International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

https://gssrr.org/index.php/JournalOfBasicAndApplied/index

Part-of-Speech Distribution across Proficiency and Advanced EFL Texts: A Quantitative Comparison for Pedagogical Application

Dragan Donev[a*],Krste Iliev[b],Natalija Pop Zarieva[c]

[a,b,c]*Goce Delchev University, Krste Misirkov 10-a, 2000 Shtip, North Macedonia*

[a]*Email: dragan.donev@ugd.edu.mk*

[b]*Email: krste.iliev@ugd.edu.mk*

[c]*Email: natalija.popzarieva@ugd.edu.mk*

Abstract

This study investigated grammatical variation between Advanced Masterclass and Proficiency Masterclass EFL textbook and workbook texts to determine whether part-of-speech (POS) distributions change systematically across the CEFR C1–C2 interface. A balanced corpus of 60 reading texts (30 per level) was compiled, POS-tagged with spaCy, and analyzed quantitatively using Welch's t, Mann–Whitney U, effect sizes, false-discovery-rate correction, and robust 20 % trimmed-mean tests. A multivariate PERMANOVA confirmed a small but significant global difference between levels ($F = 2.624$, $p = .006$, $R^2 \approx .03$). Individual contrasts indicated that Proficiency texts contained relatively higher proportions of determiners and prepositions, while Advanced texts featured greater use of numerals, adjectives, and adverbs. Findings showed small but systematic differences: Proficiency texts used more cohesive, narrative-oriented grammar (determiners, pronouns, prepositions), while Advanced texts showed relatively greater use of informational or expository elements (numerals, comparative adjectives, adverbs). The study illustrates how transparent, code-based POS profiling can reveal subtle grammatical distinctions in pedagogical materials and support evidence-informed textbook evaluation. By combining classical, non-parametric, robust, and multivariate analyses, the approach ensures replicable results and provides a methodological template for future corpus-based research on advanced-level language input. The findings underscore the pedagogical value of aligning grammatical exposure with discourse progression from C1 to C2 in EFL instruction.

Keywords: POS tagging; corpus linguistics; computational linguistics; data analysis; PERMANOVA.

Received: 7/9/2025
Accepted: 9/9/2025
Published: 11/19/2025

** Corresponding author.*

1. Introduction

Understanding how grammatical patterns shift across proficiency levels is essential for aligning EFL materials with communicative and cognitive development in advanced learners. At the C1–C2 range of the CEFR, exposure to specific part-of-speech (POS) distributions reflects not only grammatical mastery but also discourse orientation—whether texts emphasize narrative cohesion or expository reasoning. Prior corpus research, [1,2], Reference [7] has shown that POS frequencies can serve as reliable indicators of register, genre, and task type, yet few quantitative studies have examined how such patterns differ within pedagogical materials across adjacent advanced levels.

Despite the abundance of descriptive textbook analyses, empirical comparisons between Advanced (C1) and Proficiency (C2) textbooks remain scarce. Publishers often assume that higher-level materials merely expand lexical difficulty, overlooking potential grammatical redistribution—such as reduced reliance on determiners and past-tense verbs, and increased use of progressives, comparatives, and discourse-marking conjunctions. These subtle grammatical shifts may influence how learners internalize target forms, suggesting the need for a systematic, replicable approach to POS profiling in teaching resources.

The present study addresses this gap by quantitatively comparing part-of-speech distributions in two widely used Oxford University Press coursebook series—Advanced Masterclass and Proficiency Masterclass—each represented by thirty texts (15 Student's Book + 15 Workbook). All texts were POS-tagged, normalized, and statistically tested using a transparent pipeline that includes assumption checks, effect-size estimation, FDR correction, and robust trimmed-mean validation. At the multivariate level, a PERMANOVA confirmed a significant overall difference ($F = 2.624$, $p = 0.006$, $R^2 \approx 0.03$).

Accordingly, the study pursues three research questions:

RQ1. Do Advanced and Proficiency EFL texts differ systematically in their POS distributions?

