Grammatical Variation Meta-Study: A Bayesian Reanalysis

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

Abstract

A common line in sociolinguistics is that some parts of grammar (like word order) are less available for social evaluation than others, and certainly than phonological variation. The empirical record on grammatical variation seems to support this: realization variables dominate the literature, while order variables are rare. But is this asymmetry a fact about language, or a fact about research practice? This paper reanalyzes the MacKenzie and Robinson (2025) database using a Bayesian selection–outcome model. The model explicitly distinguishes the probability of being tested from the probability of reporting an effect. The key finding is that selection, not capacity, is the bottleneck in the published record. Testing rates differ sharply by journal and variation type; conditional on testing, reported success rates are high and comparable across types. The asymmetry in the published record thus reflects research-design choices more than evidence for a principled incapacity of order variables.

1 Methods

1.1 Data and coding

This reanalysis uses the open database compiled by MacKenzie and Robinson (2025), covering *Language Variation and Change* (LVC) and the *Journal of Sociolinguistics* (JSlx) from each journal's first year of publication through 2023. The unit of analysis is a variable–variety pairing. If a paper studies multiple variables or a variable across multiple varieties, each counts separately. The dataset contains 427 variable–variety observations.[1] Table 1 details the distribution of variables by journal and variation type, highlighting the scarcity of order variables, particularly in JSlx.

Note that JSlx has zero observations for order variables. This means there is no direct identifying information in JSlx for an order effect; any predictions for order-in-JSlx are driven by the model structure (fixed effects and pooling) and by the LVC data. Consequently, variation-type effects are

[1] Materials and code are available at github.com/BrettRey/grammatical-variation-metastudy.

Table 1: Distribution of variables by journal and variation type in the full dataset ($N = 427$).

Journal	Realization	Order	Both	Total
JSlx	137	0	2	139
LVC	218	41	29	288
Total	355	41	31	427

estimated largely from within-LVC comparisons, and journal effects are informed primarily by realization (and "both") variables.

Coding follows MacKenzie and Robinson (2025), which adapts the form–order–omission scheme of Mansfield et al. (2023) and treats omission as a subtype of realization. Variables are included if they express grammatical meanings or functions in more than one way. Phonetic or phonological variables, lexical or discourse-pragmatic choice, discourse or conversation structure, and code-switching are excluded. The ambiguous variables (*ING*) and (*TD*) are also excluded. Variables are classified as REALIZATION, ORDER, or BOTH, with omission treated as a subtype of realization. The label "both" is used when a variable involves realization and order (for example, the dative alternation).

Social significance is coded as "not investigated", "investigated but not found", or "found".[2] When investigated, evidence is recorded as production, perception, or metalinguistic behaviours. In modelling, "tested" corresponds to "investigated", and "found" is defined conditional on testing. "Found" means "reported as found"; the category "investigated but not found" is heterogeneous, encompassing genuinely absent effects, underpowered tests, and mismatched social-meaning targets. Because "both" is a coding convention rather than a theoretical category, a sensitivity check counts such variables in both realization and order categories; this doesn't change the qualitative selection pattern.

1.2 Why a selection model?

Sociolinguistic theory has long debated whether grammatical variables carry social meaning. To answer this, researchers typically look to the published record. But that record answers a different question: among variables that researchers chose to test, which showed effects? Variables that nobody tested are invisible to any analysis that conditions on the tested subset.

Consider the dative alternation: *I gave a book to her* vs. *I gave her a book*. Can listeners hear a social difference between these forms? Do speakers use them to signal identity? To answer this, researchers first choose to test the variable. If order variables are rarely tested – perhaps because they're harder to elicit or less salient to researchers – then even a null result in the published record tells us little about their social-meaning potential. This creates a form of survivor bias. It is akin to trying to learn about

[2]The published CSV uses NA for both "not investigated" and "investigated but not found". Following the original analysis code, NAs in the social-significance fields are recoded as "investigated but not found" when evidence type is specified, and as "not investigated" otherwise. This preserves the intended three-level outcome structure.

the fleet by inspecting only the planes that made it home. A SELECTION MODEL makes this explicit by estimating two things at once: the probability that a variable gets tested, and the probability that a tested variable shows an effect. The goal isn't to infer what would happen if *all* variables were tested, but to decompose the published record into (i) selection into testing and (ii) outcomes conditional on selection. The question, then, is not only what the literature has found, but what the literature has *chosen to look for*.

