The Gravity of Evaluation: Modelling Sparse Evaluative Bundle Frequencies in Learner Corpora with PPML

Tim Marchand
Gakushuin University, Tokyo, Japan
tim.marchand@gakushuin.ac.jp

Abstract

A persistent methodological problem in corpus-based interlanguage research concerns the appropriate statistical estimator for lexical bundle frequency outcomes. Bundle counts are non-negative, heavily right-skewed, and populated with meaningful zeros, distributional properties that distort standard log-transformed linear models. This paper argues that Poisson Pseudo-Maximum Likelihood (PPML) estimation, developed by Silva and Tenreyro (2006) for gravity-model trade flows, offers a principled solution by virtue of a structural analogy: both Zipfian corpus frequency distributions and gravity-model outcomes exhibit the same power-law sparsity that makes log-linearisation problematic, and PPML's consistency under misspecification makes it a particularly robust estimator for corpus count data. We demonstrate this through a proof-of-concept analysis of evaluative tag-bundle development in the ICNALE Japanese learner corpus (A2-B2), with bundle selection grounded in Biber's MDA framework (Larsson, Paquot and Biber, 2021) and Gries's (2024) tupling paradigm. PPML models reveal evaluative register routing, learners overuse simpler interpersonal constructions while underusing NP-internal modification, across four distinct developmental trajectories.

Keywords: PPML, learner corpus, lexical bundles, MDA, register routing, ICNALE, normalised entropy

1. Introduction

A persistent challenge in corpus-based interlanguage research concerns the appropriate statistical estimator for lexical bundle frequency outcomes. Bundle counts are non-negative, heavily right-skewed, heteroskedastic, and populated with structurally meaningful zeros, representing distributional properties inherited from Zipf's law (Zipf, 1935), which governs the rank-frequency distribution of all linguistic items from words to n-gram bundles. These properties remain difficult to accommodate within commonly used log-linear models, yet they remain the field's default, inherited largely from computational convenience. This paper argues for a principled alternative by drawing on an econometric literature that has confronted the same distributional challenge in a different domain.

Silva and Tenreyro (2006) demonstrated that log-linearising multiplicative models in the presence of heteroskedastic errors produces biased and inconsistent estimates due to Jensen's inequality: the mathematical principle that $E[\log(y)] \neq \log(E[y])$ when y is heteroskedastic. They developed Poisson Pseudo-Maximum Likelihood (PPML) estimation as a solution for gravity-model trade flows, econometric models that happen to share two distributional properties (power-law sparsity and structurally meaningful zeros), with Zipfian corpus frequency data. This paper introduces PPML to corpus linguistics through this structural analogy and provides a proof-of-concept analysis of evaluative tag-bundle development in the ICNALE Japanese learner corpus.

This proof-of-concept analysis is undertaken in service of a larger ongoing project applying the same tagging and PPML pipeline to a longitudinal below-the-line (BTL) learner corpus, whose considerably more irregular structure (incomplete proficiency metadata, uneven posting frequency across writers, and reply chains spanning multiple cohorts) will place far greater demands on any estimator than the present dataset. ICNALE is used here specifically because its balanced design is particularly favourable to random-effects estimation: each writer contributes exactly two essays of comparable length, proficiency metadata is complete, and writer identity is unambiguous. The corpus therefore provides a conservative test of PPML's utility, evaluated under conditions that, if anything, favour the mixed-effects alternative rather than the estimator this paper argues for.

This paper offers three contributions. First, methodologically, the Zipf-gravity analogy motivates PPML as a principled estimator for sparse corpus counts, with its formal justification resting on the conditional-mean argument developed in Section 2.1. Second, procedurally, a theory-driven selection process separates the operationalisation of feature categories from statistical inference, thereby preventing the circularity common to data-driven variable selection. This approach is justified both on theoretical grounds and through the holdout robustness check reported in Section 4.4. Third, substantively, the analysis identifies evaluative register routing in Japanese EFL argumentative writing, confirmed through lexical realisation analysis using normalised Shannon entropy as a measure of formulaic concentration, extracted using the MDA Tagger (Marchand, 2026).

The remainder of this paper is organised as follows. Section 2 sets out the Zipf-gravity analogy formally, situates the analysis within second language register theory, and offers a worked illustration of Zipfian sparsity drawn directly from the dataset. Section 3 describes the corpus, the MDA Tagger's tagging and bundle-extraction procedure, the bundle selection pipeline, and the analytical and validation methods, including the zero-inclusive dataset construction required for the analyses that follow. Section 4 reports the bundle models, the GLMM comparison, the developmental trajectories and the holdout robustness results. Section 5 discusses methodological, substantive, and pedagogical implications and the study's limitations, before Section 6 concludes.

2. Theoretical Background

2.1 The Zipf-Gravity Analogy and Jensen's Inequality

The field of corpus linguistics has historically relied on log-transformed count data fed into linear or log-linear models (e.g., Baayen, 2008; Gries, 2015, 2021), a practice inherited partly from computational convenience and partly from the assumption that normalisation renders frequency distributions tractable. This section develops the alternative introduced above, deriving a more appropriate estimator from a theoretical analogy between two well-established power law phenomena: Zipf's law in linguistics (Zipf, 1935, 1949) and the gravity model in economics and economic geography (Tinbergen, 1962).

Zipf's law describes the rank-frequency distribution of words in natural language corpora as following a power law: a small number of lexical items account for a disproportionate share of tokens, while the vast majority of items occupy a long, sparse tail, particularly at the level of

multiword units, n-grams, and lexical bundles (Baayen, 2001; Biber et al., 1999; Ellis et al., 2008). This distributional structure is not incidental but constitutive: it characterises language at every level of granularity from individual words to n-gram bundles, and it means that corpus frequency data are structurally non-normal, heavily right-skewed, and populated with meaningful zeros in the long tail, a property known as the Large Number of Rare Events (LNRE) phenomenon (Baayen, 2001; Evert, 2005).

In empirical economics, Tinbergen's (1962) gravity model predicts bilateral trade as proportional to economic mass and inversely proportional to geographic distance. In practice, trade data look remarkably like corpus frequencies: non-negative, heavily skewed, and full of zeros representing negligible or absent flows. This structural parallel is not superficial: both phenomena are power law distributions arising from multiplicative generative processes, and both resist the log-linearisation that would make them amenable to ordinary least squares estimation.

The conventional response to this challenge is log-transformation: if y_i are bundle counts, $\log(y_i + 1)$ or $\log(y_i/N_i)$ is modelled as a linear outcome. Jensen's inequality shows this produces biased estimates under heteroskedasticity. For a strictly concave function f such as the logarithm, Jensen's inequality states $E[f(x)] < f(E[x])$. This means $E[\log(y)] < \log(E[y])$; the expected log is not the log of the expectation. Under heteroskedasticity, this gap varies systematically across observations: $E[\log(y)|x]$ does not identify $\log(E[y]|x)$, so retransforming a log-linear model's fitted values need not recover the conditional mean of the count outcome.

This problem is not unique to econometrics: ecologists confronting the same log-transformation practice for count data have independently reached the same conclusion, noting that the transformation is undefined at zero and that heteroskedasticity typically persists even after transformation (O'Hara and Kotze, 2010).