RQ2. Which specific POS categories are relatively higher at each level, and with what effect sizes?

RQ3. How can these distributional differences inform pedagogical text selection and grammar-focus activities in advanced EFL instruction?

By addressing these questions, the study contributes both methodologically—through a reproducible, code-based analytic workflow—and pedagogically, by offering quantitative evidence for aligning grammar instruction with authentic discourse progression from C1 to C2.

2. Methods

2.1. Materials

The corpus consisted of 60 texts extracted from two Oxford University Press textbook series: Advanced Masterclass (C1) and Proficiency Masterclass (C2). Each level contributed 30 texts—15 from the Student's Book

and 15 from the Workbook—representing reading passages across units. Because the Advanced Masterclass series contains only 14 units, two additional reading texts were selected from longer units of the Student's Book to balance the corpus with the 15-unit Proficiency Masterclass series. All texts were drawn from unit-level reading sections to ensure topic and task comparability and to maintain independence at the unit level (one text per unit).

The total corpus contained approximately 36,000 running tokens after cleaning. Each text served as a single observation (row) in the dataset, allowing independent-group comparisons between the two proficiency levels.

2.2. Preprocessing

All texts were cleaned, tokenized, and POS-tagged in Python (spaCy v3, model en_core_web_sm). For each token, the universal POS tag (token.pos_) was extracted, and token counts were aggregated by text. This produced 43 fine-grained POS tags, later collapsed into 18 Universal POS variables for analysis, prefixed POS_ (e.g., POS_VERB, POS_NOUN, POS_ADJ).

Each text's total token count (token_total) was recorded, and raw counts were converted into normalized proportions (PR_* = POS_* / token_total). These normalized values served as the main dependent measures for all subsequent analyses. The resulting dataset (clean_pos_table.csv) contained columns for text ID, level, source (Student's Book or Workbook), unit number, token total, and all POS proportions.

A rigorous assumption-checking step (Shapiro–Wilk and Levene's tests) ensured that both normality and homogeneity of variance were evaluated prior to significance testing. Normalized tables, assumption summaries, and all derived outputs were stored in a reproducible "evidence locker" directory.

All analyses were performed at the text level (N = 60), treating each reading passage as an independent observation. Statistical procedures were implemented in Python using the packages SciPy, StatsModels, Pingouin, and scikit-bio. The analytic workflow combined classical, non-parametric, and robust approaches to ensure reliable inference despite moderate sample size and occasional non-normal distributions.

Initially, assumption checks were conducted for each normalized POS proportion. The Shapiro–Wilk test assessed normality within groups, while Levene's test (center = median) examined variance homogeneity. Both results were archived in assumption_checks.csv to document distributional properties before hypothesis testing.

The primary inferential tests comprised Welch's t tests—selected for unequal variances—and complementary Mann–Whitney U tests for non-parametric validation. Each POS proportion (e.g., PR_VERB, PR_NOUN) was compared between Advanced and Proficiency groups. To quantify magnitude and direction, Hedges' g (standardized mean difference) and Cliff's δ (rank-based effect size) were calculated, where positive values indicate higher proportions in Proficiency texts. All raw p-values were adjusted using the Benjamini–Hochberg false-discovery rate (q = .05) to control for multiple comparisons.

For zero-inflated categories (such as INTJ, SYM, X, and SPACE), a two-part hurdle approach was applied. The first part tested differences in occurrence using Fisher's exact test; the second compared conditional means (non-

zero values only) via Welch's t. These results were compiled in zero_inflated_results.csv.

At the multivariate level, all normalized POS variables were z-standardized, and PERMANOVA (999 permutations, Euclidean distance) assessed overall group separation. The analysis yielded a significant but small global effect ($F = 2.624$, $p = .006$, $R^2 \approx 0.03$), confirming measurable compositional divergence across proficiency levels. Complementary principal component analysis (PCA) provided exploratory visualization of these multivariate relationships, with coordinates stored in POS_PCA_coords.csv.