1.3 Modelling strategy

The model has two stages. The SELECTION STAGE asks who gets tested; the OUTCOME STAGE asks who succeeds once tested. This structure separates research-practice effects (who gets tested) from capacity effects (who shows effects once tested).

Readers familiar with mixed-effects logistic regression – the workhorse of variable-rule analysis in tools like Rbrul or Goldvarb – will recognize the outcome stage. It's the same model: a binary outcome (found or not) predicted by fixed effects (journal, variation type, year) and random effects (paper, author, language). What's new is pairing it with a selection stage that asks why some variables were tested at all.

Predictors in both stages are journal, variation type, and year. Year is z-scored (centred and scaled), so year odds ratios represent the change associated with a one-standard-deviation increase in publication year (SD $\approx$ 9.6 years in this corpus). Random intercepts for paper, first author, and language account for clustering, and correlations across stages allow the same paper or author to influence both testing and findings.

The model is fit in Stan, a probabilistic programming language for Bayesian inference. Priors are regularising (weakly informative), discouraging extreme estimates while remaining flexible enough to follow the data; full specifications appear in Appendix A. Sampling uses four chains, 4,000 iterations (2,000 warmup), and a separate two-stage model (fitting selection and outcome independently) provides comparison.

1.4 Reading the output

Results are presented as odds ratios (ORs) and predicted probabilities. An OR of 2 means twice the odds; an OR of 0.5 means half the odds. Odds ratios are notoriously hard to intuit – and often make readers wish for predicted probabilities – so I provide those too (Figure 3). A 95% CrI is an interval containing 95% of the posterior probability mass. Rather than focusing on whether an interval "excludes 1", I emphasize the predictive implications: how much does the probability of testing or finding an effect change when shifting from realization to order variables? All fixed effects are treatment-coded with JSlx and BOTH as reference levels, and year centred at the corpus mean.

Posterior predictive checks (PPCs) ask: "Can the model reproduce data like ours?" If simulated data from the fitted model cluster around the observed data, the model is capturing the key structure. If the model can't convincingly simulate our dataset, it shouldn't be trusted to explain it.

Table 2: Posterior estimates for the selection stage (probability of being tested). Odds Ratios (OR) are presented with 95% Credible Intervals (CrI).

Predictor	OR (Median)	95% CrI
Journal: LVC (vs. JSlx)	0.37	[0.17, 0.83]
Type: Order (vs. Both)	0.50	[0.21, 1.17]
Type: Realization (vs. Both)	2.33	[1.12, 4.89]
Year (per SD ≈ 9.6 years)	1.65	[1.04, 2.67]

One feature of Bayesian models with random effects is PARTIAL POOLING: estimates for groups with little data are pulled toward the overall mean. This protects against overfitting small samples. Wide intervals for sparse categories (like order variables in perception studies) should be read as "data can't resolve this", not "effect is absent."

2 Results

2.1 Model fit and predictive checks

Before interpreting the results, does the model fit the data well? Posterior predictive checks (PPCs) answer this by simulating new datasets from the fitted model and asking whether they resemble the observed data. Here, the replicated tested and found rates cluster around the observed rates overall (Figure 1) and within journal and variation-type strata (Figure 2). The observed rates fall well inside the predictive distributions, suggesting that the model captures the main selection structure without apparent overfitting. Population-level predicted probabilities make the selection patterns easier to read (Figure 3).

2.2 Selection (tested)

Table 2 summarizes the selection stage. Odds ratios (ORs) above 1 indicate higher odds of being tested.

Testing varies notably by journal, variation type, and year. First, LVC is less likely to test variables than JSlx. In plain terms, variables in LVC have about one-third the odds of being tested for social significance compared to JSlx. Second, as Table 2 shows, realization variables are more likely to be tested than the "both" reference category (OR ≈ 2.33), while order variables trend lower (OR ≈ 0.50) with substantial uncertainty. Taken together, the posterior favours higher testing odds for realization than for order, but the order estimate is imprecise. Finally, testing increases over time (OR ≈ 1.65); per SD (≈ 9.6 years), testing odds increase by about 65%.

