeks): each token use strengthens memory traces, increasing production probability, which generates more tokens in a self-reinforcing loop (Bybee & Hopper 2001). Editorial norms and genre licensing operate slowly (years to decades): copy-editing workflows and style guide updates formalize emergent patterns, creating explicit standards that then constrain future production. The fast loop generates the pattern; the slow loop crystallizes and transmits it. Perturbation experiments can distinguish their contributions: reducing frequency while maintaining editorial standards should weaken the cluster gradually, whereas removing editorial oversight while maintaining frequency should increase drift without collapse. Both stabilizers are necessary: frequency alone produces transient patterns (internet slang that fades quickly), normative pressure alone produces rigid prescriptions that speakers ignore (failed language reforms). The interaction explains the observed robustness.

These mechanisms – frequency, redundancy, and normativity – are precisely the homeostatic forces that Boyd's framework predicts for socially maintained kinds. Unlike the biophysical constraints that stabilize phonemes, constructions rely more heavily on usage-based learning and community standards. But the empirical signatures are parallel: predictable

PR curves: GUM -> EWT
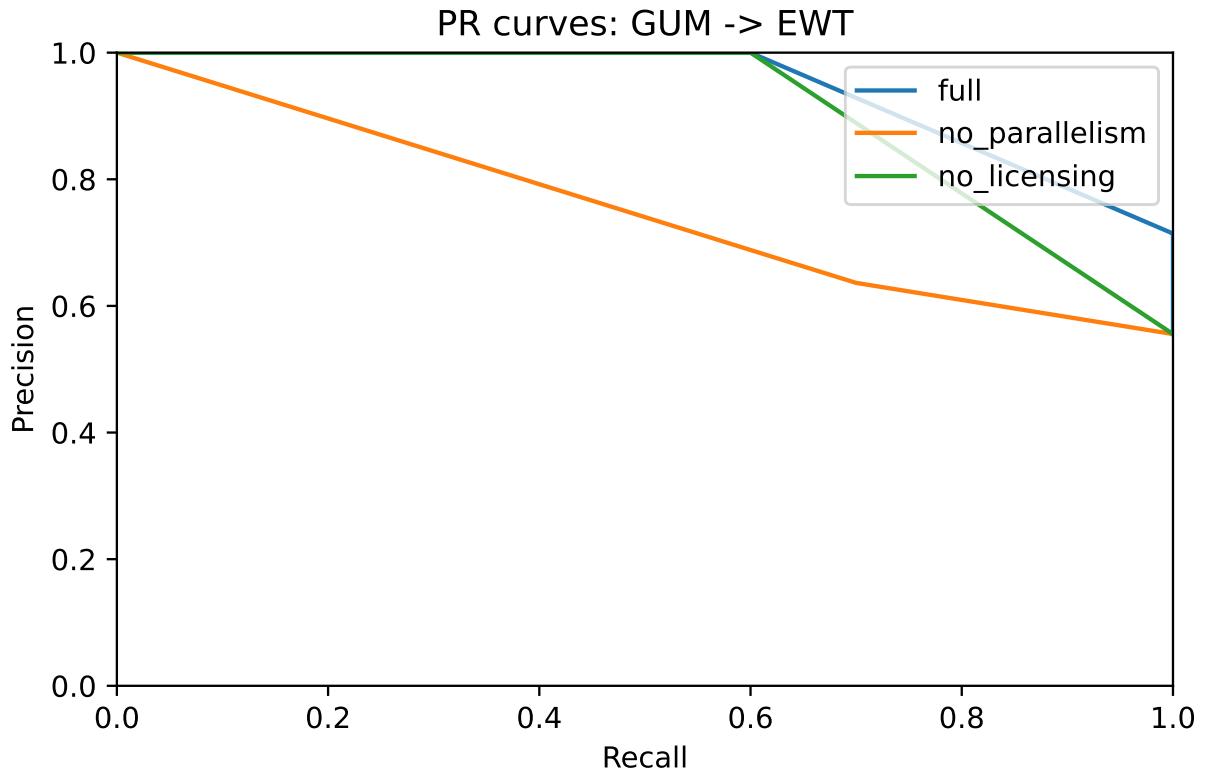
Figure 5: Projectibility and ablation for *or even*. Precision–recall curves for a regularized logistic model using the full cue bundle (anchor + parallelism + scalar/contrast cues; solid) versus ablations (drop parallelism; dashed; drop scalar/contrast cues; dotted). Model is trained on GUM and evaluated on EWT; class prevalence in the target set and train/test construction are held constant across conditions. The full bundle achieves high PR–AUC ($\geq$ 0.70); ablating parallelism produces the largest drop, with licensing cues providing smaller but reliable support.

patterns that degrade systematically when stabilizers are removed.

How do speakers acquire this cue bundle? Children learning *or even* face a distributional learning problem: from scattered instances in input, extract the recurring pattern that the anchor co-occurs with parallelism, scalar/contrast cues, and scalar meaning. Domain-general statistical learning mechanisms – tracking form-meaning co-occurrences, registering cue reliability, chunking frequent sequences – are sufficient (Tomasello 2003, Goldberg 2006). As the construction becomes entrenched through repeated activation, it gains processing advantages (faster recognition, automatic retrieval) that further stabilize it against perturbation.

This is the cognitive implementation of "frequency and entrenchment": repeated exposure → strengthened memory trace → resistance to change.

Frequency drives entrenchment through repeated activation of form-meaning mappings in memory (Bybee & Hopper 2001); cue redundancy provides fallback signals when individual cues are noisy or degraded; normative pressure operates through explicit correction (editorial changes, teaching) and implicit modelling (exposure to edited text). The ablation results in Table 1 provide evidence that these mechanisms matter: removing parallelism or scalar/contrast cues degrades performance because the construction depends on their combined contribution.

The modest sample sizes mean these results are illustrative rather than definitive; general claims about constructions-as-kinds will require broader sampling across construction types and corpora. What *or even* demonstrates is proof-of-concept: the diagnostics can be applied to constructions, and they yield interpretable signatures when they succeed.

A further question concerns constructional inheritance. *Or even* isn't isolated – it shares properties with a family of scalar additive constructions: *let alone*, *much less*, *not to mention*, *never mind*, *to say nothing of*. These patterns signal scalar extremity, but differ in register (formal vs. colloquial), cue reliability (e.g., *much less* shows weaker parallelism), and productivity. An HPC treatment of this family would ask: is the scalar-additive schema itself an HPC kind at a higher level of abstraction, with *or even* as an instantiation? Or are these distinct kinds that happen to share features? The diagnostics extend naturally: if the abstract schema projects across new instances and is maintained by identifiable mechanisms (analogical extension, paradigmatic pressure), it qualifies. Full analysis requires broader constructicon sampling and is left to future work.

The *or even* case completes the demonstration across linguistic levels. Where phonemes showed scaling laws and stability bands, and words showed cohesion under drift, constructions show cross-corpus cue bundles with measurable degradation under ablation. Each level exhibits projectibility maintained by identifiable mechanisms – the hallmark of HPC kinds

in Boyd's naturalistic framework.

5.1 From a flagship to a stratified constructional battery

The *or even* case is deliberately clean: it's frequent enough to analyse in modest corpora, it has a clear scalar-additive semantics, and it supplies an interpretable cue bundle (anchor string, syntactic parallelism, and scalar/contrast cues). But a single construction is too easy to tailor the analysis to. If the construction tier is to do more than demonstrate feasibility, the diagnostics have to survive contact with heterogeneous constructions: different cue geometries, different decoy spaces, and different plausible stabilizers. The aim isn't to argue that *all* constructions are HPC kinds, but to show that the same two tests can be run repeatedly at this tier without smuggling in a bespoke definition of success.

Accordingly, the confirmatory construction-level analysis expands from one flagship to a small, stratified battery of ten targets. Eight are treated as positive candidates; two are included as designed *brakes* cases to force the framework to say *no* (or at least *not at this grain*) when it should. The battery is chosen by three operational criteria. First, each target has to admit a non-trivial near-miss space: there have to be abundant anchor- or frame-matched strings that aren't instances, so that projectibility isn't reducible to string spotting. Second, each target has to be representable by at least three partially independent cues (so ablation is informative). Third, extraction has to be feasible in UD-style dependency parses plus shallow context windows, without hand-built sense inventories or construction-specific annotation.