Silva and Tenreyro (2006) demonstrated formally that log-linearising a multiplicative gravity model in the presence of heteroskedastic errors, the standard practice at the time, produces systematically biased and inconsistent estimates, with the bias concentrated in the sparse, near-zero cells that constitute the long tail of the distribution. Their proposed solution, Poisson Pseudo-Maximum Likelihood (PPML) estimation, models $E[y|x] = \exp(x\beta)$ directly, exploiting the Poisson likelihood as a quasi-likelihood function: consistency requires only that the conditional mean is correctly specified, not that the outcome is genuinely Poisson-distributed. This relaxes the strict *mean-equals-variance* assumption often associated with standard Poisson regression, an assumption that typically leads applied linguists to abandon it in favour of negative binomial alternatives at the first sign of overdispersion. PPML remains consistent under this kind of variance misspecification provided the conditional mean itself is correctly specified, with any resulting imprecision in the standard errors handled separately by the clustered sandwich estimator described below. This distributional flexibility matters directly for corpus data, where overdispersion often manifests as word “burstiness” (a word's tendency to recur disproportionately once it has already occurred in a text), the very property that had led some corpus linguists to reject standard Poisson models in favour of Poisson-mixture or negative binomial alternatives (Church and Gale, 1995; Winter and Bürkner, 2021). PPML directly resolves this concern through its quasi-likelihood properties. Since coefficient consistency relies solely on the correct specification of the conditional mean, burstiness-driven overdispersion does not bias point estimates; its effect on standard errors is handled robustly using the clustered sandwich estimator described in Section 3.3.

PPML has since become the default estimator for gravity-type models across empirical economics, extending well beyond bilateral trade to foreign direct investment, migration, and patent citation counts, domains united not by subject matter but by the same Zipfian statistical signature of sparse, heavy-tailed, non-negative counts. Silva and Tenreyro's result is a general claim about estimation under heteroskedastic multiplicative error, not one specific to trade economics, and the present paper argues that the same generality extends to lexical bundle frequencies.

The case for Poisson-family regression has in fact already been made within linguistics itself, independently of the econometric literature this paper draws on. Winter and Bürkner (2021) argue that count data pervades the discipline, from word frequencies to speech errors to co-speech gestures, and that Poisson regression remains surprisingly little used relative to its applicability, absent even from most statistics textbooks aimed at linguists. Using a Bayesian mixed-effects framework, they show that Poisson and negative binomial regression handle count data, including overdispersed count data, more defensibly than log-transformation, and they identify corpus linguistics specifically as a field that would benefit from wider adoption, noting that the discipline's continued reliance on tests such as chi-square violates the independence assumption corpus data routinely fail to satisfy. The present paper takes up this call for the specific case of lexical bundle frequencies, but arrives at its estimator by a different route. PPML and the Bayesian Poisson-family models Winter and Bürkner advocate are, in this sense, complementary answers to the same underlying problem, motivated from different theoretical directions but converging on the same rejection of log-transformation as the default.

2.2 Evaluative Register Routing in Learner Writing

Multi-Dimensional Analysis (Biber, 1988) provides the theoretical framework for predicting which evaluative features should show learner-NS divergence in argumentative EFL writing. Dimension 1 (involved production) loads high on first-person pronouns, private and suasive verbs, stance adverbs, and predicative adjectives. Dimension 3 (elaborated reference) loads high on attributive adjectives, nominalisations, and complex prepositional phrases. Larsson, Paquot and Biber (2021) demonstrate that learner writing in general is displaced toward Dimension 1 and away from Dimension 3 relative to expert NS writing, a pattern confirmed for Japanese and Asian EFL argumentative writing specifically (Ha, 2022; Ishikawa, 2013; Granger and Rayson, 1998; Kyle and Crossley, 2017). This pattern has been independently corroborated in ICNALE-based lexical bundle research conducted outside the MDA tradition. Sawaguchi (2024) examined four-word lexical bundles produced by Japanese A2 and B1 learners in the ICNALE and found that B1 learners deploy more varied stance bundles (e.g., *believe that it is*) and discourse bundles (e.g., *it is up to*) than A2 learners, but that referential bundles show markedly slower development that persists even at B2, a finding that directly anticipates the deepening underuse trajectory identified for the adjective-noun-preposition bundle in Section 4.2. Muşlu (2018) compared stance lexical bundle use in argumentative essays by Turkish and Japanese EFL learners against a native-speaker baseline, reporting

systematic structural and functional divergence in how the two learner groups deploy stance bundles relative to L1 English writers. Most strikingly, Sitler and Spivey (2025) title their multimodal register study of ICNALE lexical bundles after the very phrase identified in Section 4.3 as the dominant realisation of the first-person suasive bundle (*I agree with this*). Their study independently confirms its salience as a formulaic marker in Japanese EFL argumentative writing; using a Generalized Linear Mixed Model, they report that overall bundle frequency decreases as proficiency increases, a pattern broadly consistent with the B2 convergence trajectory reported below.

We characterise this overall pattern as *evaluative register routing*: learners channel evaluative meaning through the syntactically simpler and more overtly interpersonal constructions of Dimension 1 rather than the NP-internal adjectival modification of Dimension 3. This is not merely a frequency effect: as the concordance-based analysis will show, learners and NS writers frequently use structurally similar patterns for qualitatively different communicative purposes, a distinction that aggregate frequency counts cannot reveal but that lexical realisation analysis can.

2.3 Zipfian Sparsity in the Feature Bundle Data

The abstract argument in Section 2.1 is directly visible in the bundle data reported later in this paper. Table 2 shows that even the most frequent of the six modelled bundles is present in only 69.5% of files, and the least frequent in just 16.7% (Section 3.5), the mark of meaningful zeros throughout the corpus, not occasional gaps, and precisely the kind of per-file count data PPML is designed to model. Under a conventional log(count + 1) transformation applied to a count variable with this much zero mass, the resulting bias is concentrated exactly where Silva and Tenreyro (2006) predict: in the sparse, near-zero cells that dominate the distribution, understating genuine differences between the small number of files with high counts and the large number with none. This is precisely the Jensen's inequality distortion described in Section 2.1.

3. Data and Methods

3.1 Corpus and Tagging

The analysis uses the ICNALE Japanese learner sub-corpus (Ishikawa, 2013): 800 essays by Japanese university students at A2, B1, and B2 CEFR proficiency levels across two argumentative topics: part-time employment for college students (PTJ), and a proposed restaurant smoking ban (SMK). The native-speaker baseline is the ICNALE English sub-corpus: 400 essays by college-educated NS writers on the same topics. Essays average 250-300 words under a 20-40 minute time limit, controlling composition conditions across proficiency levels and providing `file_wc`, the per-file word count used as the offset variable throughout the PPML specification (Section 3.3). Writer IDs are tracked across essays, giving a nested structure of two essays per writer that the PPML cluster correction accounts for.

All texts were tagged using the MDA Tagger (Marchand, 2026), an R-based Multi-Dimensional Analysis toolkit that applies Biber's (1988) feature taxonomy to produce vertically formatted output in the form token_{{Penn<MDA>}}, with Penn Treebank tags extended by MDA functional categories in angular brackets. A representative extract from a tagged A2 essay illustrates the format in (1):

(1) *I_{{PRP<FPP1>}}*
agree_{{VBP<PUBV><SUAV><VPRT>}}
with_{{IN<PIN>}}the_{{DT}}
preceding_{{NN<GER>}}
statement_{{NN<NOMZ>}}

Here, the suasive verb agree receives the compound tag VBP<PUBV><SUAV><VPRT>, marking it simultaneously as a public verb, a suasive verb, and a present-tense verb phrase. These extended tags can appear ungainly, combining standard Penn Treebank structural tags with stacked MDA functional categories, but this apparent redundancy is useful: it preserves crucial grammatical distinctions while encoding the fine-grained functional information the evaluative bundle filter below requires. More concise versions of the feature bundles appear where brevity is a virtue, such as in Table 3.