2.3 Outcome (found | tested)

Conditional on testing, success rates are high and comparable across types. The probability of finding social significance is high (approximately 92–95% across subgroups; Figure 3) and shows wide overlap across journals and variation types. Outcome effects have intervals that span no effect, consistent with selection rather than capacity as the bottleneck. Wide credible intervals in sparse categories –

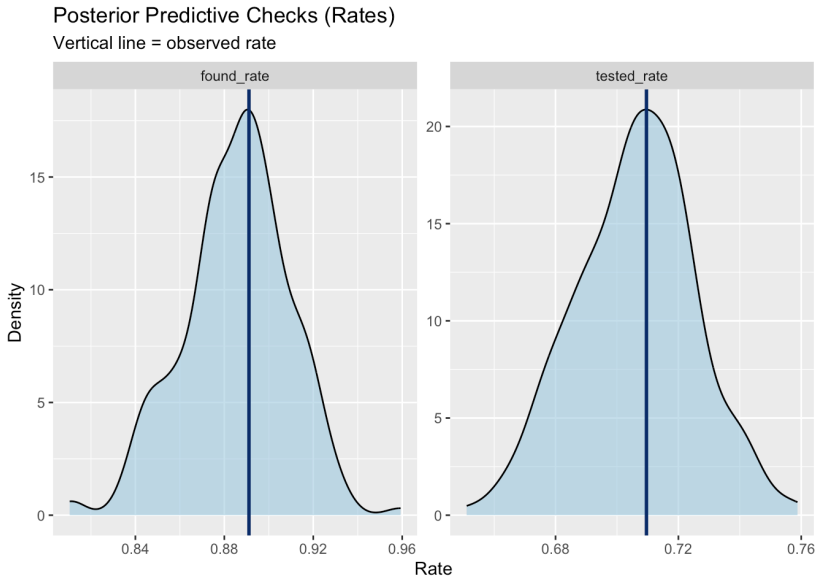

Figure 1: Overall posterior predictive check. Each histogram shows rates from 200 simulated datasets; vertical lines mark the observed rates. If the observed rates fall inside the predictive distribution, the model captures the data generating process.

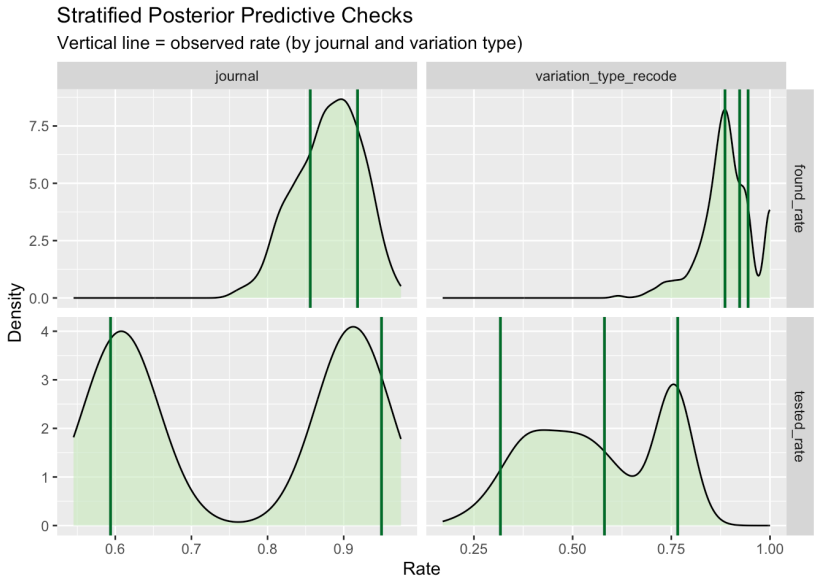

Figure 2: Stratified posterior predictive checks by journal and variation type. The same logic applies within subgroups: if observed rates (vertical lines) fall inside the simulated distributions, the model captures the subgroup structure.