To avoid corpus-by-construction confounds, a fixed *corpus palette* is pre-declared and applied uniformly across all targets. The core palette spans three UD English treebanks that jointly cover the relevant register range: GUM (multi-genre, including academic, fiction, and spoken/social), EWT (web genres including blogs, newsgroups, emails, and reviews), and GUMReddit (informal social-media dialogue). To increase sample sizes and reduce saturation, the battery is expanded to additional UD English treebanks (ESL, ESLSpok, GEN-

TLE, CHILDES, LinES, PUD, ParTUT, UniDive, ATIS, LittlePrince, Pronouns, CTeTex, PCEDT). Register becomes an explicit experimental factor rather than a per-construction contingency: every construction in the battery is evaluated over the same grid, and sparsity in particular slices is treated as distributional scope (the extent of the kind's population) rather than as a test failure. For register contrasts I use a two-bin split (edited/formal vs informal/spoken) and exclude mixed or ambiguous corpora; assignments are listed in Appendix A.

The selected constructions span four cue regimes (Table 2). Semi-schematic argument-structure patterns with lexical hooks (e.g., *way-* and *time-away*) test whether moderately schematic templates can be treated as local kinds without collapsing into verb-class semantics (Goldberg 1995, 2006). Paired-marker templates (comparative correlatives) shift the burden from lexical anchors to parallelism and clause pairing. Clausal operators with a lexical spine (*just because ... doesn't mean ...*; all-clefts) foreground discourse function and polarity, where licensing cues should matter more than argument structure. Finally, pragmatic "format" constructions (here, *X much?*) are treated as population-relative: their stabilizers are plausibly register-specific, and projection is expected to be strong only within the genres where they circulate.

Two additional targets are included explicitly to discipline the ontology. The first is *resultative* as a pooled analyst's umbrella. Run naively as a single category, it's predicted to behave as "too fat": heterogeneous subtypes should erode cross-corpus projection and wash out ablation signatures. The second is a register-local format construction (*X much?*), included to make locality concrete: the framework predicts that projectibility can succeed in the appropriate population–genre slice while failing (trivially, by sparsity) outside it. In both cases the point isn't to "rescue" a kind claim, but to show that the same diagnostics (and the same thresholds) yield informative failure modes rather than producing an "HPC" verdict by default.

Table 2: Constructional battery for the confirmatory extension of Case C. Each target is specified by a candidate-generation heuristic (to define the evaluation universe), a minimally sufficient cue bundle (for ablation), and a predicted stabilizer profile. Eight are treated as positive candidates; two are designed *brakes* cases ("too fat"; register-local).

Target	Candidate set (heuristic)	Cue bundle (illustrative) and predicted stabilizers
or even	Anchor-present strings containing *or even*	Anchor string; syntactic parallelism of contrasted heads; scalar/contrast cues in left context. Redundancy and norming predict cross-corpus stability and ablation sensitivity.
way-construction	Tokens of *way* with a possessive dependent (*my/your/his ... way*)	Possessor+*way* frame; PP path; eventive verb profile (often effortful traversal). Stabilizers: entrenchment of a schematic template, plus cue redundancy between the nominal frame and path syntax (Goldberg 1995, 2006).
time-away	Verb + object time-NP + particle *away*	Particle *away*; object is a time-denoting NP; activity predicate tendency and durative packaging. Stabilizers: conventional aspectual packaging reinforced by editorial norms and high token frequency.

Target	Candidate set (heuristic)	Cue bundle (illustrative) and predicted stabilizers
Comparative correlative	Two-clause strings with repeated *the* preceding comparative morphology	Paired *the* markers; comparative morphology in both clauses; clause adjacency/parallelism (often punctuation-delimited). Stabilizers: processing and parallelism pressures on a rigid template (cue dependence predicted to be high).
Just because ... doesn't mean ...	Strings containing *just because* or *doesn't mean* with clausal complements	Two-part operator packaging; polarity/negation; discourse contrast markers. Stabilizers: discourse function plus redundant lexical packaging (two-part spine predicts robust ablation deltas).
All-cleft (*All S is NP*)	*all* + clausal dependent + copula (*is/was*)	Cleft-like packaging; copular spine; exhaustivity cues (often reinforced by punctuation and discourse framing). Stabilizers: focus packaging and editorial regularisation of canonical phrasing.
Binominal *N of a N* (*that idiot of a doctor*)	N_1 *of a* N_2 nominal frame	Fixed *of a* frame; evaluative N_1 tendency; agreement/selectional asymmetries (N_2 as semantic head). Stabilizers: frame entrenchment plus conventional evaluative usage patterns.

Target	Candidate set (heuristic)	Cue bundle (illustrative) and predicted stabilizers
N–P–N (*day by day*)	Identical noun lemma repeated with a preposition between	Lexical identity; restricted P inventory; distributive/reciprocal semantics proxied by local syntactic environment. Stabilizers: strong formal rigidity with modest lexical productivity (predicted to split into subfamilies under "too fat" pressure if pooled indiscriminately).
Resultative (pooled) *brakes*	Broad [V Obj XP] frames with XP as AP/PP	Predicted "too fat": pooling distinct subtypes (AP-resultatives; PP/path-like; verb-class specific patterns) should weaken projection and blur ablation signatures. Stratified subkinds are predicted to restore projectibility if (and only if) mechanisms are genuinely local.
X much? brakes	Utterances ending in *much* with fragment-like syntax	Sentence-final *much*; predicate fragment; stance punctuation. Predicted register-local homeostasis: strong within informal dialogue/web-like text, weak or absent elsewhere (locality as a positive prediction, not a nuisance).

5.2   Construction battery results

To reduce tautological saturation, instance labels are defined by a stricter heuristic than the three cue proxies, so cues predict rather than define labels. The heuristics were also broadened to target structural signatures (copular all-clefts, possessive+*way* with path framing), while comparative correlatives were tightened to require genuine comparatives rather than generic *the* sequences. Under this regime, five constructions meet the minimum thresholds

(Table 3): N–P–N, resultative_adj, resultative_pooled, all-clefts, and the *way* construction. The expanded corpus palette yields 34 estimable train–test pairs for these constructions (notably across CHILDES and LinES in addition to GUM/EWT). Table 4 summarizes the full-model PR–AUC and ablation deltas for those pairs. Full PR–AUC now ranges from 0.58 to 1.00; Cue 3 ablations still produce the largest drops overall, but all-clefts show moderate sensitivity to Cue 1 and the *way* construction shows strong sensitivity to Cue 1/Cue 2.

By the preregistered decision rule, N–P–N and the *way* construction count as *provisional positives at this grain*. All-clefts are borderline (two train–test directions exceed the 0.70 target; others do not), and the resultative outcomes remain *brakes*: both the pooled set and the AP subtype depend almost entirely on Cue 3, consistent with the "too fat" prediction, while the PP/path subtype is too sparse to evaluate. Comparative correlatives no longer meet the evaluability thresholds under the tightened heuristic, and binominal *N of a N* plus the remaining targets (*or even*, *time-away*, *just because . . . doesn't mean . . .*, *X much?*) still fall below prevalence thresholds, so no kind claim is made for them in this run. Following the Gelman-style checks, the heuristic is frozen at this level; further changes would count as a new task design rather than an incremental tweak.

The implication is deliberately narrow: the construction tier does not license a blanket HPC claim. Kindhood is earned case-by-case, and at this grain it attaches to N–P–N (and provisionally to *way*), while the resultatives behave as a failure mode that supports the "too fat" diagnosis. Cue dominance varies across targets (Cue 3 for resultatives; Cue 1/Cue 2 for *way*), suggesting that multi-cue redundancy is not yet robustly established in these slices; stronger evidence for redundancy will require larger samples and the planned downsampling and covariance checks.

The homeostasis diagnostics temper the story further. Cue covariance (Table 5) is often undefined because one cue is invariant across positives, and where it is defined the remaining cue pair tends to show modest or negative association (e.g., *or even*). This points to cue specialization or trade-offs rather than robust redundancy in the current slices.