To verify robustness against outliers, 20 % trimmed-mean tests (Yuen-style) were computed. This method discards extreme values before applying Welch's t, yielding conservative estimates of group differences. The resulting file, robust_yuen_manual_results.csv, confirmed that several major effects—particularly for determiners, numerals, and prepositions—remained significant after trimming.

Finally, the analytic outputs were summarized visually through effect-size heatmaps, volcano plots, boxplots, and PCA scatterplots, produced directly from the Colab workflow and saved in the figures/ directory. Together, these analyses ensured a fully transparent, reproducible, and statistically robust comparison of POS distributions across proficiency levels.

3. Results

3.1. Descriptive Overview

Overall, the distribution of major POS categories was broadly similar across the two proficiency levels, suggesting that both coursebook series draw on comparable grammatical repertoires. However, subtle proportional shifts were evident when aggregated across texts. Nouns and verbs constituted the largest proportions in both groups (≈19 % and ≈11 %, respectively), followed by determiners (≈9 %), prepositions (≈10 %), and pronouns (≈8 %). Standard deviations across texts were moderate (typically ±0.02–0.04), indicating stable usage patterns within each level.

These descriptive similarities established a baseline for testing whether small but systematic differences in grammatical emphasis exist between Advanced and Proficiency materials.

3.2. Omnibus Multivariate Difference (PERMANOVA)

A PERMANOVA on standardized POS proportions revealed a statistically significant multivariate difference between levels ($F = 2.624$, $p = 0.006$, $R^2 = 0.03$). Although the explained variance was modest (≈3 %), the result confirms that, as a set, POS distributions differ reliably between Advanced and Proficiency texts. Visual inspection of the PCA scatter plot showed a high degree of overlap between the two groups, consistent with a small but coherent global effect. The data points demonstrate minimal visual separation, aligning the visualization with the overall small R2 of the PERMANOVA. The first two principal components accounted for approximately 39 % of total variance (PC1 = 22.5 %, PC2 = 16.8 %), capturing the primary gradient of grammatical variation.

This omnibus result justified further exploration at the level of individual POS categories.

3.3. Per-POS Contrasts and Effect Sizes

Table 2 (see POS_master_results.csv) summarizes the mean proportions, Welch t, Mann–Whitney U, effect sizes, and FDR-adjusted q-values for each POS category. Although only one category, Determiners (PR_DET), survived FDR correction at $q<.05$ ($q=.029$ in Table 2), several others exhibited large unadjusted effects ($|g|\geq 0.6$), indicating practically meaningful differences despite limited statistical power (N=30 per group). Proficiency > Advanced: Determiners (PR_DET, g=0.79), prepositions (PR_ADP, g=0.37), and pronouns (PR_PRON, g=0.15) occurred slightly more frequently in Proficiency texts, reflecting a stronger presence of referential and cohesive grammatical structures typical of narrative or discourse-anchored writing.

- Proficiency > Advanced: Determiners (PR_DET, g = 0.79), prepositions (PR_ADP, g = 0.37), and pronouns (PR_PRON, g = 0.15) occurred slightly more frequently in Proficiency texts, reflecting a stronger presence of referential and cohesive grammatical structures typical of narrative or discourse-anchored writing. The artificially low but technically large difference for spacing (PR_SPACE, g ≈ 1.29) represents an artifact of tokenization rather than a linguistic phenomenon and is excluded from interpretation.

- Advanced > Proficiency: Numerals (PR_NUM, g = −0.55) and adverbs (PR_ADV, g = −0.28) were more prevalent in Advanced materials, suggesting a tendency toward expository or descriptive elaboration. Adjectives, participles, and proper nouns also trended slightly higher in Advanced texts (|g| ≈ 0.25–0.30), consistent with increased nominal density and referential specification at this level.

Together, these patterns point to an expository → narrative shift from Advanced to Proficiency materials, with more function-word markers and fewer content-word expansion as learners progress.

