How to Study Boundary Phenomena: English Reciprocals and the Limits of Categorization

Brett Reynolds

Humber Polytechnic and University of Toronto

brett.reynolds@humber.ca

December 12, 2025

Abstract

English reciprocals (*each other* and *one another*) exemplify a fundamental challenge in linguistic categorization: how do we rigorously investigate boundary phenomena when our entire dataset consists of just two items that straddle categorical divisions? Traditional approaches risk methodological opportunism by selecting diagnostic tests that support preferred analyses. This paper proposes measuring the *stability of diagnostic ambiguity* rather than forcing binary decisions. Using a comprehensive profile of 155 morphological, syntactic, semantic, and phonological properties, I locate the reciprocals in multidimensional grammatical space and apply five converging analytical lenses: distance measurement with permutation testing, specification curve analysis across different metrics and feature sets, comparison group robustness checks, mixture calibration, and calibration against known categories. Results consistently show reciprocals occupying a boundary position between canonical pronouns and compound determinatives, with near-parity mixture weights that place them at the midpoint between anchor profiles – a location that remains stable across analytical choices. This stability of diagnostic ambiguity, rather than decisive categorization, aligns with view-

ing grammatical categories as homeostatic property clusters with genuine boundaries. The methodology provides a template for transparent investigation of small-n categorization problems, demonstrating how boundary phenomena can be studied rigorously without forcing artificial binary decisions. While reciprocals remain categorized as pronouns in standard frameworks, they occupy the peripheral boundary region alongside morphologically complex items that realize fused grammatical functions.

1 Introduction

When linguists categorize boundary phenomena – items that straddle categorical divisions – we face a compound problem. English reciprocals exemplify this challenge: not only are there just two (*each other* and *one another*), but they also exhibit properties of both pronouns and compound determinatives. How can we rigorously investigate categorical ambiguity when our entire dataset consists of two boundary-dwelling items? Small samples of clear cases pose little difficulty – English has few words with the NICER properties (Sag et al. 2020), but their status as verbs is uncontroversial. The challenge emerges when scarce data meets genuine categorial ambiguity, precisely the items linguists argue about.[1]

The traditional solution is to pick diagnostic tests and argue about them. Can it be a subject? Does it inflect for case? But this invites what Croft (2001) calls "methodological opportunism" – consciously or not, we select tests that support our preferred analysis. When we have only two items to classify, this cherry-picking becomes especially tempting.

This paper proposes a different approach: instead of seeking a decisive verdict, I measure the *stability of diagnostic ambiguity*. The interesting question isn't "which category?" but "how stable is the apparent boundary position under different measurement choices?" If grammatical categories are homeostatic property clusters (Miller 2021) whose observable behaviour is boundary-like rather than all-or-nothing, then some items should consistently

[1] Such as the categorial status of *to* (Levine 2012).

appear ambiguous no matter how we analyze them.[2]

On this homeostatic property-cluster (HPC) view, categories are stabilized by partially distinct mechanisms that sustain overlapping bundles of properties. This yields concrete, testable predictions – not arbitrary methodological desiderata but consequences of the theoretical framework itself: (E1) a boundary item should show invariance of its boundary position across reasonable analytic choices (distance metric, regularization, feature ablations); (E2) different feature families may exert opposed pulls if distinct mechanisms sustain partly conflicting bundles (cross-dimensional tension); (E3) clear anchors should calibrate cleanly under the same lenses; (E4) a one-parameter blend fit to anchor profiles should place boundary items near parity; (E5) boundary diagnoses should persist under nulls that preserve the instrument's marginals. The analyses below are organized to test these predictions directly.

I demonstrate this approach using English reciprocals as a test case. Previous work ("Quantifying the Differences Between Lexical Categories: The Case of Pronouns and Determinatives in English" 2021) has already noted their uncertain status, sitting between canonical pronouns and compound determinatives (Huddleston & Pullum 2002: henceforth *CGEL*, Ch. 5 §9.6): items like *somebody* and *anybody* that realize fused determiner–head functions (Payne, Huddleston & Pullum 2007). The fusion-of-functions architecture predicts that morphologically complex items can belong to the determinative category while serving hybrid syntactic roles, making reciprocals a perfect probe for boundary behaviour.

The approach has five components, each addressing a specific challenge in small-n categorization. First, I use a comprehensive feature profile: 155 binary properties covering morphology, syntax, semantics, and phonology and apply this to 138 words that form the universe of pronouns and determinatives as defined by *CGEL*. This makes theoretical commitments explicit rather than hiding them across cherry-picked diagnostics. It's like mapping

[2]For word-kinds, Miller (2021) distinguishes internal cognitive and external social mechanisms. In applying this template to grammatical categories, I treat the feature matrix as an operationalization of the cluster and take the relevant mechanisms to include morphological realization rules, agreement/case systems, entrenched distributional patterns, grammaticalization pathways, and community norms.

vowel space: instead of checking one or two formants, I measure everything I can.

Second, I visualize the high-dimensional feature space to build intuition about reciprocals' position. Figure 1 uses Multiple Correspondence Analysis (MCA) (Greenacre 2017), an ordination technique that arranges 155-dimensional data in 2 dimensions while preserving distances between items, much like reducing vowel measurements to an F1–F2 plot. MCA naturally handles sparse binary data through χ^2 geometry.[3] But this is purely illustrative. For actual measurement, I use Jaccard distance for reasons explained below. All inferential contrasts operate on these full-dimensional Jaccard distances (with other variants as checks), never on the low-dimensional projections.

Third, I test whether patterns could arise by chance. With only two reciprocals, any pattern might be noise. So I scramble the data 5,000 times using the quasiswap algorithm (Miklós & Podani 2004) while preserving its basic structure, asking, "how often would random scrambling produce patterns this strong?" The results of this indicates whether the reciprocals' boundary position is meaningful or accidental.

By this point, I've already made several analytical choices: Jaccard distance, specific reference items, a particular test statistic. Each choice is a fork in the path. So fourth, I map what Gelman & Loken (2013) call the "garden of forking paths" by varying every reasonable alternative and showing all results. Different distance metrics, different comparison groups, different feature sets – I try them all and display the full spectrum. If the conclusion depends heavily on arbitrary choices, it should be immediately apparent.

Fifth, I calibrate against known structure from both directions. Bottom-up: can my methods correctly recover the categories from the *CGEL* framework? Top-down: what category mixture would generate the reciprocal pattern we observe? If the machinery can't identify clear pronouns as pronouns, the whole enterprise fails.