Downsampling the training data to 25% produces only small changes in PR–AUC (max $|\Delta| \approx 0.08$; a negligible change) across the 34 estimable pairs (Table 6), suggesting that even after loosening the heuristics the cue bundle still yields stable separability on these candidate sets. As a result, the perturbation test remains weakly informative at this grain and likely requires broader samples or less restrictive label heuristics to become diagnostic. Bootstrap CIs (Table 7) reinforce the pattern: N–P–N and *way* remain entirely above the 0.70 target, all-clefts straddle the threshold, and comparative correlatives fall below evaluability under the tightened heuristic. Shuffled-label baselines (Table 8) track prevalence – especially for the high-prevalence N–P–N pairs – highlighting that some high PR–AUC values are prevalence-driven rather than cue-driven.

Table 3: Construction battery prevalence by corpus. Values are candidates (positives).

Construction	Corpus	Candidates (positives)
all_cleft	ATIS	32(0)
all_cleft	CHILDES	64(15)
all_cleft	CTETEX	8(0)
all_cleft	ESLSPOK	3(0)
all_cleft	EWT	164(29)
all_cleft	GENTLE	6(0)
all_cleft	GUM	152(25)
all_cleft	LINES	75(5)
all_cleft	LITTLEPRINCE	8(1)
all_cleft	PARTUT	23(2)
all_cleft	PUD	5(0)
binominal_of_a	ATIS	20(0)

Continued on next page

Construction	Corpus	Candidates (positives)
binominal_of_a	CHILDES	23(0)
binominal_of_a	CTETEX	4(0)
binominal_of_a	ESLSPOK	2(0)
binominal_of_a	EWT	63(1)
binominal_of_a	GENTLE	9(0)
binominal_of_a	GUM	88(3)
binominal_of_a	LINES	99(3)
binominal_of_a	LITTLEPRINCE	5(0)
binominal_of_a	PARTUT	20(1)
binominal_of_a	PUD	10(0)
comparative_correlative	CHILDES	2(2)
comparative_correlative	EWT	13(4)
comparative_correlative	GUM	9(2)
comparative_correlative	LINES	3(3)
just_because_doesnt_mean	CHILDES	1(1)
just_because_doesnt_mean	EWT	1(1)
npn	CHILDES	3(3)
npn	CTETEX	1(1)
npn	ESLSPOK	1(1)
npn	EWT	24(18)
npn	GUM	23(19)
npn	LINES	16(15)
npn	LITTLEPRINCE	3(3)
npn	PARTUT	6(6)

Continued on next page

Construction	Corpus	Candidates (positives)
npn	PUD	1(1)
or_even	EWT	18(10)
or_even	GUM	12(5)
or_even	LINES	4(1)
or_even	LITTLEPRINCE	2(0)
or_even	PARTUT	3(2)
resultative_adj	ATIS	2(0)
resultative_adj	CHILDES	144(16)
resultative_adj	CTETEX	7(0)
resultative_adj	ESL	209(0)
resultative_adj	ESLSPOK	10(0)
resultative_adj	EWT	315(28)
resultative_adj	GENTLE	26(1)
resultative_adj	GUM	272(17)
resultative_adj	GUMREDDIT	31(0)
resultative_adj	LINES	134(10)
resultative_adj	LITTLEPRINCE	10(0)
resultative_adj	PARTUT	50(2)
resultative_adj	PUD	20(1)
resultative_pooled	ATIS	3171(0)
resultative_pooled	CHILDES	14645(19)
resultative_pooled	CTETEX	348(0)
resultative_pooled	ESL	5305(0)
resultative_pooled	ESLSPOK	1128(0)

Construction	Corpus	Candidates (positives)
resultative_pooled	EWT	12226(29)
resultative_pooled	GENTLE	686(1)
resultative_pooled	GUM	10407(17)
resultative_pooled	GUMREDDIT	842(0)
resultative_pooled	LINES	4857(10)
resultative_pooled	LITTLEPRINCE	304(0)
resultative_pooled	PARTUT	2030(2)
resultative_pooled	PRONOUNS	60(0)
resultative_pooled	PUD	872(1)
resultative_pp	ATIS	1(0)
resultative_pp	CHILDES	22(3)
resultative_pp	ESL	5(0)
resultative_pp	EWT	22(1)
resultative_pp	GENTLE	4(0)
resultative_pp	GUM	15(0)
resultative_pp	GUMREDDIT	2(0)
resultative_pp	LINES	2(0)
resultative_pp	PARTUT	3(0)
time_away	CHILDES	145(3)
time_away	CTETEX	2(0)
time_away	ESLSPOK	5(0)
time_away	EWT	93(9)
time_away	GENTLE	11(0)
time_away	GUM	86(3)

Construction	Corpus	Candidates (positives)
time_away	LINES	61(3)
time_away	LITTLEPRINCE	8(0)
time_away	PARTUT	4(1)
time_away	PUD	5(0)
way_construction	ATIS	174(0)
way_construction	CHILDES	156(1)
way_construction	ESLSPOK	10(0)
way_construction	EWT	268(14)
way_construction	GENTLE	12(0)
way_construction	GUM	235(6)
way_construction	LINES	121(20)
way_construction	LITTLEPRINCE	5(0)
way_construction	PARTUT	44(2)
way_construction	PUD	10(3)
x_much	CHILDES	33(11)
x_much	ESLSPOK	46(7)
x_much	EWT	14(2)
x_much	GUM	8(2)
x_much	LINES	7(1)
x_much	LITTLEPRINCE	1(1)
x_much	PARTUT	1(0)

5.3 Candidate sets, decoys, and cue ablations

Across targets, the projectibility test is kept methodologically uniform. Each construction is evaluated on an *anchor- or frame-restricted* universe of candidates generated by an explicit heuristic (Table 2), and within that universe the task is to distinguish true instances from near misses. This restriction serves two purposes. It prevents trivial wins in which a unique lexical anchor would solve the problem, and it makes the negative class interpretable: decoys aren't arbitrary non-instances, but strings that are close enough that the constructional cue bundle has to do real work. Within that candidate set, class prevalence is reported explicitly and PR–AUC is used because it remains informative under imbalance.

With the fixed corpus palette in place, every construction undergoes the same three-step reporting protocol:

1. **Prevalence map.** For each corpus (and, where metadata permit, each genre slice), count anchor- or frame-matched candidates and validated instances. Zeros are informative: they delimit the construction's *distributional scope* rather than registering as test failures.

2. **In-register projectibility.** Where prevalence thresholds are met (Appendix A), test transfer between corpora or slices in the same register class (e.g., informal↔informal; edited↔edited). This uses the standard PR–AUC$\geq$ 0.70 target.

3. **Cross-register stress test.** Train in one register class and test in another (e.g., GUMReddit→academic GUM). For register-neutral constructions like *or even*, cross-register transfer should remain robust; for register-local constructions like *X much?*, it's predicted to degrade sharply – a falsifiable signature of locality.

Cue bundles are specified to be both linguistically interpretable and mechanically ablatable. For lexically hooked templates (*or even*, *way*, *time-away*, binominal *of*), ablations separate (i) anchor/frame material from (ii) structural packaging (e.g., possessive+*way*, PP path; time-NP objecthood; *of a* frame) and (iii) scalar/contrast cues (polarity, contrast

markers, stance punctuation). For paired-marker templates (comparative correlatives), ablations remove parallelism cues (comparative morphology; repeated *the*; clause pairing) in isolation. For pragmatic formats (*X much?*), ablations separate the overt marker (*much*) from fragment syntax and stance punctuation, with the expectation that no single cue is sufficient in naturally noisy data. The success criterion follows the flagship analysis: cross-corpus PR–AUC around or above 0.70 for the full bundle, with clear ablation sensitivity (especially to parallelism) and prevalence thresholds enforced (§2).

The homeostasis diagnostic is likewise kept parallel to the flagship. For each construction the analysis reports (i) cue covariance (e.g., pairwise ϕ coefficients and their bootstrap intervals) as a direct signature of clustering, (ii) downsampling sensitivity as a perturbation proxy (frequency as a stabilizer), and (iii) register stratification when appropriate (genre licensing as a stabilizer). The two *brakes* cases are expected to behave differently. The pooled resultative is predicted to show weak cue covariance and unstable cross-corpus transfer unless it's stratified into more local equilibria, in line with the "too fat" failure mode (§6). The *X much?* format is predicted to show the opposite pattern: strong internal covariance and projection within the relevant population slice, paired with sharp attenuation outside it. In both cases, the outcome is informative either way. A pooled umbrella that projects robustly without stratification would be evidence *for* a higher-level kind; a register-local format that projects widely across corpora would undermine the locality claim and force a different mechanism story.