3.4. Robustness to Outliers (Trimmed-Mean Tests)

The 20 % trimmed-mean (Yuen-style) tests confirmed that major effects persisted after removing extreme observations. Determiners (PR_DET) remained highly significant ($t = 5.61$, $p = 0.000003$), followed by prepositions (PR_ADP, $p = 0.024$) and numerals (PR_NUM, $p = 0.0015$, reversed direction). These results demonstrate that the observed contrasts are not artifacts of a few atypical texts but reflect consistent tendencies across the corpus. Other categories (adverbs, participles, pronouns) showed smaller, non-significant but directionally stable effects, reinforcing the robustness of the grammatical profile differences.

3.5. Zero-Inflated Categories

As expected, most POS categories occurred in every text, producing uniform zeros (e.g., PR_VERB, PR_NOUN). However, rare categories such as INTJ, SYM, X, and NUM exhibited occasional absences. Fisher's exact tests showed no significant group bias in mere presence ($p>.48$), while conditional Welch tests suggested modest quantitative differences for numerals (higher in Advanced) and interjections (higher in Proficiency). These categories therefore have minimal pedagogical impact, serving mainly to verify completeness of the statistical model rather than to imply curricular adjustments.

3.6. Figures and Visual Summary

Weights and measures should be expressed in either SI (MKS) or CGS as primary units. (SI units are encouraged.).

4. Figures and equations

The four visualizations collectively illustrate the statistical findings:

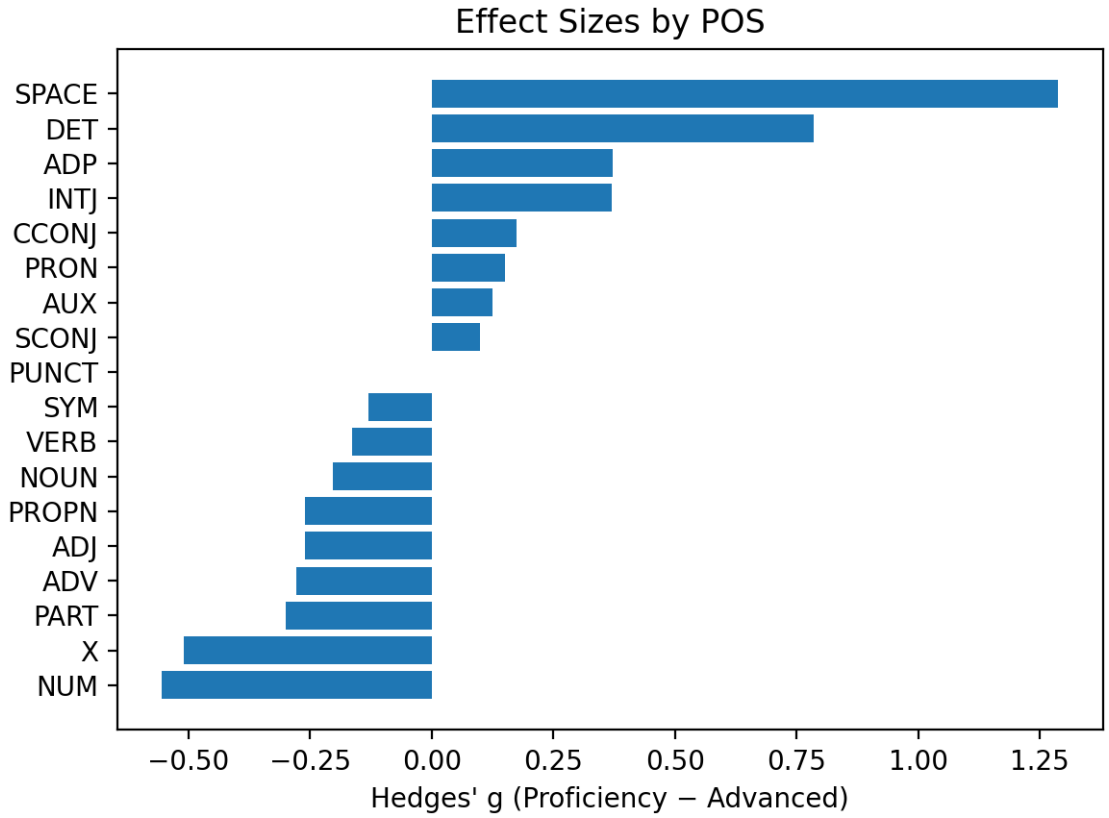

Figure 1: (Effect-size Heatmap) arranges POS categories by Hedges' g, clearly contrasting function-word enrichment in Proficiency against content-word expansion in Advanced