Two limitations constrain this entire analysis. First, everything depends on the hand-

[3]MCA's χ^2 distance weights deviations from independence: rare shared features contribute more to distance calculations than common absences, making it appropriate for grammatical matrices where most features are absent for most words.

coded feature matrix – if the instrument is flawed, so are the conclusions. Second, with $n = 2$, statistical power is inherently limited. I address these constraints through transparency and multiple robustness checks, but I can't eliminate them.

The obvious objection is that elaborate statistics with only two reciprocals is futile. A phonetic calibration analogy dissolves that worry. Once a language's vowel space is mapped, a single token can be located relative to established categories without collecting more tokens of that type. A lone production near the /i/–/ɪ/ boundary can be characterized as boundary-dwelling – but the token *is* either /i/ or /ɪ/; there's a fact of the matter. Our acoustic measurements simply can't resolve which. A lone instance of the front rounded vowel [y] in a loan name can be placed outside English's native inventory by its F1–F2–F3 coordinates. The 155-feature matrix plays the same role here: it's the calibrated space, with the same resolution limits. The analyses below don't try to show that reciprocals constitute their own category (which $n = 2$ cannot support); they locate them in the calibrated space defined by clear *pronoun* and *determinative* anchors, ask by permutation whether the observed configuration is unusual under the matrix's marginals, and use mixture calibration and specification curves to check whether that location is stable across analytic choices.

The payoff isn't a definitive recategorization – it's a template for investigating boundary phenomena honestly. When we can't increase our sample size (English won't grow more reciprocals any time soon), we can at least be transparent about what the available evidence actually shows. The methods are deliberately portable: other researchers can apply this same toolkit to their own small-n categorization problems with minimal adaptation.

2 The Challenge of Measuring Grammatical Similarity

When we ask whether reciprocals are pronouns or determinatives, we're really asking how to measure grammatical similarity. Traditional approaches pick a handful of diagnostic tests and declare success when items pattern together. That invites methodological opportunism.

To avoid this trap, I use the calibrated instrument from the introduction: a 155-feature feature matrix that treats each word form as a point in a high-dimensional space whose axes are grammatical properties.

The features come from "Quantifying the Differences Between Lexical Categories: The Case of Pronouns and Determinatives in English" (2021), who coded 232 binary features across 138 word forms. I keep only the 155 that apply to more than one word form – this filters out idiosyncratic properties unique to individual items, leaving features that actually help compare categories. They span morphology (66), syntax (50), semantics (36), and phonology (3). With binary coding, every word form becomes a 155-bit vector.

Again, the vowel analogy is useful. Once you've mapped a language's vowel space, you can place any new token relative to the established categories. Here, the 155 features are the acoustic dimensions; canonical pronouns and compound determinatives are the reference categories; the 2D ordination is like an F1–F2 plot for visual orientation; and Jaccard distance stands in for perceptual distance.

I use Jaccard because in sparse binary data, shared absences tell you almost nothing – the fact that neither *each other* nor *the* can be plural doesn't reveal much about their relationship. What matters are shared presences and conflicts. Jaccard ignores the uninformative zeros and focuses on what counts. The MCA plot in Figure 1 is just for intuition; all the actual contrasts use full 155-dimensional distances.

For comparison, I use two predeclared sets. First, canonical pronouns like *he*, *her*, and *herself*. Second, the compound determinatives from Huddleston & Pullum (2002: Ch. 5 §9.6) – items like *anybody* and *everyone* that realize fused determiner-head functions (Payne, Huddleston & Pullum 2007). If morphologically complex reciprocals pattern with anything, it's probably with these hybrids rather than with simple pronouns.

From here the pipeline is straightforward: plot for orientation, compute distances, shuffle to test significance, vary specifications to check robustness, and calculate mixture weights to summarize position. Each step tackles a specific challenge while keeping the choices