The upshot is that the construction tier can be treated with the same epistemic discipline as the phoneme and word tiers. A construction counts as an HPC kind only if (i) its cue bundle supports out-of-sample inference in a way that survives explicit decoy baselines, and (ii) that bundle is plausibly maintained by stabilizers whose signatures are visible as cue covariance and predictable perturbation sensitivity. The battery isn't an attempt to catalogue English constructions; it's a stress test of whether the diagnostics behave coherently across heterogeneous targets, and whether the failure modes are empirically tractable rather

than merely philosophical.

6   Where HPC fails: thin, fat, and negative classes

The diagnostics in §1 and §2 are symmetric: they license kindhood when a category projects and when stabilizers with the right signatures can be named; they also tell us when to say *no*. Because kinds are discovered empirically rather than stipulated – with their boundaries determined by where mechanisms maintain covariance – failure isn't a metaphysical verdict about the string or pattern in the abstract. It's a claim that here and now, we lack the predictive grip and identifiable mechanisms that HPC requires under Boyd's view.

Some categories are too thin. Nonce coinages and campaign-season blends that never diffuse, idiolectal "style phonemes," and child-only overregularizations within adult Standard English don't pass the projectibility test in the relevant population–time slice: held-out prediction collapses, and the stabilizers one would expect to bind properties (frequency, entrenchment, community norms) are either absent or act in the opposite direction. In our terms, the decision rule fails in both halves: PR–AUC sits at baseline and no credible mechanism–signature pairing survives sensitivity checks. These cases are explananda for learning or diffusion, but not kinds.

Some categories are too fat. Cross-linguistic umbrellas such as "resultative" or "ditransitive" pool patterns maintained by different morphosyntactic resources, cue reliabilities, and norming regimes. The pooled set can look impressive descriptively, but the mechanism story is disjunctive: dispersion in one language, selectional licensing in another, constructional analogy in a third (Croft 2001, Haspelmath 2010). When the projectibility assay is run at this granularity, cross-corpus prediction drifts toward family- or area-specific quirks, ablations fail to show stable redundancy, and effects wash out under lineage-pruning. The right move, on a Boyd-style account, is to localize the ontology: retain language-internal equilibria as kinds and treat the global umbrella as an interest-relative taxonomy rather than a

single HPC kind. In some cases, stratification restores projection: if the pooled resultative is split into AP-resultative subtypes vs. PP-path subtypes, each subtype may show the cue covariance and cross-corpus transfer that the pooled category lacks. That stratified outcome would confirm the "too fat" diagnosis while identifying the local kinds it concealed.

Some categories are merely negative. Complement classes – "ungrammatical strings", "all exceptions to rule R" – are defined by what they're not rather than what they are. Such sets are defined by failure to meet norms, not by a family of co-instantiated, causally sustained properties. They don't project – there's no stable covariance to learn – and they don't admit a non-accidental mechanism story. By the criteria in §2, they fail the homeostasis test by construction. Miller's mechanism-first treatment of word-kinds is instructive here: the relevant covariation is sustained by internal and external stabilizers; a complement class has no such base. A caveat: *structured subsets* within a complement class can be HPC candidates in their own right. L2 transfer clusters (e.g., persistent article-drop errors among speakers of articleless L1s) may show stable covariance, predictable error profiles, and mechanism–signature pairings that the grab-bag "all L2 errors" lacks. The framework doesn't rule out kindhood for negative-seeming sets; it demands the same evidence of projectibility and stabilizer-consistent signatures that any candidate must meet.

There are borderline zones. "Cranberry" elements – bound morphemes that appear in just one or two words, like *cran-* in *cranberry* – can form local kinds if a distributional and prosodic profile stabilizes across a small family of items (e.g., the Latinate bound root *-ceive/-cept* in *receive*, *perceive*, *conception* shows recurrent selectional frames and stress behaviour, whereas *cran-* in *cranberry* does not and so degenerates into an analyst's one-off grouping). Base-identical exponents in morphology can be treated as kinds when the distributional covariance and paradigm pressure produce reliable out-of-sample inferences (e.g., base plurals in *sheep*, *deer* and base-identical preterite in *hit*, *set* may predict paradigm-internal contrasts and error patterns; absent that, a "zero form" is often a bookkeeping device). In phonology, allophonic habits become kinds when they're normed and extend across the speech community (e.g.,

North American English intervocalic /t, d/ → [ɾ] (*water*) or contextually dark [ɫ]) whereas fleeting stylistic effects (an emphatic release on a single token, a momentary creak on one syllable) do not. The diagnostics are agnostic about representation: what matters is whether predictive signatures persist under the robustness checks we have fixed in advance (§2).

The practical value of the failures is twofold. First, they prevent over-generalization: we avoid declaring "everything is an HPC" by tying kindhood to recomputable tests and to stabilizer–signature pairs. Second, they guide analysis: when a proposal fails, the failure mode – thin, fat, or negative – indicates which lever to pull next (look for diffusion and norms; localize the ontology; abandon complement classes). That division of labour is the point of bringing an HPC realism to language: the categories that travel well do so because mechanisms keep enough of their properties together, and the ones that don't travel are precisely those where such homeostasis is missing.

The failure taxonomy also blocks the pathologies Rubin (2008) raises for moral HPCs: isolated goods, structural mismatches, and weak induction. Only clusters with causally important stabilizers and demonstrated projectibility earn kindhood here.

7   Predictions and disconfirmers

The diagnostics and thresholds specified in §2 generate falsifiable predictions. This section makes explicit what would disconfirm the HPC account and presents two cross-cutting perturbation tests.

The HPC claim generates clear disconfirmers. If the /y/ effect collapses under lineage-pruning – it fails. If semantic drift produces no cohesion loss – it fails. If ablations don't degrade construction recognition – it fails. The full specification in §2 lists additional conditions: inventory scatter when controlled, random temporal errors, single-cue sufficiency. But the principle is the same. Each test has declared thresholds; each failure forces retreat to thinner accounts.

Two experiments test homeostatic maintenance directly. First, frequency downsampling should weaken stabilization: reducing construction tokens to 25% should degrade PR–AUC by ≥ 0.10 or reduce cue rates by $\geq 20\%$. Second, scaling should generalize: just as /y/ appears preferentially in larger vowel inventories, rare constructional variants should concentrate in corpora with larger constructicon repertoires (monotonic increase across quartiles with non-overlapping CIs at extremes).

These tests operationalize the core claim: linguistic categories qualify as HPCs when they project via identifiable mechanisms. Where thresholds are met, kindhood is warranted; where not, we should prefer thinner or more local accounts.

8 General discussion

The claim advanced here is deliberately modest. It isn't a new architecture of grammar, nor an insistence that every analyst's category is a kind. It's a method for deciding when a linguistic category *earns* kindhood by homeostatic property-cluster lights: it has to project to held-out data, and the projection has to be underwritten by stabilizers whose signatures we can name and check.

Across tiers the story is the same. Phoneme inventories show a stability band and a scaling relation for /y/ that survive genealogical and areal controls. Some words drift while retaining enough covariance for out-of-sample recovery, while others do not. The *or even* construction travels across corpora and loses predictive grip when a stabilizing cue is removed. Those are the observable faces of homeostasis at the population–time scales analysed here (§3–5).

The pay-off is twofold. First, predictive grip: the diagnostics in §1 and decision rules in §2 force us to say in advance what counts as success and what would change our minds. Success isn't a rhetorical gloss ("striking regularity") but concrete measures – slopes with uncertainty intervals, mass within bands, cross-corpus areas under curve, ablation deltas,

and calibration metrics that readers can recompute.

Second, an ontology with brakes: kinds are discovered through evidence rather than declared by fiat, and they are local equilibria (maintained by specific mechanisms in specific populations) rather than universal essences (Boyd 1991, 1999). That stance blocks overreach. Cross-linguistic umbrellas like "resultative" fail as single kinds because they pool heterogeneous mechanisms. Thin proposals (nonce items, one-off blends) and complement classes lack the stabilizer base that projectibility requires (§6). In between lie the categories that travel: their properties cohere because mechanisms keep them together.