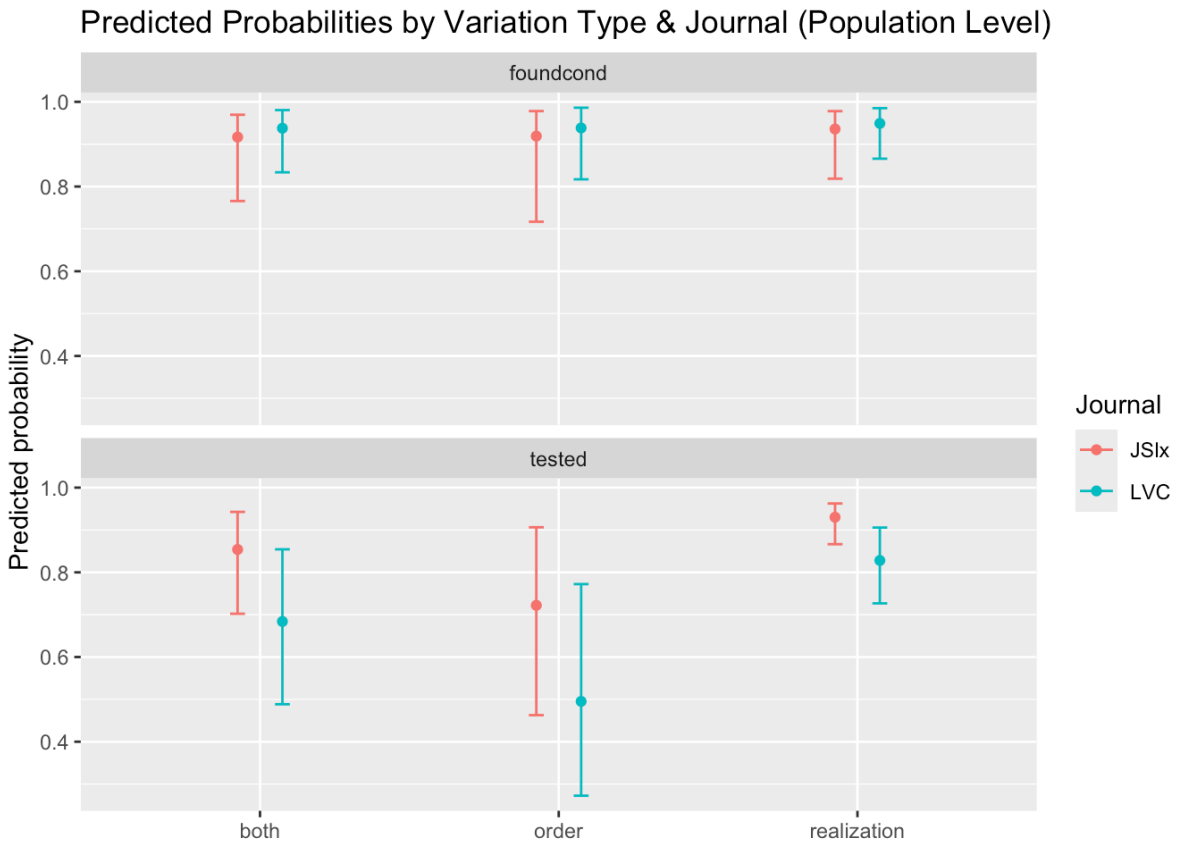

Figure 3: Population-level predicted probabilities of testing (top) and finding significance (bottom) by journal and variation type. Points show posterior medians; error bars show 95% credible intervals. Non-overlapping intervals suggest reliable differences; overlapping intervals suggest uncertainty about whether groups differ.

order variables, perception and metalinguistic domains – reflect data limitations, not absence of effect. Because "found" tracks what is reported in print, these high success rates should be read as properties of the published record (and its filters), not as direct estimates of underlying social-meaning capacity.

2.4 Robustness

A good analysis shouldn't depend on arbitrary modelling choices. Several robustness checks confirm that the main findings are stable:

The separate two-stage model yields the same qualitative pattern (Table 3): selection effects are stable, outcome effects are weak. The centred-intercept sensitivity run – which uses priors centred on observed rates rather than agnostic priors – reproduces the same conclusions.

An extended model adds a journal × variation-type interaction in the selection stage (full estimates in Appendix A). If institutional framing requirements drive selection, JSlx might show a larger gap between realization and order than LVC. The interaction coefficients represent the multiplicative change in the LVC effect for each variation type relative to the reference (both). These coefficients

Table 3: Odds ratio comparison: Joint model vs. separate two-stage model. Estimates are posterior medians [95% CrI]. Both models show the same qualitative pattern: selection effects replicate; outcome effects are weak and uncertain.