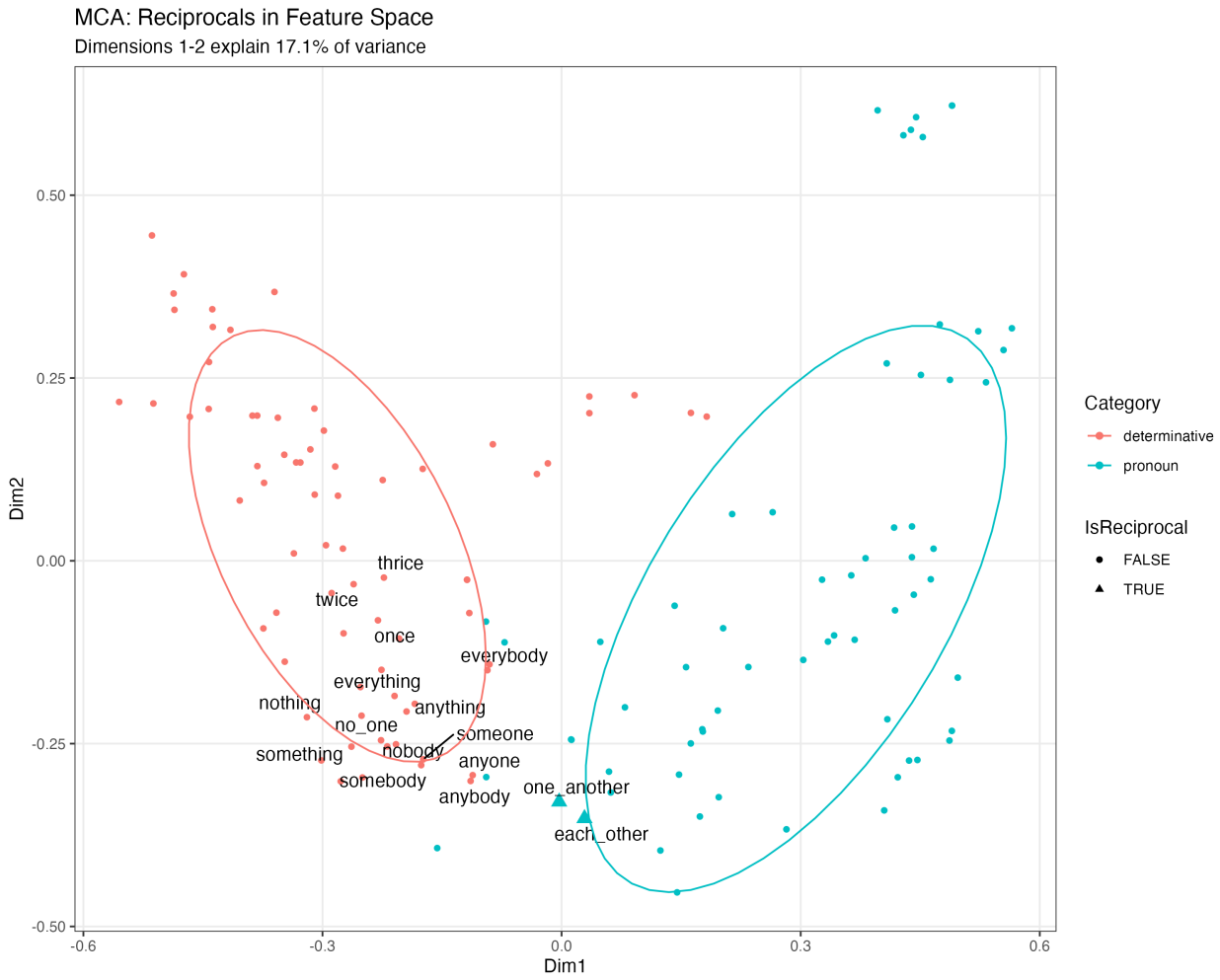


Figure 1: Reciprocals in grammatical feature space visualized with Multiple Correspondence Analysis. Pronouns (cyan) and determinatives (red) form regions; the compound determinatives sit at the interface; *each other* and *one another* (triangles) fall in that same interface.

transparent.

3 Testing Whether Patterns Could Arise by Chance

Having established how to measure distances, I faced a statistical challenge: with only two reciprocals in English, any apparent pattern might simply be noise. I addressed this through permutation testing, specifically using a technique called "quasiswap". The logic is this: each word in the dataset has a certain number of properties (e.g., *each other* has 23), and each property applies to a certain number of words (e.g., `inflects for case` applies

to 75). These totals reflect real facts about English – that pronouns tend to have case distinctions, that most words can't be gradable, and so on. But what if we kept these basic facts while scrambling which specific words have which specific properties while holding the totals constant? If the reciprocals' proximity to determinatives is meaningful, it should disappear under such scrambling. If it persists, we might just be seeing an artifact of how common or rare certain features happen to be.

The quasiswap algorithm (Miklós & Podani 2004) performs exactly this controlled scrambling. It rearranges the feature matrix through thousands of small swaps, always preserving how many features each word has and how many words have each feature. So, after 5,000 such rearrangements, how often do scrambled datasets produce distance patterns as extreme as what we actually observe? This tests whether the specific combination of features in reciprocals – not just their total number – drives their intermediate position.

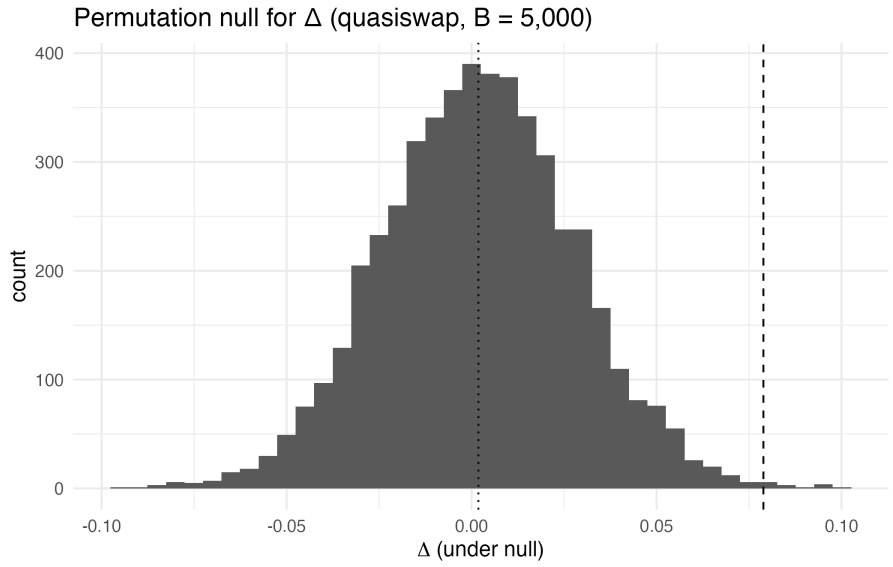


Figure 2: Testing whether the reciprocals' intermediate position could arise by chance. The histogram shows distance patterns from 5,000 scrambled versions of the data; the dashed vertical line marks the position of the unscrambled observation. The pattern appears in only 0.6% of scrambled datasets, suggesting it's not a statistical fluke.

The results are shown in Figure 2. For my primary comparison – using carefully matched sets of six pronouns and six compound determinatives – the observed pattern appeared

in only 0.6% of scrambled datasets ($p = 0.006$). This suggests the reciprocals' intermediate position isn't merely an artifact of which properties happen to be common or rare in English; there's something systematic about their specific combination of features.

But two caveats apply. First, I'll reiterate that, with only two reciprocals, any statistical test has limited power. Second, my choice of which six pronouns and six determinatives to compare matters. Different selections might yield different results, a sensitivity I explore in Section 5. Despite these limitations, the permutation test provides initial evidence that reciprocals' boundary position reflects genuine grammatical structure rather than statistical noise.[4]

4 The Garden of Forking Paths

As I point out in the introduction, I've made several analytic choices: Jaccard as the primary distance, specific reference items, a particular test statistic. Each choice is a fork in the path and an opportunity for bias to accumulate. Exploring only one path risks over-interpreting a pattern that might disappear under equally reasonable alternatives.

Specification-curve analysis (Simonsohn, Simmons & Nelson 2020) addresses this directly. Instead of defending a single pipeline, I run many plausible pipelines and show all results. Here, I vary the distance metric (Jaccard, Dice, and an IDF-weighted Jaccard), the feature set (all features, or blockwise ablations such as "no morphology", "no syntax", etc.), and the weighting method. The plot labels this "alpha" – a technical term for regularization. The "ridge" method keeps all features weighted; "elastic05" allows the algorithm to exclude uninformative features. Each combination yields a value of Δ, where $\Delta = \bar{d}(\text{reciprocals}, \text{compound determinatives}) - \bar{d}(\text{reciprocals}, \text{pronouns})$, the mean distance from the reciprocals to compound determinatives minus the mean distance from the reciprocals to pronouns.

Figure 3 shows two clear patterns. First, the sign of Δ is stable across distance metrics and

[4]Implementation in `03_matched_set_robustness.R`; null draws saved as CSV/RDS for reproducibility.