The stabilizers form a stack, not a single cause. At the phoneme tier, biophysical constraints carve the design space, developmental learning binds cues, sociocultural norms transmit inventories (Stevens 1989, Liljencrants & Lindblom 1972, Lindblom 1990, Ekström et al. 2025). At the lexical tier, frequency entrenches forms, editorial standards enforce conventions, usage communities police extensions. At the constructional tier, cue redundancy protects against noise, normative pressure corrects deviations, genre licensing regulates distribution. The mechanisms shift in their balance: articulatory constraints weigh heavily for phonemes, frequency and norms for words, cue redundancy and editorial pressure for constructions. But at every tier, multiple forces interact: body, cognition, and society always co-stabilize. The logic remains: covariance maintained by identifiable forces at the relevant timescale. This stack is exactly what a Boyd-style naturalism predicts for socially scaffolded kinds (Boyd 2000, Khalidi 2013).

The HPC framework carries theoretical commitments. Treating constructions as kinds commits to storage and form-meaning pairing (compatible with Construction Grammar, HPSG, Sign-Based CxG, or lexicalist frameworks); treating words as kinds commits to lexical entries with associated property clusters (contra extreme proceduralism). The diagnostics themselves – projectibility and homeostasis tests – could in principle be applied within other frameworks, but the substantive claim that linguistic categories *are* HPC kinds entails specific ontological commitments. The value isn't framework-neutrality but disciplined realism:

kinds are discovered through evidence, not stipulated by theory.

There are limits. PHOIBLE counts depend on coding choices and coverage; UD parses vary by genre and version; corpus composition affects both the drift and construction results. The paper addresses these in the small – alternative inventory codings, lineage pruning and macro-area controls, string-only baselines, ablation and calibration checks – but they remain sources of uncertainty. The scope is also intentionally narrow: English for the construction case; contemporary written registers for the word case; a PHOIBLE snapshot for phonology. The point isn't universality but a disciplined procedure that can be extended and falsified elsewhere.

Three extensions seem especially promising. One is developmental and modelling work that turns stabilizers into mechanisms with dynamics. Iterated-learning models can test which combinations of cue redundancy, frequency, and norm enforcement are sufficient for the observed signatures, and child-directed or learner corpora can implement the frequency perturbations preregistered in §2. Another is cross-linguistic generalization at the right grain: language-internal construction families with comparable cue bundles, analysed with the same projectibility and homeostasis tests. Both directions preserve the epistemic discipline that makes HPC realism useful in cognitive science: kinds aren't declared by fiat but by patterns that survive baselines, controls, and perturbations.

A third is "general social agents". The approach offers a low-cost testbed for HPC-style predictions without opening a new empirical front. Recent work from Manning & Horton (2025) builds theory-grounded LLM agents from small human datasets and validates them across distinct settings; in preregistered studies these agents transfer out of domain and beat both off-the-shelf baselines and equilibrium benchmarks across large families of novel games. Used cautiously, such agents could pre-screen the frequency-downsampling and repertoire-size predictions here and surface disconfirmers before running new human studies.

Finally, a word on alternatives. Purely distributional views get part of the way – clusters can be found at every tier – but they lack a story about *why* some clusters are projectible

while others evaporate. The mechanism-first stance advanced here supplies that story and demands discriminating checks. Essentialist views capture projectibility by stipulation but have little to say about drift, diversity, and social maintenance. The present approach keeps the realism while naturalizing it: kinds are whatever supports reliable inference because stabilizers – biophysical, developmental, social – keep enough of the relevant properties together (Miller 2021, Boyd 1991, 1999). The figures and tables in this paper are small demonstrations of that general point. Where the signed effects and thresholds are met, a kind claim is warranted; where they aren't, the label should be withheld. That discipline, and not any particular representation, is the contribution.

Table 4: Construction battery results for estimable train–test pairs. Cue ablations report the change in PR–AUC relative to the full model (positive values indicate drops).

Construction	Direction	n_{tr}	n_{te}	PR–AUC	ΔC1	ΔC2	ΔC3
all_cleft	childes→ewt	64(15)	164(29)	0.614	0.146	0.000	0.348
all_cleft	childes→gum	64(15)	152(25)	0.580	0.143	0.000	0.341
all_cleft	ewt→childes	164(29)	64(15)	0.830	0.046	0.000	0.536
all_cleft	ewt→gum	164(29)	152(25)	0.580	0.143	0.000	0.341
all_cleft	gum→childes	152(25)	64(15)	0.830	0.046	0.000	0.536
all_cleft	gum→ewt	152(25)	164(29)	0.614	0.146	0.000	0.348
npn	ewt→gum	24(18)	23(19)	1.000	0.000	0.000	0.174
npn	gum→ewt	23(19)	24(18)	1.000	0.000	0.000	0.250
resultative_adj	childes→ewt	144(16)	315(28)	0.933	0.000	0.000	0.844
resultative_adj	childes→gum	144(16)	272(17)	0.895	0.000	0.000	0.832
resultative_adj	childes→lines	144(16)	134(10)	1.000	0.000	0.000	0.925
resultative_adj	ewt→childes	315(28)	144(16)	1.000	0.000	0.000	0.889
resultative_adj	ewt→gum	315(28)	272(17)	0.895	0.000	0.000	0.832
resultative_adj	ewt→lines	315(28)	134(10)	1.000	0.000	0.000	0.925
resultative_adj	gum→childes	272(17)	144(16)	1.000	0.000	0.000	0.889
resultative_adj	gum→ewt	272(17)	315(28)	0.933	0.000	0.000	0.844
resultative_adj	gum→lines	272(17)	134(10)	1.000	0.000	0.000	0.925
resultative_adj	lines→childes	134(10)	144(16)	1.000	0.000	0.000	0.889
resultative_adj	lines→ewt	134(10)	315(28)	0.933	0.000	0.000	0.844
resultative_adj	lines→gum	134(10)	272(17)	0.895	0.000	0.000	0.832
resultative_pooled	childes→ewt	14645(19)	12226(29)	0.879	0.000	0.000	0.793
resultative_pooled	childes→gum	14645(19)	10407(17)	0.680	0.000	0.000	0.621
resultative_pooled	childes→lines	14645(19)	4857(10)	1.000	0.000	0.000	0.926
resultative_pooled	ewt→childes	12226(29)	14645(19)	0.905	0.000	0.000	0.790
resultative_pooled	ewt→gum	12226(29)	10407(17)	0.680	0.000	0.000	0.621
resultative_pooled	ewt→lines	12226(29)	4857(10)	1.000	0.000	0.000	0.926
resultative_pooled	gum→childes	10407(17)	14645(19)	0.905	0.000	0.000	0.790
resultative_pooled	gum→ewt	10407(17)	12226(29)	0.879	0.000	0.000	0.793
resultative_pooled	gum→lines	10407(17)	4857(10)	1.000	0.000	0.000	0.926
resultative_pooled	lines→childes	4857(10)	14645(19)	0.905	0.000	0.000	0.790
resultative_pooled	lines→ewt	4857(10)	12226(29)	0.879	0.000	0.000	0.793
resultative_pooled	lines→gum	4857(10)	10407(17)	0.680	0.000	0.000	0.621
way_construction	ewt→lines	268(14)	121(20)	1.000	0.428	0.240	0.000
way_construction	lines→ewt	121(20)	268(14)	0.887	0.656	0.208	0.063

Table 5: Cue covariance (phi) among positive instances in the construction battery. Bootstrap 95% CIs are reported in the CSV.

Construction	Corpus	n_{pos}	ϕ_{12}	ϕ_{13}	ϕ_{23}
all_cleft	CHILDES	15	–	0.327	–
all_cleft	EWT	29	–	–	–
all_cleft	GUM	25	–	0.261	–
npn	EWT	18	–	–	–
npn	GUM	19	–	–	–
npn	LINES	15	–	–	–
or_even	EWT	10	–	–	-0.535
resultative_adj	CHILDES	16	–	–	–
resultative_adj	EWT	28	–	–	–
resultative_adj	GUM	17	–	–	–
resultative_adj	LINES	10	–	–	–
resultative_pooled	CHILDES	19	–	–	–
resultative_pooled	EWT	29	–	–	–
resultative_pooled	GUM	17	–	–	–
resultative_pooled	LINES	10	–	–	–
way_construction	EWT	14	–	–	–
way_construction	LINES	20	–	–	–
x_much	CHILDES	11	–	–	-0.418

Table 6: Downsampling sensitivity for estimable construction pairs (train set reduced to 25%).