		Joint model	Two-stage model
Stage	Predictor	OR [Lo, Hi]	OR [Lo, Hi]
Tested	Intercept	5.48 [2.07, 14.75]	7.71 [2.50, 25.36]
	LVC (vs JSlx)	0.37 [0.17, 0.83]	0.30 [0.14, 0.68]
	Order (vs Both)	0.50 [0.21, 1.17]	0.46 [0.19, 1.11]
	Realization (vs Both)	2.33 [1.12, 4.89]	1.87 [0.84, 4.05]
	Year (z-scored)	1.65 [1.04, 2.67]	1.63 [1.03, 2.59]
Found	Intercept	10.76 [3.31, 35.24]	17.86 [4.80, 67.86]
	LVC (vs JSlx)	1.34 [0.63, 2.83]	1.06 [0.46, 2.27]
	Order (vs Both)	1.04 [0.42, 2.55]	0.97 [0.39, 2.45]
	Realization (vs Both)	1.36 [0.60, 3.06]	0.88 [0.37, 2.09]
	Year (z-scored)	1.27 [0.73, 2.22]	1.20 [0.69, 2.05]

have wide intervals spanning no effect (order:LVC OR ≈ 0.54, 95% CrI 0.23–1.29; realization:LVC OR ≈ 0.66, 95% CrI 0.30–1.47), so the data don't clearly support differential selection by journal.

A piecewise-linear year effect (two interior knots) tests whether selection trends are non-linear. The spline coefficients point in the same direction (ORs 1.19, 1.72, 1.32; all intervals spanning 1), consistent with the primary model's linear year assumption rather than distinct period effects.

Summary: Selection effects are robust across modelling choices. The outcome effects remain weak and uncertain regardless of how the model is specified, strengthening the conclusion that selection – not capacity – drives the published pattern.

3 Discussion

3.1 Selection vs. outcome

The main asymmetry is in selection. Journals, variable type, and time shape which variables are tested, but tested variables show high and overlapping probabilities of social significance. This supports a selection-driven account of the literature: the bottleneck is which variables are studied, not whether tested variables yield positive findings once investigated. This conclusion concerns publication and research-design practices rather than the underlying capacity of grammatical variables to carry social meaning. In that sense, the record is informative not only about social meaning, but about what researchers have chosen to look for.

3.2 Implications for the Grammatical Invisibility Principle

The Grammatical Invisibility Principle (GIP) holds that grammatical variables are less accessible to social evaluation than phonological variables. MacKenzie and Robinson (2025) conclude that for realization variables, robust evidence of social significance exists; for order variables, absence of evidence

doesn't equal evidence of absence. The selection model sharpens this: *conditional on testing*, both realization and order variables succeed at comparable rates (92–95%). This high success rate likely reflects a strong filter: once researchers choose to test a variable, they overwhelmingly report finding effects. A success rate near 90% is either a triumph of sociolinguistic detection or a clue about which results survive the trip to print; the model can't separate these. This suggests that the tested variables may be positively selected for plausibility. In other words, this likely represents the order-variable honour roll, not the whole graduating class. The finding undercuts the strong version of the GIP – that order variables *can't* carry social meaning – while remaining agnostic about whether order variables are *harder to perceive* or *less frequently recruited*. On a causal-network view, the relevant question is not whether order variables *can* carry social meaning, but which causal pathways (e.g., perceptual salience, frequency, register mediation) make them projectible targets for social evaluation. The current literature is sparse precisely where those pathways would need to be mapped.

3.3 Relation to prior descriptive work

The descriptive goal of MacKenzie and Robinson (2025) is to catalogue grammatical variables and assess the state of evidence for social significance in each domain. I retain their coding scheme and tallies without modification. The points of agreement are substantial: realization variables dominate the literature; order variables are understudied; perception and metalinguistic domains remain sparse.

The focused revision is methodological. MacKenzie and Robinson devote considerable discussion (their Section 5) to selection into testing, including the file drawer problem and the fact that LVC doesn't require a social-significance framing. They explicitly warn against treating the scarcity of order-variable evidence as evidence that such effects don't exist, emphasizing that the relevant studies have largely not been done. What I add is an explicit model for this selection mechanism. Rather than treating untested variables as missing data to be ignored, the joint model estimates the probability of testing and conditions the outcome analysis on testing. This formalizes what MacKenzie and Robinson discuss narratively.