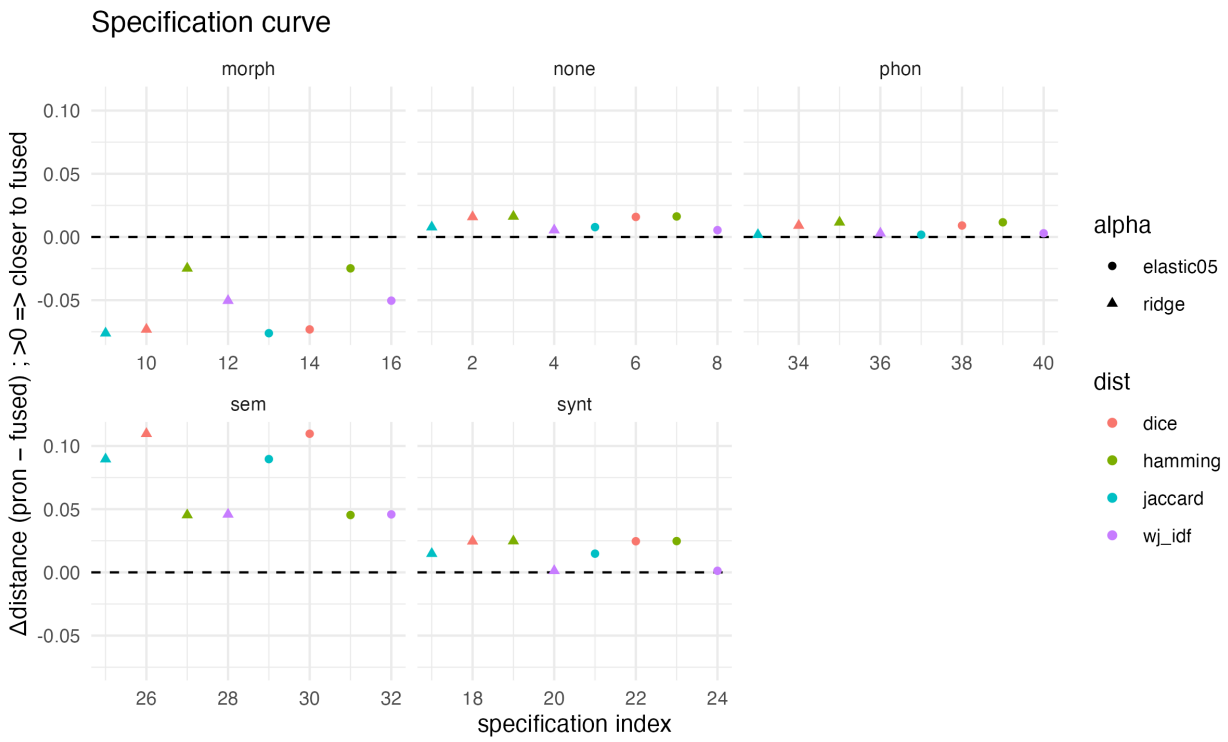


Figure 3: Specification curve for Δ. Each point shows the observed contrast under different combinations of weighting method (alpha for shape), distance metric (dist for colour), and feature block exclusion (panels). Positive values indicate the reciprocals are closer to compound determinatives; negative values indicate they're closer to pronouns; zero marks the boundary. The dashed line is the pronoun–determinative dividing line.

weighting methods: most specifications land on the same side of zero. Second, only one block flips the sign – when morphology is *excluded* (panel "morph"), $\Delta < 0$ and the reciprocals look more pronoun-like. With all features included (panel *none*) $\Delta > 0$, and removing semantics pushes Δ further positive (more determinative-like). Removing syntax or phonology leaves small, positive contrasts near the baseline. This pattern suggests that morphological features carry most of the determinative-ward signal, whereas semantic features contribute a pronoun-ward pull; syntax and phonology are comparatively weak in this contrast.

This invariance has theoretical significance. If category boundaries are sharp but located at epistemically inaccessible thresholds, we should expect the *existence* of a boundary to be robust across reasonable metrics even when its precise location cannot be pinpointed. The specification curve tests exactly this: Δ may shift in magnitude across specifications, but its

sign – indicating which side of the boundary reciprocals fall on – remains stable across most of them. That's the empirical signature of a real boundary whose exact location eludes finite measurement.

What the specification curve does *not* address is design sensitivity from the particular 6+6 matched sets used in the permutation contrast; that is investigated separately by rotating the matched sets in Section 5.

5 The Importance of Comparison Groups

I had chosen specific pronouns and determinatives for comparison based on theoretical considerations, but different choices may yield different contrasts. To assess design sensitivity, I repeated the analysis 100 times, each time randomly selecting six pronouns and six compound determinatives from the eligible pools while holding all other steps fixed.

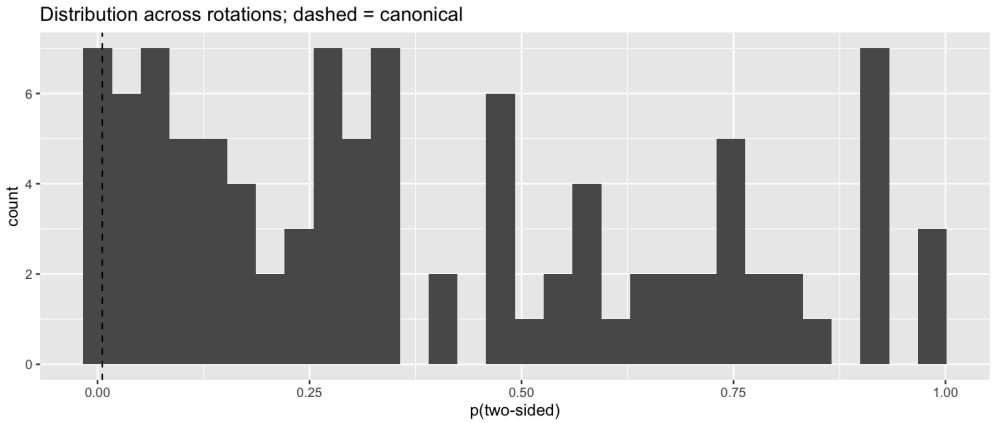


Figure 4: Sensitivity of Δ to the choice of comparison groups. The histogram shows permutation tail areas for Δ across 100 rotations of six pronoun and six compound determinative anchors. The dashed line marks the prespecified comparison set (Δ tail area = 0.006). Most rotations yield larger tail areas (median = 0.309); only 13% fall below 0.05.

Figure 4 shows wide variation across rotations.[5] The prespecified set yields a permutation tail area of 0.006; across rotations the median tail area is 0.309, and 13 out of 100 runs fall below 0.05. I treat 0.05 as a reference line rather than a decision rule. With such small n,

[5] Implementation: `03_matched_set_robustness.R`; the specific canonical items are listed in `matched_subset_manifest.txt`.

significant results are especially vulnerable to Type S (sign) and Type M (magnitude) errors (Gelman & Carlin 2014), so the more informative approach is to summarise the distribution of Δ across rotations rather than to spotlight a single tail area. The substantive point is that Δ is design-sensitive to the choice of comparators: the prespecified set lands near the lower tail (rank 3/100), but many equally reasonable comparator sets don't produce extreme tail areas. Accordingly, I focus on the effect itself – the distance contrast Δ – and on how Δ varies across rotations, using the rotations as a descriptive multiverse (Steegen et al. 2016) that quantifies uncertainty due to researcher degrees of freedom rather than as a filter for statistical significance.