Construction	Direction	n_{tr}	n_{te}	PR–AUC	Δ down
all_cleft	childes→ewt	64(15)	164(29)	0.614	0.030
all_cleft	childes→gum	64(15)	152(25)	0.580	0.038
all_cleft	ewt→childes	164(29)	64(15)	0.830	0.003
all_cleft	ewt→gum	164(29)	152(25)	0.580	0.007
all_cleft	gum→childes	152(25)	64(15)	0.830	0.011
all_cleft	gum→ewt	152(25)	164(29)	0.614	0.023
npn	ewt→gum	24(18)	23(19)	1.000	0.000
npn	gum→ewt	23(19)	24(18)	1.000	0.000
resultative_adj	childes→ewt	144(16)	315(28)	0.933	-0.000
resultative_adj	childes→gum	144(16)	272(17)	0.895	0.000
resultative_adj	childes→lines	144(16)	134(10)	1.000	0.000
resultative_adj	ewt→childes	315(28)	144(16)	1.000	0.000
resultative_adj	ewt→gum	315(28)	272(17)	0.895	0.000
resultative_adj	ewt→lines	315(28)	134(10)	1.000	0.000
resultative_adj	gum→childes	272(17)	144(16)	1.000	0.000
resultative_adj	gum→ewt	272(17)	315(28)	0.933	-0.000
resultative_adj	gum→lines	272(17)	134(10)	1.000	0.000
resultative_adj	lines→childes	134(10)	144(16)	1.000	0.000
resultative_adj	lines→ewt	134(10)	315(28)	0.933	-0.000
resultative_adj	lines→gum	134(10)	272(17)	0.895	0.000
resultative_pooled	childes→ewt	14645(19)	12226(29)	0.879	0.000
resultative_pooled	childes→gum	14645(19)	10407(17)	0.680	-0.000
resultative_pooled	childes→lines	14645(19)	4857(10)	1.000	0.000
resultative_pooled	ewt→childes	12226(29)	14645(19)	0.905	0.000
resultative_pooled	ewt→gum	12226(29)	10407(17)	0.680	-0.000
resultative_pooled	ewt→lines	12226(29)	4857(10)	1.000	0.000
resultative_pooled	gum→childes	10407(17)	14645(19)	0.905	0.000
resultative_pooled	gum→ewt	10407(17)	12226(29)	0.879	0.000
resultative_pooled	gum→lines	10407(17)	4857(10)	1.000	0.000
resultative_pooled	lines→childes	4857(10)	14645(19)	0.905	0.000
resultative_pooled	lines→ewt	4857(10)	12226(29)	0.879	0.000
resultative_pooled	lines→gum	4857(10)	10407(17)	0.680	-0.000
way_construction	ewt→lines	268(14)	121(20)	1.000	0.003
way_construction	lines→ewt	121(20)	268(14)	0.887	0.004

Table 7: Bootstrap CIs (2.5/97.5%) for construction battery PR–AUC and ablation deltas.

Construction	Direction	PR–AUC (CI)	ΔC1 (CI)	ΔC2 (CI)	
all_cleft	childes→ewt	0.584 [0.354, 0.758]	0.116 [-0.074, 0.264]	0.000 [0.000, 0.000]	0.325 [0.
all_cleft	childes→gum	0.555 [0.334, 0.743]	0.120 [-0.065, 0.258]	0.000 [0.000, 0.000]	0.318 [0.
all_cleft	ewt→childes	0.818 [0.585, 0.980]	0.049 [-0.038, 0.178]	0.000 [0.000, 0.000]	0.530 [0.
all_cleft	ewt→gum	0.593 [0.401, 0.756]	0.153 [0.047, 0.287]	0.000 [0.000, 0.000]	0.342 [0.
all_cleft	gum→childes	0.836 [0.629, 0.969]	0.042 [-0.040, 0.149]	0.000 [0.000, 0.000]	0.530 [0.
all_cleft	gum→ewt	0.612 [0.445, 0.769]	0.147 [0.034, 0.275]	0.000 [0.000, 0.000]	0.346 [0.
npn	ewt→gum	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.175 [0.
npn	gum→ewt	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.251 [0.
resultative_adj	childes→ewt	0.925 [0.833, 1.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.835 [0.
resultative_adj	childes→gum	0.889 [0.700, 1.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.828 [0.
resultative_adj	childes→lines	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.924 [0.
resultative_adj	ewt→childes	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.889 [0.
resultative_adj	ewt→gum	0.885 [0.727, 1.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.823 [0.
resultative_adj	ewt→lines	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.927 [0.
resultative_adj	gum→childes	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.889 [0.
resultative_adj	gum→ewt	0.933 [0.818, 1.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.846 [0.
resultative_adj	gum→lines	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.925 [0.
resultative_adj	lines→childes	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.890 [0.
resultative_adj	lines→ewt	0.928 [0.838, 1.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.839 [0.
resultative_adj	lines→gum	0.890 [0.736, 1.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.829 [0.
resultative_pooled	childes→ewt	0.877 [0.765, 0.972]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.791 [0.
resultative_pooled	childes→gum	0.680 [0.464, 0.850]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.622 [0.
resultative_pooled	childes→lines	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.927 [0.
resultative_pooled	ewt→childes	0.913 [0.768, 1.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.800 [0.
resultative_pooled	ewt→gum	0.677 [0.500, 0.842]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.616 [0.
resultative_pooled	ewt→lines	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.926 [0.
resultative_pooled	gum→childes	0.909 [0.778, 1.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.796 [0.
resultative_pooled	gum→ewt	0.885 [0.750, 0.973]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.800 [0.
resultative_pooled	gum→lines	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.928 [0.
resultative_pooled	lines→childes	0.901 [0.750, 1.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.788 [0.
resultative_pooled	lines→ewt	0.879 [0.756, 0.976]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.794 [0.
resultative_pooled	lines→gum	0.675 [0.467, 0.858]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.617 [0.
way_construction	ewt→lines	1.000 [1.000, 1.000]	0.435 [0.230, 0.649]	0.245 [0.087, 0.459]	-0.000 [0.
way_construction	lines→ewt	0.893 [0.764, 1.000]	0.650 [0.489, 0.823]	0.210 [0.101, 0.378]	0.063 [0.

Table 8: Shuffled-label PR–AUC baselines for the construction battery.

Construction	Direction	Full PR–AUC	Shuffle mean	Shuffle CI
all_cleft	childes→ewt	0.614	0.297	[0.138, 0.614]
all_cleft	childes→gum	0.580	0.294	[0.123, 0.580]
all_cleft	ewt→childes	0.830	0.429	[0.180, 0.830]
all_cleft	ewt→gum	0.580	0.298	[0.123, 0.580]
all_cleft	gum→childes	0.830	0.421	[0.180, 0.830]
all_cleft	gum→ewt	0.614	0.314	[0.138, 0.614]
npn	ewt→gum	1.000	0.905	[0.826, 1.000]
npn	gum→ewt	1.000	0.889	[0.750, 1.000]
resultative_adj	childes→ewt	0.933	0.524	[0.089, 0.933]
resultative_adj	childes→gum	0.895	0.529	[0.062, 0.895]
resultative_adj	childes→lines	1.000	0.597	[0.075, 1.000]
resultative_adj	ewt→childes	1.000	0.578	[0.111, 1.000]
resultative_adj	ewt→gum	0.895	0.516	[0.062, 0.895]
resultative_adj	ewt→lines	1.000	0.519	[0.075, 1.000]
resultative_adj	gum→childes	1.000	0.360	[0.111, 1.000]
resultative_adj	gum→ewt	0.933	0.363	[0.089, 0.933]
resultative_adj	gum→lines	1.000	0.408	[0.075, 1.000]
resultative_adj	lines→childes	1.000	0.622	[0.111, 1.000]
resultative_adj	lines→ewt	0.933	0.541	[0.089, 0.933]
resultative_adj	lines→gum	0.895	0.533	[0.062, 0.895]
resultative_pooled	childes→ewt	0.879	0.035	[0.002, 0.106]
resultative_pooled	childes→gum	0.680	0.029	[0.002, 0.680]
resultative_pooled	childes→lines	1.000	0.038	[0.002, 0.097]
resultative_pooled	ewt→childes	0.905	0.140	[0.001, 0.905]
resultative_pooled	ewt→gum	0.680	0.073	[0.002, 0.680]
resultative_pooled	ewt→lines	1.000	0.115	[0.002, 1.000]
resultative_pooled	gum→childes	0.905	0.074	[0.001, 0.905]
resultative_pooled	gum→ewt	0.879	0.069	[0.002, 0.879]
resultative_pooled	gum→lines	1.000	0.065	[0.002, 1.000]
resultative_pooled	lines→childes	0.905	0.047	[0.001, 0.134]
resultative_pooled	lines→ewt	0.879	0.060	[0.002, 0.879]
resultative_pooled	lines→gum	0.680	0.028	[0.002, 0.059]
way_construction	ewt→lines	1.000	0.409	[0.128, 1.000]
way_construction	lines→ewt	0.887	0.272	[0.041, 0.887]

A Statistical specifications and robustness checks

This appendix provides complete technical specifications for the analyses in the main text. All thresholds and decision rules used here are listed explicitly.