Following tagging, a bundle-extraction routine slides a fixed-width window across the tag sequence of each file to generate all observed n-gram tag combinations, tallying their frequency per file. The default bundle length is set to three tags (trigrams), following established practice in lexical bundle research (Biber et al., 2004) to target the level of granularity at which functional sequences become interpretable as recurrent phraseological patterns rather than incidental co-occurrences.

A key design choice here is defining bundles over tag sequences rather than surface words. For the sequence in (1), the extraction routine yields the feature bundle template in (2):

(2) *{{PRP<FPP1>}}*
{{VBP<PUBV><SUAV><VPRT>}} {{IN<PIN>}}

This buys two practical advantages: it pools low-frequency lexical variants into functionally coherent groups, and it links the extracted units directly to established MDA grammatical categories. Because extraction operates purely on tag sequences, (2) is agnostic to its surface realization, matching *I agree with*, *I agree on*, and *I insist on* alike. This allows the same structural template to be examined both quantitatively, as a frequency outcome, and qualitatively, via the concordancing functionality described in Section 3.4.

Finally, the extracted bundle inventory is filtered to retain only evaluative bundles, operationalized as any tag sequence containing at least one adjectival (JJ), adverbial (RB), or suasive verb (SUAV) tag in any functional variant. This criterion aligns with the study's focus on evaluative register routing, where adjectives and adverbs function as primary carriers of stance and assessment, and suasive verbs function as explicit stance markers within Biber's MDA framework.

3.2 Theory-Driven Feature Bundle Selection

Feature bundle selection follows a theory-driven two-stage procedure. In the first stage, nine functional categories predicted to show learner-NS divergence are specified *a priori* from MDA theory and learner corpus research: four Dimension 3 underuse categories (attributive NP modification, prepositional evaluative NP, attributive NP with nominalisation, adjective-noun-preposition) and five Dimension 1 overuse categories (sentence-boundary stance adverbial, amplified passive, predicative adjective frame, wh-suasive frame, first-person suasive verb). In the second stage, Gries's (2024) tupling paradigm confirms empirical realisation: weighted log-odds $|z| \geq 2.58$ (Monroe, Colaresi and Quinn, 2008), Gries's (2008) dispersion measure DP $\leq$ 0.75 at writer level, and corpus-wide frequency ≥ 10. DP is calculated as DP = $0.5 \times \Sigma|\text{observed proportion} - \text{expected proportion}|$ across writers, where the expected proportion reflects each writer's share of total corpus word count; lower

values indicate even dispersion across the writer population rather than concentration in a small subset of prolific users. To illustrate: the focal bundle of Model 1 ({{DT}} {{JJ<JJ>}} {{NN<NN>}}; henceforth M1) returns a weighted log-odds of -5.68 in the JPN-vs-NS comparison, far exceeding the $|z| \geq 2.58$ threshold and confirming it as one of the most distinctive underused categories in the corpus. Its writer-level DP of 0.33 indicates the bundle, where it occurs, is fairly evenly spread across the learner population rather than concentrated in a handful of prolific writers, and its corpus-wide total of 1,895 tokens comfortably exceeds the frequency floor. Seven of the nine predicted bundle types pass these criteria; the DP computation, and the reason two do not, is described in Section 3.5.

3.3 PPML Estimation

For each selected bundle, a PPML model is fitted using `fixest::feglm()` (Bergé, 2018) with the quasipoisson family and standard errors clustered by `writer_id`. The regression specification (`raw_count ~ level * topic + offset(log(file_wc))`) includes total word count as an exposure offset, ensuring bundle occurrences are modelled as a rate per word rather than an absolute count driven by essay length. Clustering treats a writer's essays as potentially correlated while assuming independence across different writers; the resulting sandwich estimator, so named for the form of its variance-covariance matrix, remains valid without requiring the count distribution's variance to be correctly specified, the robustness property motivating PPML's use in Section 2.1.
Level has four categories (NS, A2, B1, B2) and topic two (PTJ, SMK), generating an intercept plus seven predictor coefficients per model; Benjamini-Hochberg (BH) correction (Benjamini and Hochberg, 1995) is applied across these seven terms within each model, treating the seven inferential comparisons per bundle as the prespecified family of tests for that outcome. For comparison, each bundle is additionally fitted as a Generalised Linear Mixed Model (GLMM) with writer random intercept using lme4 (Bates et al., 2015), with both Poisson and negative binomial families: the frequentist implementation of the count-appropriate regression approach that Winter and Bürkner (2021) and Gries (2015) argue corpus linguistics should adopt in place of log-transformation, and the most readily available route to that approach given existing familiarity with lme4-based mixed models in the field.

3.4 Concordance-Based Analysis

To examine the lexical instantiation of each bundle, the MDA Tagger's concordancing function was used to extract KWIC concordances for all hits in both JPN and NS sub-corpora, with a five-token window on each side of the node (Hunston and Francis, 2000). Because the tagger retains the surface token alongside every tag, retrieving a concordance for a given tag-bundle template requires no separate lexical lookup: the same tagged files that supply the quantitative frequency counts in Section 3.1 directly supply their qualitative concordance lines. The top five most frequent lexical realisations were identified from the complete hit set before random sampling. Lexical diversity was quantified using two complementary measures: type-token ratio (TTR = types/tokens) and normalised Shannon entropy (H_norm; Pielou, 1966). For a bundle with k distinct realisations each occurring with probability p_i, Shannon entropy ($H = -\Sigma\, p_i \log_2(p_i)$; Shannon, 1948) and normalised entropy ($H_{norm} = H / \log_2(k)$) were calculated. H_norm ranges from 0 (single realisation dominates) to 1 (uniform distribution across all types). Unlike TTR, H_norm weights by frequency: a bundle where 98% of hits share one type will have H_norm ≈ 0 regardless of how many rare types exist, providing a robust metric for distributional skewness (Tweedie and Baayen, 1998; Gries, 2008).

3.5 Feature Bundle Sparsity

Table 2 reports, for each of the six modelled bundles, the proportion of files in which it occurs at all (Range). Even the most frequent bundle, M1, is present in only 69.5% of files; the least frequent, M4, in just 16.7%. This confirms empirically the Zipfian sparsity argued for on theoretical grounds in Section 2.1: absence is not missing data but a substantive part of each bundle's frequency distribution, and every regression model reported in Section 4 is fitted on the full file-by-bundle cross-tabulation, including these zero counts, rather than on the subset of files in which a bundle happens to occur.
Two further theoretically motivated bundles are excluded from the confirmatory analysis on related grounds: the wh-suasive frame (present in only 3.7% of files) and the nominalised NP frame (12.3%) both exceed Gries's (2008) dispersion threshold ($DP \leq 0.75$), indicating both are concentrated in a small subset of writers rather than being stable, corpus-wide register features. A third bundle, the amplified passive, is excluded at the model-fitting stage for an unrelated reason: it occurs in only one of the two essay topics. This is a different kind of restricted attestation from the wh-suasive frame's population-level absence discussed below: a topic-specific gap most plausibly reflects the essay prompts themselves rather than any distributional or developmental property of learner or native-speaker language, so it is not discussed further here. Although excluded from the confirmatory regression analysis, the nominalised NP frame is examined qualitatively in Section 4.3, where its concordance profile offers independent support for the writer-level dispersion evidence presented here, while wh-suasive frame is examined in Supplementary Materials as a case of estimation under even more extreme sparsity: the complete absence of NS occurrences causes both PPML and GLMM to encounter separation.