6 Simulating the Boundary

To build intuition about what these statistical patterns mean grammatically, I constructed a simple generative model. Think of grammatical categories as recipes: pronouns combine ingredients in certain proportions, determinatives in others. A boundary item would use a mixed recipe – perhaps 70% pronoun ingredients, 30% determinative. The two category profiles are estimated from the anchor items with the reciprocals held out; the exercise asks how much of each profile best predicts a reciprocal's observed properties.[6]

The simulator implements this idea with a mixture weight w that ranges from 0 to 1. When $w = 1$, the simulator generates features like a typical pronoun profile; when $w = 0$, like a typical determinative profile. For values in between, it mixes the two patterns – at $w = 0.7$, for instance, it's 70% pronoun-like and 30% determinative-like. Read w as a predictive index of boundary position, not as a literal claim about composition.

I then asked: what mixture weight would make the simulator best match the reciprocals' observed feature patterns under the same scoring rule? This is like asking: if reciprocals are intermediate between the two anchor profiles, what blend best predicts their properties?

[6]Concretely, the category profiles are fit on the anchors only; for a given weight w the blend's predictive score for a reciprocal's 155-bit feature vector is computed under a proper scoring rule (log loss), and $\hat{w}$ is

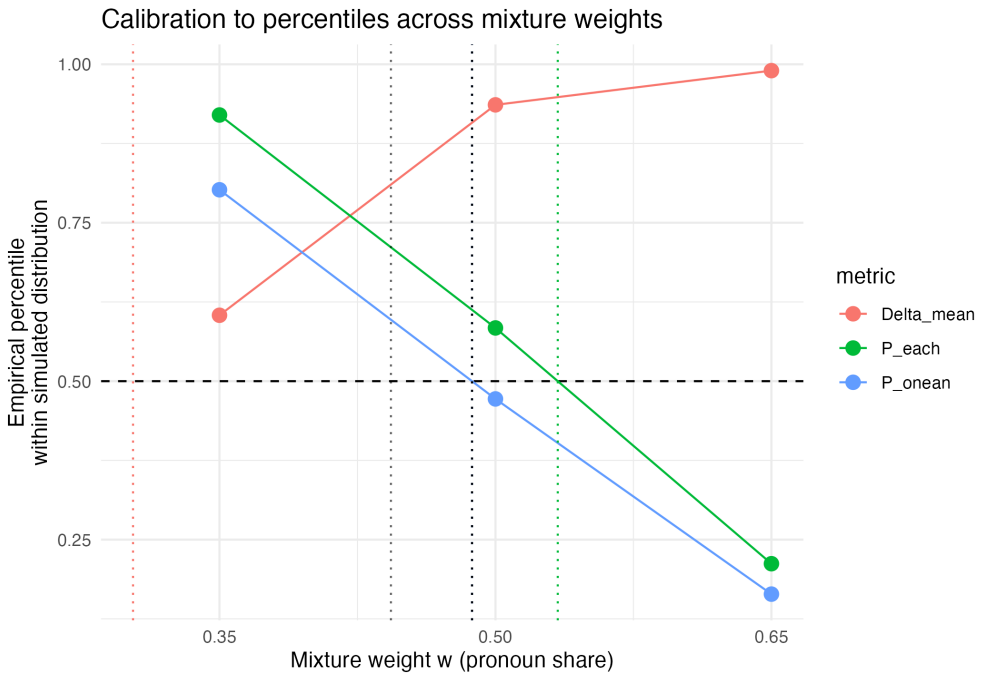


Figure 5: Predictive blending of pronoun and compound determinative profiles. Curves show predictive fit as the mixture weight w varies; dashed lines mark the best-fitting weights for each reciprocal. Both are near the midpoint: $\hat{w} = 0.534$ for *each other*, $\hat{w} = 0.487$ for *one another*.

Figure 5 reveals the answer: both reciprocals sit almost exactly at the midpoint. To match the properties of *each other*, the best blend is about 0.53 pronoun-like; for *one another*, about 0.49 pronoun-like.

This provides another lens on the same phenomenon. Whether we visualize grammatical space, measure distances, run permutation contrasts, or calibrate this simple blend, reciprocals consistently appear at the boundary – not 0.9 pronoun or 0.8 determinative, but located at the midpoint between the two anchor profiles.

Two caveats apply. First, reciprocals aren't literally probability mixtures of two kinds; the simulator provides an interpretable, one-number summary of where they sit relative to the anchor profiles. Second, reasonable ways of scoring the match, or modest feature ablations, shift the weights slightly but keep both reciprocals near 0.5, in line with the earlier specification-curve pattern.

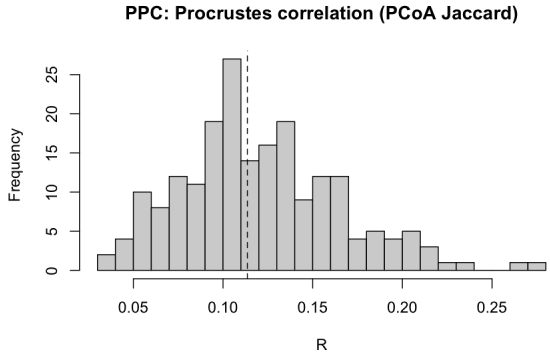
the maximizer. Implementation in `04_weight_calibration.R`; calibration data in the repository.

7 Checking the Instruments

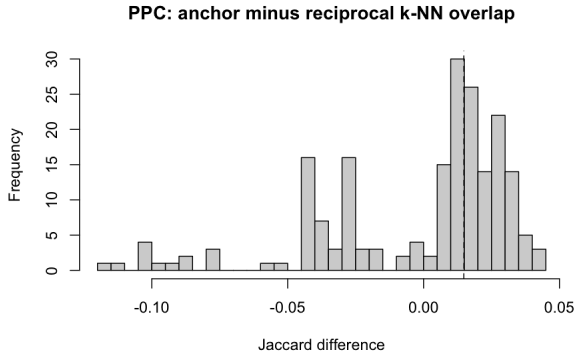
Throughout these analyses, I have treated the feature matrix as a measurement model – but what if the measurements themselves are problematic? I first ask whether the statistical machinery can recover known structure. This is like checking a scale with a kilogram weight and a feather before weighing something ambiguous.