A.1 Phoneme-level specifications

Model specification. The /y/ presence model uses logistic regression with fixed effects for language family and macro-area (dummy codes) plus centred vowel inventory size.

Cross-validation. 10-fold CV using GroupKFold by language family to prevent leakage across related languages. Evaluation metric: ROC–AUC (area under receiver operating characteristic curve).

Inventory selection sensitivity. For languages with multiple inventories, we compute the statistics under three selection rules (largest, smallest, and 10 fixed-seed random draws). Primary figures use the largest-inventory rule; sensitivity outputs are recorded in `out/summary_tables_selection.csv` and `out/y_model_sensitivity.csv`. Under the grouped CV protocol, /y/ AUC ranges roughly 0.65–0.72 across these selection rules, so the result is near-threshold but robustly above chance.

Success criteria. All have to be met: (1) positive inventory effect with 95% CI excluding zero; (2) 10-fold CV AUC ≥ 0.70; (3) Mann–Kendall trend test $p < 0.01$. Trend significance is computed with a Mann–Kendall-style statistic (normal approximation) and a permutation null (1000 permutations); (4) effect persists across three specifications: family-effects-only, family+area effects, and lineage-pruned samples (all including the inventory predictor).

A.2 Word-level specifications

Target selection. Targets are high-drift adjectives from HistWords COHA lemma SGNS (Hamilton, Leskovec & Jurafsky 2016): the top decile of average cosine displacement among non-stop, non-proper adjectives (volstats list), with documented drift terms (*nice, sick, gay,*

Table 9: Word-level drift summary (COHA lemma SGNS, 1900–2000). Cells show mean with 95% bootstrap CI. Project acc is per-lexeme word-identity accuracy on held-out decades; macro-F1 (aligned) = 0.835; macro-F1 (neighbor-jaccard) = 0.736; shuffled-label baselines mean = 0.025 (95% CI [0.000, 0.082]) and 0.023 (95% CI [0.000, 0.072]).

Group	n	Neighbor overlap	Self cosine	Project acc	Project acc (J)	Temporal
Targets	20	0.200 [0.122, 0.282]	0.535 [0.500, 0.575]	0.708 [0.575, 0.833]	0.575 [0.408, 0.725]	1.714 [1.5...
Controls	16	0.640 [0.607, 0.680]	0.776 [0.764, 0.788]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.034 [1.0...

awful) forced in when available. Minimum average frequency is 1×10^{-6} (about 1 per million). Controls are drawn from the bottom decile of displacement and matched on POS and log-frequency (±0.5, relaxed to POS-only when necessary). Some canonical drift items (e.g., *egregious*, *broadcast*) are excluded because the SGNS release supplies zero vectors.

Register metadata. The HistWords COHA lemma release provides year but no genre or register metadata in the distributed stats (freqs/volstats/pos), so register contrasts are not available for this run.

Cohesion and temporal locality. Cohesion is the mean top-50 neighbour overlap between adjacent decades. Temporal locality is the mean absolute error (in decades) between each decade and its nearest-neighbour decade in embedding space.

Projectibility. Two classifiers: (1) aligned prototype classifier trained on 1900–1940 and tested on 1950–2000; (2) alignment-free neighbour-set classifier using Jaccard similarity to training-decade neighbour unions. Metrics: macro-F1 across all lexemes and per-lexeme accuracy (targets vs controls). Shuffled-label baselines use 200 permutations for each classifier. Outputs: `out/word_drift_summary.csv`, `out/word_drift_lexemes.csv`, `out/word_drift_shuffle.csv`, `out/word_drift_jaccard_shuffle.csv`.

A.3 Construction-level specifications

Corpus palette. Construction-battery analyses use the UD English palette (GUM, EWT, GUMReddit, ESL, ESLSpok, GENTLE, CHILDES, LinES, PUD, ParTUT, UniDive, ATIS, LittlePrince, Pronouns, CTeTex, PCEDT). Register is treated as an explicit experimen-

tal factor via the three-step protocol in §5.3. We use a two-bin register split for contrast tests: **edited/formal** (PUD, PCEDT, CTeTex, LittlePrince) and **informal/spoken** (EWT, GUMReddit, CHILDES, ESLSpok). Mixed or ambiguous corpora (GUM, GENTLE, LinES, ParTUT, ESL, UniDive, ATIS, Pronouns) remain in the overall battery but are excluded from register-contrast comparisons.

Minimum evaluability thresholds. For the construction battery, PR–AUC and ablation metrics are computed only when both train and test sets contain ≥20 anchor-present candidates and ≥10 validated positive instances. When thresholds aren't met, prevalence is reported and the construction–corpus pair is marked "not estimable" rather than "projection failed". The flagship *or even* analysis falls below this bar, so its metrics are reported as illustrative and accompanied by explicit prevalence counts.

Features. (1) Anchor: binary presence; (2) Parallelism: UPOS match between contrasted heads; (3) Licensing: presence of *not*, *-n't*, *no*, *never*, *hardly*, *without*, *even* within 5 tokens left.

Model and evaluation. L2-regularized logistic regression (C=1.0). Evaluation restricted to anchor-present candidates (true *or even* vs. strings containing "or even" but lacking parallelism or explicit scalar cues; punctuation-only contexts count as positive only when the left head is adjacent to the anchor). Sample sizes in Table 1 reflect the anchor-present evaluation set.

Success criteria. (1) Cross-corpus PR–AUC around or above 0.70; (2) Ablations reduce PR–AUC, with the largest drops expected when parallelism is removed; (3) Anchor-only baseline on the anchor-present evaluation set equals prevalence. Calibration and shuffled-label baselines are reserved for the larger construction battery.

Battery robustness checks. For the construction battery we compute bootstrap CIs for full PR–AUC and ablation deltas (200 resamples of train and test with replacement) and shuffled-label baselines (200 permutations of train labels) to assess overfitting and prevalence effects. Outputs: `out/cx_battery_bootstrap.csv` and `out/cx_battery_shuffle.csv`.

Manual spot-checks. We manually inspect a stratified sample (10 positives and 10 negatives per newly estimable construction) and record judgments in `out/cx_battery_spotcheck_annotated.` summary notes in `notes/cx_battery_spotcheck.md`.

A.4 Multiple testing and inference

Primary outcomes (no correction). /y/ slope and AUC (Case A); average cohesion and F1 across target/control pairs (Case B); cross-corpus PR–AUC and mean ablation delta (Case C).

Secondary outcomes. Benjamini–Hochberg correction at $q = 0.10$ for: (1) individual family medians; (2) multiple vowel comparisons; (3) individual lexeme metrics.

Uncertainty. Bootstrap CIs (2000 resamples) for: family medians, AUC metrics, classification metrics. Permutation tests (1000 permutations) specifically for Mann–Kendall trend statistics.

A.5 Perturbation experiments

Frequency downsampling. Reduce construction tokens by 75%, 50%, 25% via stratified sampling. Recompute cue covariance (phi coefficients) and PR–AUC. Success: ≥ 0.10 drop in PR–AUC or $\geq 20\%$ reduction in parallelism/licensing rates at 25% sample.

Constructicon scaling. Bin corpora by construction type count (quartiles). Estimate P(rare variant) per bin with Wilson CIs. Success: monotonic increase with non-overlapping CIs for extreme quartiles.