3.6 Holdout Robustness Check

The theory-driven selection procedure in Section 3.2 addresses the circularity risk of data-driven dependent variable selection in principle. As a further, direct empirical check, the learner writer pool (excluding the NS reference group) was randomly partitioned into an 80% discovery set (320 writers) and a 20% holdout set (80 writers); the six bundle models were refitted on the holdout writers alone, on the same zero-inclusive dataset, with NS writers retained in both partitions. Coefficient signs and magnitudes that replicate on this unseen partition indicate the trajectory findings reflect genuine population-level patterns rather than artefacts of the specific writer sample used for bundle selection; results are reported in Section 4.4.

4. Results

4.1 Bundle Models and GLMM Comparison

Table 2 summarises the six valid bundle models. Comparison with GLMMs, fitted on the same zero-inclusive dataset used for the PPML models, shows that every model converges without a singular fit for both the Poisson and negative binomial families: random intercept variance is estimated as a substantive, non-zero quantity throughout,

Table 2: Feature bundles, example realisations, corpus range, direction, and developmental trajectory

Model: Feature bundle	Example	Range	Dir.	Trajectory
M1: *{{DT}} {{JJ<JJ>}} {{NN<NN>}}*	a good idea	69.5%	Under	Partial recovery
M2: *{{.}} {{RB<RB>}} {{,}}*	. So ,	54.2%	Over	Sustained overuse
M3: *{{IN<PIN>}} {{DT}} {{JJ<JJ>}}*	at the same	40.2%	Under	U-shaped, persists
M4: *{{PRP<FPP1>}} {{VBP<PUBV><SUAV><VPRT>}} {{IN<PIN>}}*	I agree with	16.7%	Over	Sustained overuse
M5: *{{JJ<JJ>}} {{NN<NN>}} {{IN<PIN>}}*	major cause of	47.7%	Under	No longer detectable by B2
M6: *{{PRP<PIT>}} {{VBZ<BEMA><VPRT>}} {{JJ<PRED>}}*	it is important	48.9%	Over	Sustained overuse, strong topic effect

ranging from approximately 0.10 for the adjective-noun-preposition bundle to approximately 0.37 for the sentence-boundary adverbial bundle.

Point estimates agree closely between PPML and GLMM Poisson throughout, typically within a small fraction of a standard error (see Supplementary Materials for details). This confirms that the underlying signal is robust to estimator choice. Standard errors show no systematic direction of inflation: the ratio of GLMM to PPML standard errors is sometimes above 1 and sometimes below across the level-effect terms, consistent with two different but both valid approaches to variance estimation applied to correctly specified data, rather than the uniform pattern of GLMM under-estimation that an incompletely specified dataset would produce. Point estimates for the headline bundle M1, for example, are near-identical under both estimators (A2: b = -0.978 PPML vs -0.996 GLMM; B1: b = -0.915 vs -0.927; B2: b = -0.718 vs -0.727), with standard errors of comparable magnitude at every level, as Figure 1 illustrates directly across all seven model terms and all three estimators, including the negative binomial specification not otherwise discussed here.

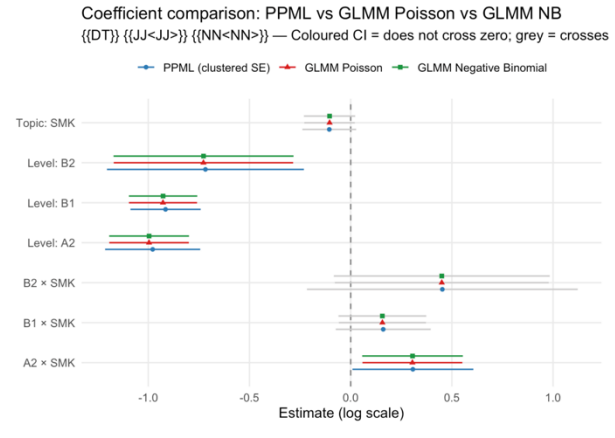

Figure 1: Coefficient comparison across three estimators for M1 {{DT}} {{JJ<JJ>}} {{NN<NN>}}. Points are estimates; lines are 95% CIs.

This convergence extends specifically to the choice of distributional family: as Figure 1 shows for M1, GLMM Negative Binomial estimates track PPML and GLMM Poisson closely even where genuine overdispersion is present. This holds across the full bundle set: across all six bundles and all level-effect terms, GLMM Poisson and GLMM Negative Binomial point estimates differ by no more than 0.007 on the log scale, with a mean absolute difference of 0.002 (full coefficient tables in Supplementary Materials). This behaviour neatly demonstrates the core logic behind PPML: provided the conditional mean is specified correctly, coefficient estimates do not hinge on whether the true distribution is strictly Poisson.

4.2 Developmental Trajectories

The PPML models reveal a coherent pattern of evaluative register routing. Four trajectory types emerge across the six valid models, each illustrated below with authentic instances drawn from the tagged corpus and retrieved via the MDA Tagger's concordancing function (Section 3.4). Fitted on the zero-inclusive dataset, point estimates and standard errors are larger throughout than an earlier, incompletely specified iteration of this analysis reported, since the corrected models estimate the incidence rate of each bundle across the entire writer population, including writers who never produce it, rather than the rate conditional on already having done so at least once.

Because the PPML model uses a log link, coefficients are on the log incidence-rate scale: exponentiating a coefficient gives the multiplicative change in the expected bundle rate relative to the NS reference level. For M1, for example, the three learner-level coefficients (A2: b = -0.978; B1: b = -0.915; B2: b = -0.718) exponentiate to approximately 0.38, 0.40, and 0.49, meaning the expected rate of this bundle is roughly 62% lower than the NS baseline at A2, 60% lower at B1, and 51% lower at B2 (cf. Figure 2a). Equally, the coefficients for an overuse feature bundle like M2 (A2 = 1.10; B1 = 1.14; B2 = 1.01) correspond to expected higher usage rates of approximately 300%, 313%, and 275% respectively (Figure 2b).