I fit a simple model that predicts each binary feature from the item's category label with modest adjustment terms and weak regularization, and then draw replicated datasets from the fitted model. If the category signal is real, the replicates should partly reproduce the observed structure without overfitting.[7]

I then ask two questions. Global geometry: do replicated datasets preserve the large-scale arrangement (which items cluster with which) under the same ordination? Local neighbourhoods: do they preserve nearest-neighbour relations around anchors?



(a) Predictive check for global structure. Histogram: distribution of the Procrustes statistic under replicated datasets; vertical line: observed statistic.



(b) Predictive check for local structure. Histogram: distribution of k-NN overlap around anchors (k fixed across runs); vertical line: observed overlap.

Figure 6: Predictive checks. (a) Global structure: distribution of Procrustes fit statistics across replicated datasets (histogram) with the observed statistic marked (vertical line). (b) Local structure: distribution of nearest-neighbour overlaps under replication, with the observed overlap marked.

[7] Concretely, for feature j on item i, $y_{ij} \sim \text{Bernoulli}(\pi_{ij})$ with $\text{logit}\,\pi_{ij} = \alpha_j + \beta_j \mathbf{1}\{c_i = \text{pronoun}\}$ and ridge regularization on β_j; draws are parametric bootstrap replicates (posterior draws if a Bayesian fit's used). Reciprocals are held out wherever predictions for them are reported.

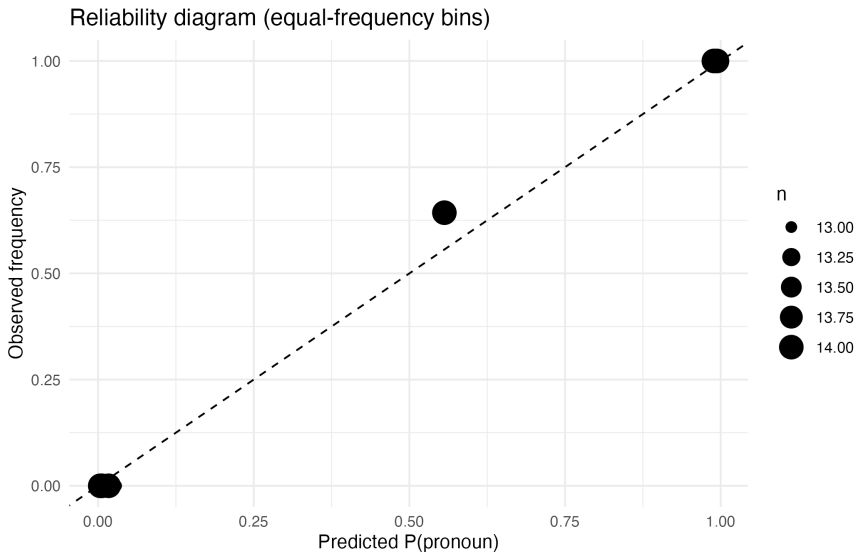


Figure 7: Reliability diagram under cross-validation, following Niculescu-Mizil & Caruana (2005). Predicted probabilities (x-axis) track observed frequencies (y-axis), showing approximate calibration: when the model predicts 70% probability for the pronoun label, about 70% of held-out items are pronouns.

The model passes these basic predictive checks (Figures 6a–7). Replicates maintain reasonable global geometry and preserve about 65.5% of local neighbourhood structure around anchors. The reliability diagram shows that when the classifier assigns, say, 70% probability to the pronoun label, the realised frequency on held-out items is near that level – probabilities are roughly calibrated rather than systematically over- or under-confident.[8]

Perfect reproduction is neither expected nor desired. The checks indicate that the measurement model carries real signal about categorical structure: clear pronouns and clear compound determinatives are distinguishable under the same machinery, which makes the reciprocals' ambiguity informative rather than an artefact of a broken instrument.

[8]For the reciprocals, the classifier assigns near-chance pronoun probabilities (0.485 for *each other*, 0.467 for *one another*); masked predictive log-likelihood is identical under either labelling (-54.240), and per-item cross-entropy is 0.371 versus a chance baseline $\ln 2 \approx 0.693$.

8 What We Learn About Grammatical Boundaries

Taken together, the results instantiate the HPC-derived expectations. (E1) holds: the sign of the distance contrast is stable across metrics and penalties and is only modest in magnitude; permutation extremity is design-sensitive but the qualitative boundary diagnosis persists across rotations. (E2) holds: morphology contributes most of the determinative-ward pull while semantics contributes a pronoun-ward pull, with syntax and phonology comparatively weak in this contrast. (E3) holds: the same machinery separates anchors and yields calibrated supervised probabilities for them. (E4) holds: the predictive blend places both reciprocals near parity. (E5) holds: the quasiswap null that preserves row and column totals rarely reproduces the observed pattern on the prespecified anchors. Under an HPC reading, this pattern – stable ambiguity with cross-dimensional tension and clean anchor behaviour – is the characteristic signature of a genuine boundary rather than a failure to decide.

This pattern matches the homeostatic property-cluster view of categories. The reciprocals exhibit stable ambiguity with cross-dimensional tension rather than a decisive recategorization or a failure of measurement. Morphology supplies most of the determinative-ward pull; semantics supplies a smaller pronoun-ward pull; syntax and phonology contribute little to this particular contrast. Within this instrument, additional statistical machinery is unlikely to yield a clean assignment because the conflict is real rather than an artefact of a specific pipeline; changing the conclusion would require different measurements or external evidence, not a different threshold.

The methodological lesson is general. Boundary items should be expected – and tested – to show stable ambiguity under reasonable analytic choices. Single diagnostics or cherry-picked features force binary decisions; comprehensive profiles, prespecified lenses, and summaries that propagate design uncertainty instead map the topology of grammatical space, including its boundary regions, and make clear what is robust within the chosen measurement model.

9 Discussion

On an HPC view, success at a boundary isn't a decisive recategorization but a stable pattern of ambiguity accompanied by cross-dimensional conflict. That is the profile reciprocals exhibit here. Across lenses that make different invariance assumptions yet interrogate the same measurements, reciprocals sit at the edge of the pronoun region in ordinations, show mixed nearest neighbours, attract near-chance supervised probabilities with tied predictive fit (0.485, 0.467; log-likelihood −54.240 under either labelling), and display a blockwise split with morphology leaning determinative-ward,[9] semantics leaning pronoun-ward, and syntax/phonology contributing little.

Read through the HPC lens, the mixture calibration reproduces decisive separation for anchors while placing reciprocals at the boundary with near-parity placement, consistent with the ordination, distance, permutation, and supervised results. Because all lenses interrogate the same measurements, agreement is internal robustness rather than independent convergence.