A.6 Software and versions

R 4.3.1 (phoneme analyses): lme4 1.1-34, ggplot2 3.4.2, boot 1.3-28. Python 3.10.12 (word/construction): scikit-learn 1.3.0, pandas 2.0.3, numpy 1.24.3, gensim 4.3.1, stanza 1.5.0. Random seeds are fixed per script (42 unless otherwise noted; construction battery bootstrap/shuffle checks

use 17). Complete session info in repository `SESSION.txt`.

Acknowledgements

Table 10: Per-lexeme drift metrics (targets vs matched controls). Project accuracy is word-identity accuracy on held-out decades.

word	group	pos	avg_freq	drift_score	neighbor_overlap	self_cosine	project_acc	tempor
dutch	target	ADJ	2.20e-05	0.527	0.234	0.473	0.667	2.000
handy	target	ADJ	9.97e-06	0.520	0.010	0.480	1.000	1.909
mexican	target	ADJ	2.69e-05	0.513	0.258	0.487	0.500	1.636
grim	target	ADJ	1.95e-05	0.513	0.138	0.487	0.333	2.182
alien	target	ADJ	2.27e-05	0.509	0.158	0.491	0.667	2.364
underground	target	ADJ	1.47e-05	0.507	0.080	0.493	1.000	1.636
elder	target	ADJ	2.12e-05	0.506	0.164	0.494	0.667	1.727
uncertain	target	ADJ	1.73e-05	0.503	0.066	0.497	0.500	1.636
cuban	target	ADJ	1.42e-05	0.502	0.042	0.498	0.667	2.273
savage	target	ADJ	2.29e-05	0.500	0.218	0.500	0.333	1.273
bold	target	ADJ	1.94e-05	0.497	0.122	0.503	0.333	2.000
lesser	target	ADJ	1.34e-05	0.491	0.026	0.509	1.000	1.091
amazing	target	ADJ	1.66e-05	0.488	0.134	0.512	1.000	1.727
neutral	target	ADJ	1.49e-05	0.485	0.044	0.515	0.333	1.727
perpetual	target	ADJ	7.07e-06	0.485	0.036	0.515	1.000	2.545
ridiculous	target	ADJ	1.76e-05	0.485	0.110	0.515	1.000	1.818
gay	target	ADJ	2.87e-05	0.442	0.328	0.558	0.167	1.182
awful	target	ADJ	3.57e-05	0.334	0.548	0.666	1.000	1.545
sick	target	ADJ	6.74e-05	0.275	0.598	0.725	1.000	1.000
nice	target	ADJ	9.81e-05	0.209	0.686	0.791	1.000	1.000
wooden	control	ADJ	3.54e-05	0.255	0.556	0.745	1.000	1.000
ancient	control	ADJ	5.57e-05	0.254	0.518	0.746	1.000	1.091
supreme	control	ADJ	4.88e-05	0.248	0.600	0.752	1.000	1.000
willing	control	ADJ	5.81e-05	0.243	0.662	0.757	1.000	1.000
yellow	control	ADJ	6.90e-05	0.238	0.672	0.762	1.000	1.182
annual	control	ADJ	4.24e-05	0.237	0.518	0.763	1.000	1.000
available	control	ADJ	6.98e-05	0.236	0.598	0.764	1.000	1.000
loud	control	ADJ	4.04e-05	0.235	0.754	0.765	1.000	1.182
narrow	control	ADJ	6.68e-05	0.228	0.560	0.772	1.000	1.000
pale	control	ADJ	5.19e-05	0.224	0.720	0.776	1.000	1.091
sudden	control	ADJ	6.94e-05	0.218	0.688	0.782	1.000	1.000
bright	control	ADJ	7.82e-05	0.212	0.728	0.788	1.000	1.000
thick	control	ADJ	6.06e-05	0.206	0.602	0.794	1.000	1.000
tall	control	ADJ	7.80e-05	0.189	0.648	0.811	1.000	1.000
dark	control	ADJ	1.87e-04	0.183	0.732	0.817	1.000	1.000
able	control	ADJ	1.87e-04	0.178	0.682	0.822	1.000	1.000

References

Boas, Hans C. & Ivan A. Sag (eds.). 2012. *Sign-based construction grammar* (CSLI Lecture Notes 193). Stanford, CA: CSLI Publications.

Boyd, Richard N. 1988. How to be a moral realist. In Geoffrey Sayre-McCord (ed.), *Essays on moral realism*, 181–228. Ithaca, NY: Cornell University Press.

Boyd, Richard N. 1991. Realism, anti-foundationalism and the enthusiasm for natural kinds. *Philosophical Studies* 61(1). 127–148. https://doi.org/10.1007/BF00385837.

Boyd, Richard N. 1999. Homeostasis, species, and higher taxa. In Robert A. Wilson (ed.), *Species: new interdisciplinary essays*, 141–186. Cambridge, MA: MIT Press.

Boyd, Richard N. 2000. Kinds as the "workmanship of men": Realism, constructivism, and natural kinds. In Julian Nida-Rümelin (ed.), *Rationalität, realismus, revision / rationality, realism, revision: proceedings of the 3rd international congress of the gesellschaft für analytische philosophie*, 52–89. Berlin: De Gruyter. https://doi.org/10.1515/9783110805703.52.

Bybee, Joan L. & Paul J. Hopper. 2001. Introduction to frequency and the emergence of linguistic structure. In Joan L. Bybee & Paul J. Hopper (eds.), *Frequency and the emergence of linguistic structure* (Typological Studies in Language 45), 1–24. Amsterdam: John Benjamins. https://doi.org/10.1075/tsl.45.01byb.

Croft, William. 2001. *Radical construction grammar: syntactic theory in typological perspective*. Oxford: Oxford University Press.

Ekström, Axel G., Claudio Tennie, Steven Moran & Caleb Everett. 2025. The phoneme as a cognitive tool. *Topics in Cognitive Science*. Advance online publication. https://doi.org/10.1111/tops.70021.

Fillmore, Charles J., Paul Kay & Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language* 64(3). 501–538.

Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure* (Cognitive Theory of Language and Culture). Chicago: University of Chicago Press.

Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language.* Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199268511.001.0001.

Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th annual meeting of the association for computational linguistics*, 1489–1501. Berlin. https://doi.org/10.18653/v1/P16-1141.

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

Heyes, Cecilia. 2018. Précis of Cognitive Gadgets: The Cultural Evolution of Thinking. *Behavioral and Brain Sciences* 42. https://doi.org/10.1017/S0140525X18002145.

Khalidi, Muhammad Ali. 2013. *Natural categories and human kinds: Classification in the natural and social sciences.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139207071.

Kuhl, Patricia K., Karen A. Williams, Francisco Lacerda, Kenneth N. Stevens & Björn Lindblom. 1991. Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics* 50(2). 93–107. https://doi.org/10.3758/BF03212211.

Liljencrants, Johan & Björn Lindblom. 1972. Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language* 48(4). 839–862. https://doi.org/10.2307/411991.

Lindblom, Björn. 1990. Explaining phonetic variation: A sketch of the H&H theory. In William J. Hardcastle & Alain Marchal (eds.), *Speech production and speech modelling,*

403–439. Dordrecht: Kluwer Academic Publishers. https://doi.org/10.1007/978-94-009-2037-8_16.

Manning, Benjamin S. & John J. Horton. 2025. General social agents. Most recent draft: 2025-09-03. Manuscript, MIT; MIT & NBER.

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

Moran, Steven, Daniel McCloy & Richard Wright (eds.). 2019. *PHOIBLE 2.0.* https://phoible.org. Accessed 2025-09-01. Jena.

Pollard, Carl & Ivan A. Sag. 1994. *Head-driven phrase structure grammar.* Stanford, CA: CSLI Publications.

Rubin, Michael. 2008. Is goodness a homeostatic property cluster? *Ethics* 118(3). 496–528.

Silveira, Natalia, Timothy Dozat, Marie-Catherine de Marneffe, Samuel R. Bowman, Miriam Connor, John Bauer & Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the ninth international conference on language resources and evaluation (lrec'14).* Reykjavik.

Stevens, Kenneth N. 1989. On the quantal nature of speech. *Journal of Phonetics* 17. 3–45.

Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition.* Cambridge, MA: Harvard University Press.

Zeldes, Amir. 2017. The GUM corpus: Creating multilayer resources in the classroom. In *Proceedings of the 2017 conference on empirical methods in natural language processing: system demonstrations.* Copenhagen.