4.2.1 Partial recovery (M1)

The feature bundle for M1 is the most distinctive on the weighted log-odds selection statistic introduced in Section 3.2 (-5.68). Underuse is significant at all three learner levels, narrowing steadily from A2 to B2: A2 (b = -0.978, p.adj < .001), B1 (b = -0.915, p.adj < .001), B2 (b = -0.718, p.adj=.009), but it remains significant even at B2. Concordance inspection reveals the mechanism: the dominant JPN realisation is discourse-organisational rather than evaluative, as shown in (3):

(3) *I have two reasons.* ***The first reason*** *is that I think smoker... (W_JPN_SMK_A2_0_066)*

Here, first functions as an ordinal sequencer. Native-speaker writers occupying the equivalent structural slot instead produce genuinely evaluative noun phrases, as in (4):

(4) *..but they might be* ***a little bit*** *better at controlling their bank... (W_ENS_PTJ_XX1_062)*

4.2.2 Sustained overuse (M2, M4, M6)

Instead of tapering off as learners expand their argumentative repertoire, overuse remains essentially flat right through B2. M2 ({{.}} {{RB<RB>}} {{,}}, A2: b = 1.10; B1: b = 1.14; B2: b = 1.01; all p.adj < .001) is driven by discourse sequencing formulae such as (5):

(5) *...many merit of supplementing it.* ***So*** *, I agree to have a... (W_JPN_PTJ_A2_0_121)*

M4, the first-person suasive bundle, shows the largest overuse magnitude of any bundle (A2: b = 1.86; B1: b = 2.17; B2: b = 1.88; all p.adj < .001), driven almost entirely by a single formula positioned overwhelmingly at an essay's opening, illustrated in (6):

(6) ***I agree with*** *this statement... (W_JPN_PTJ_A2_0_114)*

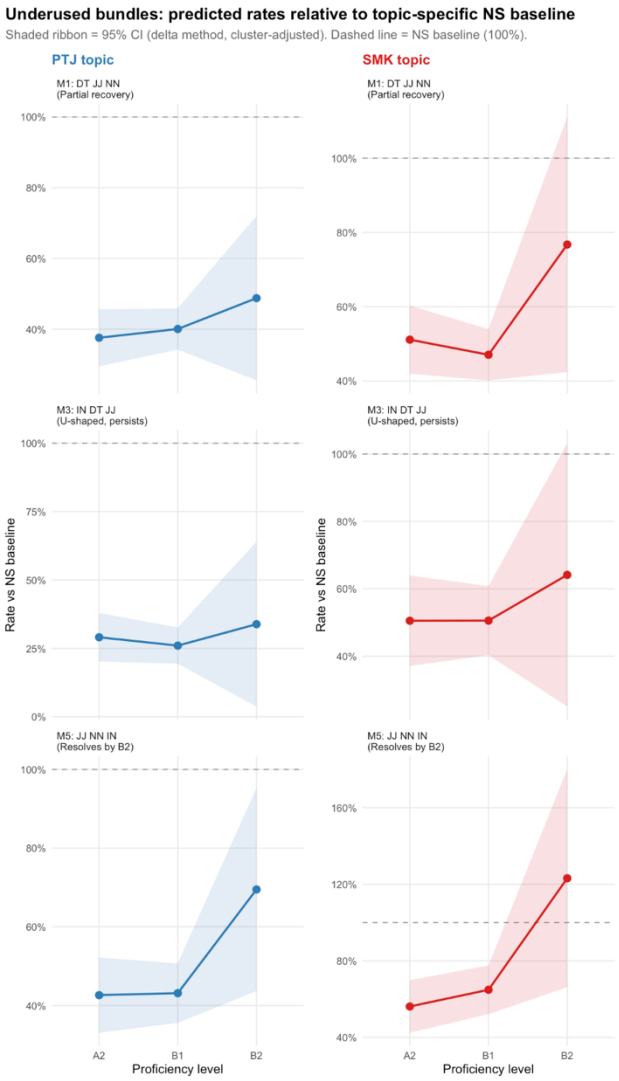

Figure 2a: Developmental trajectories of underused bundles relative to topic-specific NS baseline

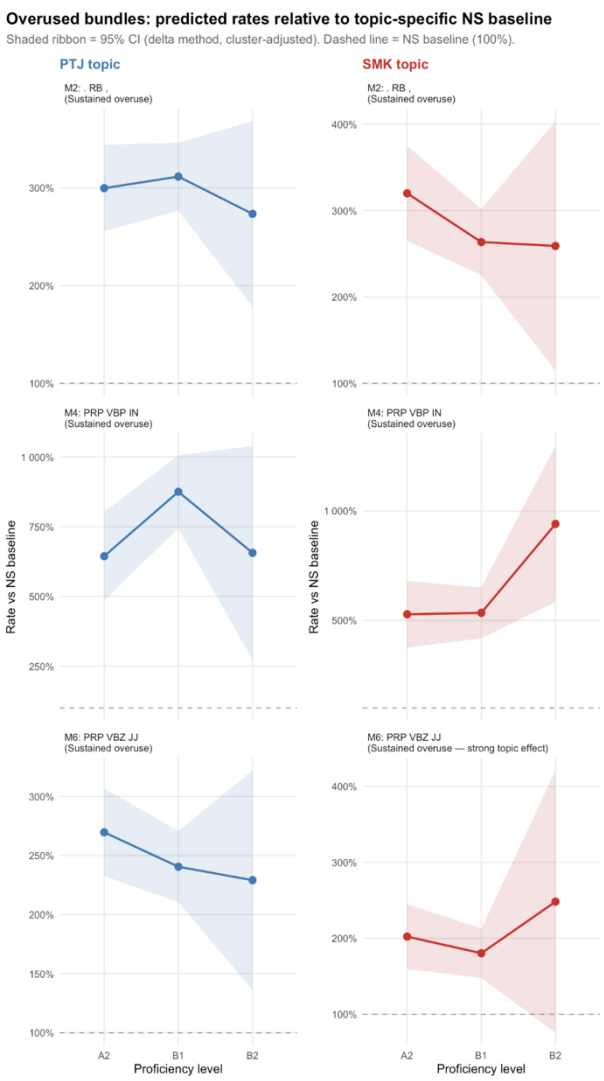

Figure 2b: Developmental trajectories of overused bundles relative to topic-specific NS baseline

M6, the predicative adjective frame as in (7), likewise shows sustained overuse (A2: b = 0.99; B1: b = 0.88; B2: b = 0.83; all p.adj < .001) and, uniquely among the six bundles, a strong topic effect (clustered Wald $\chi^2 = 23.81$, $p < .001$; full restricted-vs-full model comparison in Supplementary Materials).

(7) *I agree that **it is important** for college students to have... (W_JPN_PTJ_A2_0_063)*

4.2.3 U-shaped, persisting (M3)

M3 has the deepest underuse of any bundle in the set, widening at B1 before a partial B2 recovery that leaves the deficit still significant: A2 (b = -1.24, p.adj < .001), B1 (b = -1.35, p.adj < .001), B2 (b = -1.08, p.adj = .026). Significant topic interactions (A2xSMK, p.adj = .024; B1xSMK, p.adj = .001) indicate this U-shape is itself topic-conditioned. (8) shows an example of M3 from the SMK prompt response.

(8) *...when some people aresmoking **in the same** restaurants. In such a... (W_JPN_SMK_A2_0_063)*

4.2.4 No Longer Detectable by B2 (M5)

Strong, essentially flat underuse at A2 (b = -0.852, p.adj < .001) and B1 (b = -0.841, p.adj < .001), with the difference no longer statistically detectable by B2 (b = -0.364, p.adj = .099): the model fails to reject equality with the NS baseline at this level, which indicates insufficient evidence of a difference rather than established equivalence. This is nonetheless the only bundle among the six for which B2 learners are not statistically distinguishable from NS writers at conventional thresholds. Where learners do produce this pattern, as in (9), the evaluative adjective is followed by a content noun and a prepositional complement, which may be considered a structurally demanding configuration:

(9) *I think that the most **important thing for** a college student and a... (W_JPN_PTJ_B1_1_172)*

Figures 2a and 2b visualise the trajectory types as predicted rates relative to topic-specific NS baselines, with the dashed, line marking the 100% NS baseline as a visual reference: a trajectory whose confidence interval overlaps this line indicates the corresponding learner estimate is not statistically distinguishable from the NS reference at that level, while one whose interval remains clearly separated indicates a persisting, if narrowing, gap. Figure 2a shows this directly for the three underused bundles: M5's confidence interval overlaps the baseline by B2, while M1 and M3 both approach it without overlapping, visually confirming the partial-recovery and no-longer-detectable-by-B2 labels assigned above.