The blockwise pattern has a natural mechanistic interpretation. On an HPC view, different homeostatic mechanisms sustain different portions of the property cluster. Morphological realization rules and agreement systems – mechanisms that maintain the determinative basin – account for the morphology-ward pull. Interpretive mechanisms governing reference tracking, binding, and argument structure – mechanisms that maintain the pronoun basin – account for the semantics-ward pull. The cross-dimensional tension is not a quirk of the feature matrix; it reflects genuinely distinct causal pathways exerting opposed pressures on the reciprocals' categorical location.

The findings are consistent with multiple theoretical frameworks that predict stable ambiguity at category boundaries. What distinguishes the HPC interpretation is the cross-dimensional tension: different feature families pulling in different directions.

The diagnostic ambiguity documented here is epistemic, not ontological. On the view

[9] This likely reflects pronouns' tendency to inflect for case, which reciprocals resist.

developed in this paper, reciprocals are pronouns or they are determinatives – there's a fact of the matter. The boundary between the two categories is sharp, located at some threshold in the feature space we can't finitely specify. What our instruments reveal is proximity to that boundary: near-chance classifier probabilities, near-parity mixture weights, and cross-dimensional tension are the signatures of an item sitting close to a sharp edge, not evidence that no edge exists. The situation is analogous to a microscope at its resolution limit: two points either overlap or they don't, but below the diffraction threshold our optics can't tell which. The 155-feature matrix has an analogous resolution floor – and reciprocals fall within it.

If categories were merely fuzzy – lacking any underlying sharp boundary – we'd expect uniform gradience across dimensions rather than systematic opposed pulls. If boundaries were sharp but arbitrarily located, we'd expect greater sensitivity to metric choice than the specification curve reveals. The observed pattern – stable boundary position combined with systematic cross-dimensional conflict – is the characteristic signature of homeostatic property clusters: distinct mechanisms maintaining partially overlapping property bundles, with boundary items caught where the bundles diverge.

This interpretation also clarifies what the results don't show. The analyses don't accumulate independent evidential weight; all procedures reuse the same hand-coded matrix. Agreement across lenses therefore counts as internal robustness against method-specific artifacts rather than as independent convergence. Nor do the numbers license recategorization. On the predeclared matched set the permutation contrast is extreme ($p = 0.006$), but across rotated comparison sets it centres on non-significance (median $p = 0.309$, $13\% < 0.05$). Under an HPC reading, those very features – weak directionality, non-decisive tests, and stability of the qualitative diagnosis across reasonable specifications – are precisely what a genuine boundary is expected to look like given the target population of two types.

The fusion-of-functions architecture provides a complementary grammatical backdrop (Payne, Huddleston & Pullum 2007). If category and function are distinct primitives and

fused determiner–head structures are grammatically real, then morphologically complex items can be determinatives by category while serving fused functions in NP. That architecture predicts pressure toward categorical edges when different mechanisms sustain partially conflicting property bundles. The present measurements are consistent with such pressure for the reciprocals: they behave pronoun-like on some feature classes and determinative-like on others. At the same time, retaining the canonical *CGEL* pronoun analysis as the prior default is methodologically appropriate; the data here don't overturn it, and the framework shown here makes explicit what kinds of evidence would do so.

Seen more broadly, the study's contribution is methodological and interpretive rather than classificatory. It operationalizes a way to examine boundary phenomena that avoids selective diagnostic shopping (Croft 2001): prespecify a small set of sensible lenses; interpret agreement as internal robustness on one dataset; make prior theoretical commitments explicit; and report the full specification surface rather than only favourable summaries. Inconclusive tests cease to be mere failures when they co-occur with cross-dimensional conflict and a stable ambiguity pattern; they become data about the structure of categories if those categories are homeostatic clusters (Miller 2021).

9.1 Limitations, scope conditions, and disconfirmation

All findings are conditional on a single, theory-informed, hand-coded binary measurement model. Features are binarised, single-coder, and *CGEL*-dependent; this raises construct-validity and reliability concerns. I partly offset these with blockwise analyses and metric choice motivated by ordination, but stronger guarantees would require multi-annotator coding with reliability estimates or corpus-derived distributional features that reduce theory-dependence. Target scarcity constrains inferential resolution: English has two reciprocal types, so any null-hypothesis test contrasting reciprocals with comparators has low power, and supervised categorization produces only two probabilities of primary interest. Multiplicity and researcher degrees of freedom are addressed by prespecifying a small, theory-led set of

lenses and reporting all analyses; nonetheless, specification choices remain scope conditions on the results. External validity is restricted to English and to the *CGEL*-based inventory; cross-linguistic generalization is a matter for new data.

The framework also supplies clear disconfirmation criteria for this case. Outcomes that would count against the boundary diagnosis include: alignment of morphology, syntax, semantics, and phonology on the same pole; calibrated supervised probabilities near 1.0 under one labelling together with superior predictive fit; disappearance of ambiguity under reasonable distance metrics or regularization choices; and stability of those outcomes under feature-subset resampling. None of these criteria is met here.

9.2 Implications and outlook

The present case is intended to illustrate how to study boundary phenomena responsibly when categories are treated as homeostatic clusters. The objective isn't to force a verdict on reciprocals but to refine the diagnostics that register boundary behaviour on the very measurement model already in use.

A first priority is to characterize dependence rather than to "repair" individual codings in a large, noisy matrix. What matters is which parts of the measurement model underwrite the boundary diagnosis. Influence diagnostics for distance-based summaries (feature leverage on Jaccard neighbourhoods), blockwise ablations, and adversarial recoding of plausible high-leverage features jointly map an internal robustness surface: which modest, theory-motivated perturbations do and don't alter the qualitative outcome. This keeps the focus on stability of ambiguity rather than on accumulating additional evidence.

A second priority is calibration against control items. Clear pronouns and clear compound determinatives should behave as anchors under the same lenses: high-confidence, aligned outcomes for anchors and ambiguity for reciprocals. Reporting anchor performance turns robustness from a general claim into a concrete diagnostic. If ambiguity appears for anchors, the pipeline is uninformative; if it localizes to reciprocals, the boundary reading

gains credibility without appealing to data beyond the current inventory.

A third priority is to use token-level evidence, where available, as a diagnostic of boundary dispersion rather than as a route to a type-level decision. Treat the lexeme as a distribution over constructions and modification profiles, and compare simple, predeclared dispersion summaries (e.g. entropy over syntactic frames; divergence from anchor frame distributions) to the anchors. Under a cluster view, a boundary item is expected to show broader, more heterogeneous contextual support than clear-category anchors. A hierarchical analysis can pool information without conflating multiple types; the goal is to quantify dispersion as a property of a boundary item, not to recast token counts as votes.