The model's selection estimates converge with their descriptive findings. In this dataset, among production studies, LVC finds social significance at 92% (155/169) compared to JSlx at 83% (87/105). This 9-point gap supports their argument that LVC's lack of a social-significance framing requirement doesn't inflate publication bias – researchers who look for effects overwhelmingly find them.

The novelty is formalization rather than conceptual revision. The selection–outcome decomposition provides a principled statistical framework for claims that MacKenzie and Robinson already make informally. The difference in emphasis is slight: the high baseline success rate (92–95%) is the story, and the between-type comparison is null in part because of ceiling effects and limited order data.

3.4 Limitations and future work

Order variables are rare and heavily concentrated in LVC – the journal that is *less* likely to test for social significance. This constrains journal comparisons. Could the data distinguish a 15-point difference in

success rates between realization and order? No. The baseline is too high, the order-variable sample too small. The null is underpowered, not confirmed. Also, the binary "found" outcome collapses heterogeneity in evidence strength (p-values, authors' qualitative judgments), and high success rates likely reflect publication bias. Finally, because the database indexes published studies, the model can't address file-drawer effects where testing occurs but null results aren't published. The model describes the published record, which acts as a filter on the underlying linguistic reality.

MacKenzie and Robinson also call for collaboration with formal syntacticians to ground the realization/order distinction theoretically. This might predict sociolinguistic differences beyond what the current selection model captures. Future work should expand the corpus beyond two journals, target perception and metalinguistic studies for order variables, and explore whether formal-syntactic classifications improve predictive power.

Acknowledgements

ChatGPT 5.2 and Claude Opus 4.5 assisted with analysis and drafting, but I'm responsible for all claims and any errors.

A Model specification

The joint selection–outcome model has two stages. For observation i:

Selection stage:

$$\text{tested}_i \sim \text{Bernoulli}(\text{logit}^{-1}(\eta_i^s)) \tag{1}$$

$$\eta_i^s = \alpha^s + \mathrm{x}_i\boldsymbol{\beta}^s + u^s_{\text{paper}[i]} + u^s_{\text{author}[i]} + u^s_{\text{language}[i]} \tag{2}$$

Outcome stage (conditional on $\text{tested}_i = 1$):

$$\text{found}_i \mid \text{tested}_i = 1 \sim \text{Bernoulli}(\text{logit}^{-1}(\eta_i^o)) \tag{3}$$

$$\eta_i^o = \alpha^o + \mathrm{x}_i\boldsymbol{\beta}^o + u^o_{\text{paper}[i]} + u^o_{\text{author}[i]} + u^o_{\text{language}[i]} \tag{4}$$

where x_i is the row of the fixed-effects design matrix containing journal, variation type, and year (z-scored).

Random effects: For each grouping factor $g \in \{\text{paper}, \text{author}, \text{language}\}$:

$$\begin{pmatrix} u^s_{g[i]} \\ u^o_{g[i]} \end{pmatrix} \sim \text{MVN}\left(\mathrm{o}, \boldsymbol{\Sigma}_g\right), \quad \boldsymbol{\Sigma}_g = \begin{pmatrix} (\sigma^s_g)^2 & \rho_g\sigma^s_g\sigma^o_g \\ \rho_g\sigma^s_g\sigma^o_g & (\sigma^o_g)^2 \end{pmatrix} \tag{5}$$

Priors:

$$\alpha^s, \alpha^o \sim \text{Normal}(0, 1.0) \tag{6}$$

$$\beta^s_k, \beta^o_k \sim \text{Normal}(0, 0.5) \tag{7}$$

$$\sigma^s_g, \sigma^o_g \sim \text{Exponential}(3) \tag{8}$$

$$\mathrm{L}_g \sim \text{LKJ-Cholesky}(2) \tag{9}$$

where L_g is the Cholesky factor of the correlation matrix for grouping factor g.

References

MacKenzie, L., & Robinson, M. (2025). Spelling out grammatical variation. In D. Duncan & M. Robinson (Eds.), *English sociosyntax: Theory, evidence, approaches* (pp. 59–95). De Gruyter Mouton.

Mansfield, J., Leslie-O'Neill, H., & Li, H. (2023). Dialect differences and linguistic divergence: A cross-linguistic survey of grammatical variation. *Language Dynamics and Change*, *13*, 232–276.