Figure 2b shows the three overused bundles remaining well above the baseline at every proficiency level, with M4's markedly larger y-axis scale, extending toward 1,000%,

visually foregrounding the disproportionate magnitude of its overuse relative to M2 and M6. The two M6 plots also make its topic-conditioned trajectory directly visible: a shallow decline across proficiency in the PTJ topic against a marked rise toward B2 in the SMK topic, a shape difference not apparent from the coefficient table alone. Across all six trajectory plots, the confidence ribbons widen visibly toward B2, reflecting the reduced precision of the smaller B2 subsample, a pattern also evident in the wider holdout-partition estimates discussed in Section 4.4.

Considered together, this is a stronger and more surprising finding than trajectory heterogeneity alone would suggest: five of the six bundles show a register gap that persists essentially undiminished, or, for the U-shaped bundle, only partially resolved, through the most advanced proficiency level studied. M5 stands alone as the one bundle for which B2 learners are no longer statistically distinguishable from NS writers, making it a natural focus for future work on what specifically differentiates its developmental trajectory from the other five.

4.3 Concordance-Based Lexical Realisation Analysis

Table 3 reports type-token ratio and normalised entropy (H_norm) for all six modelled bundles and the excluded nominalised NP frame, across JPN and NS sub-corpora, computed across the complete hit set before random sampling.

Table 3: Type-token ratio and normalised entropy by bundle and group, plus the excluded nominalised NP frame

Bundle	Group	Types	Tokens	TTR	H_norm
M1 DT JJ NN	JPN	541	939	0.576	0.924
	ENS	758	1074	0.706	0.948
M2 . RB ,	JPN	71	1283	0.055	0.640
	ENS	51	211	0.242	0.842
M3 IN DT JJ	JPN	193	355	0.544	0.905
	ENS	299	456	0.656	0.949
M4 SUAV PIN	JPN	4	360	0.011	0.062
	ENS	1	26	0.038	0.000
M5 JJ NN IN	JPN	318	444	0.716	0.954
	ENS	339	410	0.827	0.978
M6 it is JJ	JPN	139	753	0.185	0.618
	ENS	60	160	0.375	0.812
DT JJ NOMZ	JPN	47	53	0.887	0.981
	ENS	105	136	0.772	0.940

The H_norm values reveal a spectrum from near-total formulaic lock to genuine productive variety. M4 (JPN H_norm = 0.062) sits at the extreme formulaic end: across 360 JPN occurrences only four distinct realisations occur, with I agree with accounting for approximately 98.6% of JPN hits (Figure 3; full concordance data in Supplementary Materials). NS usage is, if anything, more formulaically concentrated still: all 26 NS occurrences instantiate the identical single realisation, I agree with (H_norm = 0 by definition, given only one type). This shared lock nonetheless has a different status for each group: at nearly 14 times the JPN rate, I agree with functions as a genuinely canonical, load-bearing opening formula for learners, whereas its low absolute frequency among NS writers suggests the opposite: a construction rarely reached for at all, implying NS writers more typically signal agreement through constructions that fall outside this specific tag-bundle definition. The formulaic concentration observed for NS is therefore better read as evidence of this bundle's marginal status in L1 English speakers' usage than as a genuine parallel to the learner pattern.

The productive pole of the spectrum is equally instructive, and concerns a bundle excluded from the regression analysis itself. The nominalised NP frame {{DT}} {{JJ<JJ>}} {{NN<NOMZ>}}, excluded from the confirmatory PPML models under the corrected DP criterion (Section 3.5), shows a markedly different concordance profile from the six regression bundles: H_norm = 0.981 in JPN, with nearly every occurrence instantiating a distinct adjective-nominalisation pairing, as in (9):

(10) *...do part time job is **a good opportunity** of experiencing communities... (W_JPN_PTJ_A2_0_051)*

alongside single-occurrence types such as a serious situation and the financial condition. This near-uniform distribution independently corroborates the writer-level dispersion evidence that excluded this bundle from the regression analysis: a small number of learners produce it, and do so productively rather than formulaically; this reflects the same underlying pattern that a high, DP-failing dispersion score and a high lexical entropy score are both, from different angles, detecting.

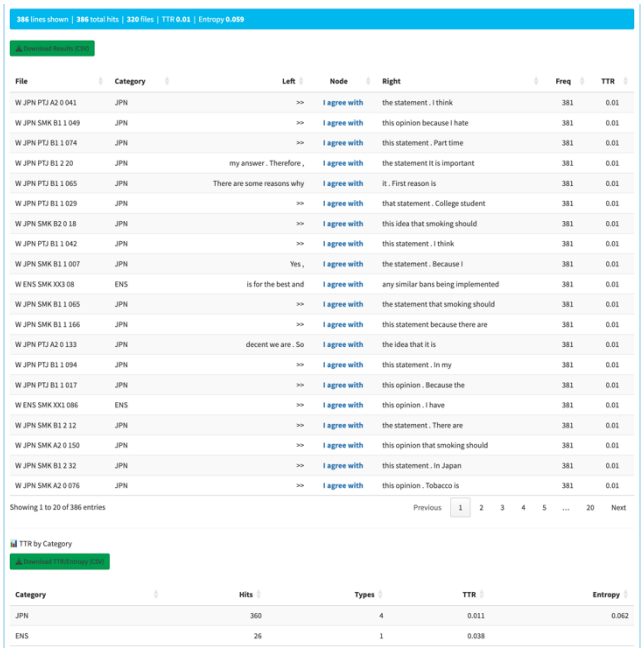

Figure 3: Screenshot of KWIC concordance for M4, JPN and ENS combined, sorted by frequency, illustrating near-total formulaic lock across both populations.

Both analyses are sensitive to the long-tailed structure of the data, but they address different questions: PPML models the conditional mean of bundle incidence, whereas H_norm describes the concentration of a bundle's lexical realisations conditional on occurrence. The two measures are not formal analogues of one another, though both reflect the same underlying Zipfian skew that motivates careful treatment of frequency data throughout this paper.

M1 further illustrates discourse organisational hijacking: despite relatively high H_norm (JPN 0.924, NS 0.948), the top JPN realisations are discourse connectives, ordinal sequencers such as The first reason and The second reason, and relational phrases such as the other hand, where the adjective slot carries organisational rather than evaluative meaning. In the NS concordance, the equivalent structural positions are more often occupied by genuine evaluative

NPs, such as a good idea and the real world. The comparatively small H_norm gap between groups (0.924 vs 0.948) understates this qualitative divergence, since discourse connectives cluster at essay-structural positions rather than being spread evenly across the text, a distributional pattern that aggregate entropy, computed across the full hit set, does not directly capture.