These steps remain within English and within the existing theoretical commitments. They raise evidential quality by making the dependence structure of the present results explicit, by checking behaviour against anchors, and – where feasible – by quantifying boundary-like dispersion in usage. That programme is congruent with the study's aim: to show how boundary cases can be investigated rigorously under an HPC lens without presuming that a decisive recategorization is either necessary or available.

A fourth priority, testable with independent data, is to collect acceptability judgments on constructions that differentiate pronouns from determinatives, predicting that reciprocals show higher judgment variance than clear anchors in exactly those contexts. If distance-to-boundary (as measured here) predicts judgment variance, that would confirm the stable-ambiguity pattern reflects proximity to a discrete threshold rather than inherent category fuzziness.

While this analysis focuses on English reciprocals within established theoretical frameworks, field linguists face analogous challenges with boundary phenomena in understudied languages. The core insight – that stable ambiguity can be more informative than forced categorization – applies even without elaborate statistical machinery. A field linguist documenting, say, Oceanic classifiers that show both nominal and verbal properties, might adapt this approach by: (1) documenting the full range of conflicting properties rather than privi-

leging select diagnostics, (2) explicitly comparing to clear anchors in each category, and (3) acknowledging boundary status in reference materials rather than forcing binary decisions. The statistical framework presented here may be impractical for initial documentation, but the principle of transparent, multi-lens investigation of boundary phenomena remains valuable.

10 Conclusion

If grammatical categories are homeostatic property clusters (Miller 2021) whose boundaries are sharp but epistemically inaccessible – producing apparent fuzziness near category edges – we need methods that can detect and characterize those boundary regions – especially when we're working with tiny samples. This paper offers one such method, demonstrated on English reciprocals but designed for broader application.

The core insight is simple: stop trying to force binary decisions on boundary phenomena. Instead, measure whether the diagnostic ambiguity itself is stable. Across every lens I applied – distance measurements, statistical classifiers, permutation tests, specification curves, generative models – reciprocals consistently appeared at the boundary between pronouns and compound determinatives. The best-fitting mixture weight places them almost exactly at the midpoint between anchor profiles, and this location barely budges no matter how I analyze them.

This stability of diagnostic ambiguity is informative. It tells us that reciprocals aren't just hard to classify due to noisy data or poor methods – they sit genuinely close to the category boundary, in the region where our instruments lose resolution. Their morphology pulls them toward the compound determinatives of Payne, Huddleston & Pullum (2007), their semantics toward pronouns. Within this measurement model, additional machinery is unlikely to overturn the boundary reading without new measurements or external evidence. But remember: this is internal robustness on a single instrument, not independent

convergence across datasets.

For linguists facing similar problems, here's the practical checklist:

1. Build comprehensive feature profiles (don't cherry-pick diagnostics)

2. Let visualization guide distance metrics (use ordination first)

3. Test patterns against scrambled baselines (especially with small n)

4. Vary specifications systematically (show all results)

5. Calibrate against clear cases (verify known structure)

The same two limitations I noted initially remain. Everything depends on the hand-coded feature matrix – transparent and theory-grounded, but still one person's coding. And with only two reciprocals, we're measuring position in calibrated space rather than making population inferences – the analyses characterize where these specific items fall, not whether reciprocals as a class differ from other categories. These constraints are inherent to the problem, not failures of method.

But within these constraints, the approach delivers something valuable: a way to study boundary phenomena that's both rigorous and honest about uncertainty. We don't need to pretend every word fits neatly into a predetermined box. Some words, like reciprocals, live at the boundaries, and that's a grammatical fact worth documenting carefully.

The methods transfer directly to other small-n categorization problems. Whether you're investigating modal auxiliaries, discourse particles, or any other small group with uncertain categorical status, this same toolkit applies. The code is available, the procedures are documented, and the logic is transparent.

The reciprocals remain categorized as pronouns in standard grammars including *CGEL*, and nothing here overturns that categorization. What I've shown is that they're pronouns in the way that peripheral members belong to categories – technically included but far from the prototype, occupying the same boundary region as the compound determinatives

tahat Payne, Huddleston & Pullum (2007) identifies as realizing fused grammatical functions. That's not a failure of categorization; it's a successful characterization of grammatical reality.

References

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

Gelman, Andrew & John Carlin. 2014. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science* 9(6). 641–651. https://doi.org/10.1177/1745691614551642.

Gelman, Andrew & Eric Loken. 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "*p*-hacking" and the research hypothesis was posited ahead of time. Unpublished manuscript, Columbia University. https://sites.stat.columbia.edu/gelman/research/unpublished/p_hacking.pdf.

Greenacre, Michael. 2017. *Correspondence analysis in practice*. 3rd edn. Boca Raton, FL: Chapman & Hall/CRC. https://doi.org/10.1201/9781315369983.

Huddleston, Rodney & Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Levine, Robert D. 2012. Auxiliaries: *To*'s company. *Journal of Linguistics* 48(1). 187–203. https://doi.org/10.1017/S002222671100034X.

Miklós, István & János Podani. 2004. Randomization of presence–absence matrices: Comments and new algorithms. *Ecology* 85(1). 86–92. https://doi.org/10.1890/03-0101.

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

Niculescu-Mizil, Alexandru & Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on machine learning (icml)*, 625–632. https://doi.org/10.1145/1102351.1102430.

Payne, John, Rodney Huddleston & Geoffrey K. Pullum. 2007. Fusion of functions: The syntax of *once*, *twice* and *thrice*. *Journal of Linguistics* 43(3). 565–603. https://doi.org/10.1017/S002222670700477X.

Quantifying the Differences Between Lexical Categories: The Case of Pronouns and Determinatives in English. 2021. *Cadernos de Linguística* 2(3). e399. https://doi.org/10.25189/2675-4916.2021.v2.n3.id399.

Sag, Ivan A., Rui P. Chaves, Anne Abeillé, Bruno Estigarribia, Dan Flickinger, Paul Kay, Laura A. Michaelis, Stefan Müller, Geoffrey K. Pullum, Frank van Eynde & Thomas Wasow. 2020. Lessons from the English auxiliary system. *Journal of Linguistics* 56(1). 87–155. https://doi.org/10.1017/S002222671800052X.

Simonsohn, Uri, Joseph P. Simmons & Leif D. Nelson. 2020. Specification curve analysis. *Nature Human Behaviour* 4(11). 1208–1214. https://doi.org/10.1038/s41562-020-0912-z.

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman & Wolf Vanpaemel. 2016. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science* 11(5). 702–712. https://doi.org/10.1177/1745691616658637.