The remaining bundles occupy intermediate positions on the spectrum. M2 (JPN H_norm = 0.64) has a small type inventory used with moderate, not extreme, concentration, unlike M4's near-total lock, with NS again showing greater diversity (H_norm = 0.842). M6 (JPN H_norm = 0.618) is similarly semi-formulaic, intermediate between M4's extreme concentration and the more productive bundles, with NS again showing higher diversity (H_norm = 0.812). M3 and M5, the two underused bundles not otherwise discussed in detail, both show high H_norm in JPN (0.905 and 0.954 respectively) despite their low corpus frequency, indicating that on the occasions learners do produce them, they do so with genuine lexical variety. This supports the Section 5.2 interpretation that underuse for these bundles reflects rarity of deployment or syntactic difficulty rather than formulaic avoidance.

4.4 Robustness to the Selection Procedure

The direction of every coefficient for the feature bundle M1 is preserved between the full-sample and holdout-only models, and all three terms remain significant in both. Underuse at A2 intensifies from b = -0.978 in the full sample to b = -1.28 in the holdout (both p.adj < .001); underuse at B1 remains significant though somewhat attenuated (b = -0.915 full vs b = -0.848 holdout, both p.adj < .001); and underuse at B2, significant in the full sample (b = -0.718, p.adj = .004), is estimated with a larger coefficient in this particular holdout partition (b = -0.872, p.adj = .007). All three level effects therefore replicate in both sign and statistical significance on this independent 20% partition of the writer population.

Across the full set of six models, 39 of 42 level-effect coefficients (92.9%) preserve the same sign between the full-sample and holdout-only estimates. All three sign changes occur in level × topic interaction terms that were already non-significant in the full sample (adjusted p-values of .853, .853, and .668); none occur in a main effect, and none touch a coefficient that features in the trajectory typology reported above. This provides evidence that the principal trajectory findings are not highly sensitive to this particular writer partition, consistent with the theoretical argument in Section 3.2 that theory-driven selection mitigates the circularity risk of data-driven dependent variable choice.

5. Discussion

5.1 Methodological Implications for Corpus Linguistics

The introduction of PPML to corpus linguistics through the Zipf-gravity analogy carries implications beyond the specific analysis reported here. The Jensen's inequality argument establishes that log-linearisation is not merely suboptimal but formally biased under the distributional conditions that characterise corpus count data, concentrated in the Zipf long tail, precisely the region containing the most theoretically interesting bundles. This theoretical case for PPML is independent of the GLMM comparison reported in Section 4.1: PPML provides a distributionally flexible alternative to log-linearisation regardless of whether a comparable mixed-effects specification converges, and its consistency does not depend on correctly identifying a random-effects variance structure.

These findings bear on a methodological debate that extends well beyond the present dataset. GLMMs have become the default tool in corpus linguistics precisely because they promise to model the nested structure of writer-level data explicitly, in contrast to the implicit corrections offered by clustered standard errors. The analysis in Section 4.1 shows this promise is not, in general, incompatible with Zipfian bundle-frequency data: GLMM converges cleanly once the modelling data is correctly specified as a zero-inclusive dataset, a precondition that is easily overlooked given how the bundle-extraction routine described in Section 3.1 generates output only where a bundle actually occurs. When GLMMs throw singular fit warnings on corpus data, researchers should not immediately assume the count distribution is incompatible with random effects. Frequently, the explanation is simpler: the dataset is missing explicit zero counts for files where a bundle never occurred.

5.2 Substantive Implications for Register Development

The four developmental trajectories reveal that register convergence is neither uniform nor unidirectional across feature types, a finding that complicates simple linear models of second language register acquisition: five of the six bundles show divergence from the NS baseline that remains significant through B2, with M5 the sole bundle for which the difference is no longer statistically detectable by B2. The sustained overuse of M2, M4, and M6 suggests that dependence on a small number of essay-opening or predicative formulae does not recede within the CEFR range studied, in tension with restructuring accounts of interlanguage development in which surface-level overuse of an early-acquired form gives way to more varied production as competing forms become available (Kellerman, 1985; Lightbown, 1983). Whether this reflects a genre convention robust to proficiency development, or a CEFR ceiling below the point at which restructuring would occur, cannot be determined from the A2-B2 range examined here.

The partial recovery of M1 and the persisting U-shape of M3, by contrast, indicate that syntactically more demanding features of the target register narrow their gap from the NS baseline without closing it within the range studied. The U-shape observed for M3 is of particular theoretical interest: rather than a simple monotonic approach to the target norm, underuse of the prepositional evaluative NP intensifies at B1 before a partial B2 recovery that leaves the deficit still significant. This is broadly consistent with a developmental account in which intermediate learners, having newly acquired the capacity for more elaborated clausal organisation, including the very discourse-sequencing adverbials responsible for the overuse pattern in M2, temporarily reallocate processing resources away from NP-internal modification, a transfer-of-processing-load effect broadly compatible with usage-based accounts of L2 syntactic development (Kyle and Crossley, 2017). M5 stands apart as the only bundle among the six for which B2 learners are no longer statistically distinguishable from NS writers, despite showing underuse of similar magnitude to M1 and M3 at A2 and B1. What distinguishes this pattern from the other five bundles' sustained or only partially narrowing divergence is not established by the present analysis and is flagged as a specific target for future work. Together, these patterns suggest that the register routing

account developed in Section 2.2 is not a static description of learner divergence from a target norm but a genuinely developmental phenomenon, with the specific developmental path varying by bundle in ways an averaged, register-wide divergence score would obscure.

5.3 Implications for Instruction and Assessment

These trajectories carry direct implications for register-aware instruction, extending the pedagogical sequencing proposed by Sawaguchi (2024) for referential and stance bundles in argumentative writing. Sawaguchi's teaching-order recommendations, introducing basic referential and objective stance bundles at A2/B1 and reserving advanced referential bundles for B2, presuppose that referential bundle acquisition proceeds incrementally across this range. The present findings complicate this assumption for NP-internal adjectival modification specifically: the persisting deficits of M1 and M3 indicate that elaborated referential features do not close the gap with NS writers within the CEFR range studied, suggesting explicit instruction targeting these bundle types may need to persist, or intensify, beyond B1 rather than being front-loaded. A concrete application follows directly from the MDA Tagger's concordancing output: a consciousness-raising task built around the M1 concordance lines in Figure 4, asking learners to sort JPN and NS realisations by function (organisational sequencer versus evaluative noun phrase) rather than by grammatical form alone, would make the underlying misalignment directly noticeable in a way frequency counts cannot.

The concordance-based evidence further suggests instruction should distinguish structural from functional acquisition. Sitler and Spivey (2025) report, using a GLMM specification, that overall bundle frequency decreases as proficiency increases; the present analysis finds a more heterogeneous pattern: frequency reduction only for M5, sustained frequency for the other five bundles. This finding suggests that feature bundle-level analysis may be necessary to characterise developmental patterns that a single aggregate frequency trend may obscure. Where formulaic overuse does persist, as for M2, M4, and M6, the fade-out that might be expected from restructuring accounts is not observed within the normalised entropy data either (Section 4.3): low lexical diversity in these bundles' realisations persists alongside their sustained frequency, meaning learners are not simply producing the same narrow lexical range less often, but continuing to produce it at a stable rate. This indicates that pedagogical intervention for these bundles should target lexical range directly, rather than anticipating that increased proficiency alone will diversify a structural pattern's realisation.

5.4 Limitations

Several limitations qualify these findings. First, the analysis is restricted to two argumentative topics within a single learner corpus; the theory-driven selection and holdout procedures (Sections 3.2, 4.4) mitigate concerns specific to this dataset, but the trajectories require replication across further topics and populations. Second, the CEFR range studied (A2-B2) does not extend to C1/C2, leaving open whether the sustained divergence observed for M2, M3, M4, and M6 eventually resolves at higher proficiency. Third, normalised entropy is sensitive to sample size at low frequencies; estimates for bundles with fewer than approximately 30 tokens, including the nominalised NP bundle excluded under the corrected DP criterion (Section 3.5) but discussed qualitatively in Section 4.3, should be interpreted with caution. Fourth, the single-partition holdout check does not constitute a full cross-validation; a k-fold procedure would give a more complete picture of estimate stability. Fifth, the concordance analysis examined only the top five realisations of each bundle; a fuller functional classification, along the lines of Sitler and Spivey's (2025) taxonomy, would strengthen the qualitative account offered here. Finally, the GLMM comparison reported in Section 4.1 uses frequentist estimation (lme4::glmer()), rather than the Bayesian estimation Winter and Bürkner (2021) specifically recommend. A preliminary Bayesian comparison on the wh-suasive frame, the bundle most affected by separation, confirms that weakly informative priors resolve the estimation failure frequentist methods encountered there, a form of regularisation distinct in kind from the log(count+1) practice critiqued in Section 2.1 (Supplementary Materials); a full Bayesian replication across all six modelled bundles remains beyond the present scope but would be a natural extension.

6. Conclusion

This paper has introduced PPML to corpus linguistics through the Zipf-gravity analogy, offering a theoretical case grounded in Jensen's inequality, together with a proof-of-concept analysis, tagged, bundle-extracted, and concordanced throughout using the MDA Tagger, identifying evaluative register routing in Japanese EFL writing across four developmental trajectory types, five of six bundles showing divergence from the NS baseline persisting largely undiminished through the most advanced proficiency level studied. Comparison with GLMM, correctly specified on a zero-inclusive dataset, converges with PPML on every substantive conclusion, while the theory-driven tupling procedure, validated by a holdout robustness check, provides a template for principled DV selection that separates operationalisation from inference. The normalised entropy measure further reveals that sustained frequency in the overused bundles is not accompanied by diversification in lexical realisation, a finding with direct implications for register-aware writing pedagogy. ICNALE's balanced design is unusually favourable to random-effects estimation, and the present results should be read accordingly: they demonstrate that PPML is a viable, theoretically well-grounded mean-model approach for sparse bundle-frequency data, converging with correctly specified GLMMs under conditions that, if anything, favour the mixed-effects alternative. Whether PPML's consistency advantage becomes practically decisive, rather than merely theoretically available, is a harder question this conservative test cannot itself resolve; the ongoing BTL project, with its considerably less regular writer-level structure, is the natural setting in which to test this directly. Future work should extend this theory-driven, entropy-informed pipeline to that corpus and to additional proficiency ranges, and examine whether the register routing pattern identified here generalises to other evaluative registers beyond argumentative essay writing.

Supplementary Materials

Supplementary materials mentioned in Sections 3.5, 4.1, 4.2, 4.3 can be found at: https://timmarchand.github.io/mda_tagger/supplementary/index.html

Bibliographical References

Baayen, R. H. (2001). Word Frequency Distributions. Kluwer.

Baayen, R. H. (2008). Analyzing Linguistic Data: A Practical Introduction to Statistics Using R. Cambridge University Press.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), 1-48.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate. Journal of the Royal Statistical Society: Series B, 57(1), 289-300.

Bergé, L. (2018). Efficient estimation of maximum likelihood models with multiple fixed-effects: The R package FENmlm. CREA Discussion Paper 13.

Biber, D. (1988). Variation across speech and writing. Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). Longman Grammar of Spoken and Written English. Longman.

Church, K. W. and Gale, W. A. (1995). Poisson mixtures. Natural Language Engineering, 1, 163-190.

Ellis, N. C., Simpson-Vlach, R., and Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. TESOL Quarterly, 42(3), 375-396.

Evert, S. (2005). The Statistics of Word Cooccurrences: Word Pairs and Collocations. PhD thesis, Institut für Maschinelle Sprachverarbeitung, University of Stuttgart.

Granger, S. and Rayson, P. (1998). Automatic profiling of learner texts. In S. Granger (Ed.), Learner English on Computer (pp. 119-131). Longman.

Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. International Journal of Corpus Linguistics, 13(4), 403-437.

Gries, S. T. (2015). Some current quantitative problems in corpus linguistics and a sketch of some solutions. Language and Linguistics, 16(1), 93-117.

Gries, S. T. (2021). Statistics for Linguistics with R: A Practical Introduction (3rd ed.). De Gruyter Mouton.

Gries, S. T. (2024). Frequency, Dispersion, Association, and Keyness: Revising and tupleizing corpus-linguistic measures. Benjamins.

Ha, Y. (2022). Syntactic complexity in EFL writing: Within-genre topic and proficiency effects. CALL-EJ, 23(1), 187-205.

Hunston, S. and Francis, G. (2000). Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English. Benjamins.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. Learner Corpus Studies in Asia and the World, 1, 91-118.

Kellerman, E. (1985). If at first you do succeed... In S. M. Gass and C. G. Madden (Eds.), Input in Second Language Acquisition (pp. 345-353). Newbury House.

Kyle, K. and Crossley, S. (2017). Assessing syntactic sophistication in L2 writing. Language Testing, 34(4), 513-535.

Larsson, T., Paquot, M., and Biber, D. (2021). On the importance of register in learner writing. In E. Seoane and D. Biber (Eds.), Corpus based approaches to register variation (pp. 235-258). Benjamins.

Lightbown, P. M. (1983). Exploring relationships between developmental and instructional sequences in L2 acquisition. In H. W. Seliger and M. H. Long (Eds.), Classroom Oriented Research in Second Language Acquisition (pp. 217-245). Newbury House.

Marchand, T. (2026). MDA Tagger: A Multi-Dimensional Analysis toolkit [Software]. https://github.com/timmarchand/mda_tagger

Monroe, B., Colaresi, M., and Quinn, K. (2008). Fightin' words. Political Analysis, 16(4), 372-403.

Muşlu, M. (2018). Use of stance lexical bundles by Turkish and Japanese EFL learners and native English speakers in academic writing. Gaziantep University Journal of Social Sciences, 17(4), 1319-1336.

O'Hara, R. B. and Kotze, D. J. (2010). Do not log-transform count data. Methods in Ecology and Evolution, 1(2), 118-122.

Sawaguchi, R. (2024). Potential of L1 and L2 corpora to identify target lexical bundles for argumentative essay writing. Asia Pacific Journal of Corpus Research, 5(1), 1-21.

Silva, J. M. C. S. and Tenreyro, S. (2006). The log of gravity. The Review of Economics and Statistics, 88(4), 641-658.

Sitler, T. and Spivey, M. (2025). “I agree with this”: A multi-modal register analysis of lexical bundles in the ICNALE. Asia Pacific Journal of Corpus Research, 6(1), 25-45.

Tinbergen, J. (1962). Shaping the world economy. Twentieth Century Fund.

Winter, B. and Bürkner, P.-C. (2021). Poisson regression for linguists: A tutorial introduction to modelling count data with brms. Language and Linguistics Compass, 15(11), e12439.

Zipf, G. K. (1935). The psycho-biology of language. Houghton Mifflin.

Zipf, G. K. (1949). Human Behavior and the Principle of Least Effort. Addison-Wesley.