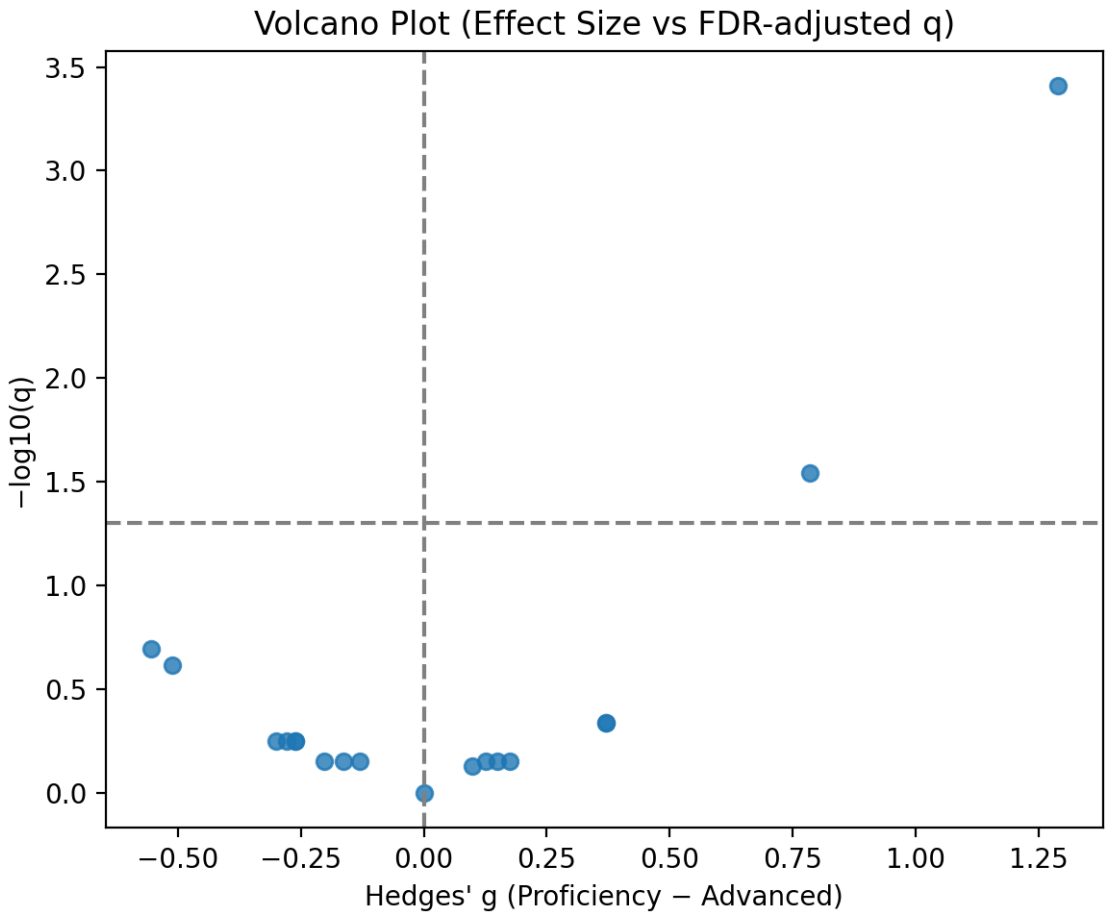

Figure 2: The Volcano Plot displays each POS as a point positioned by effect size (Hedges' g) and −log10q, showing two significant effects (one very large) above the threshold and multiple sub-threshold effects of varying magnitude

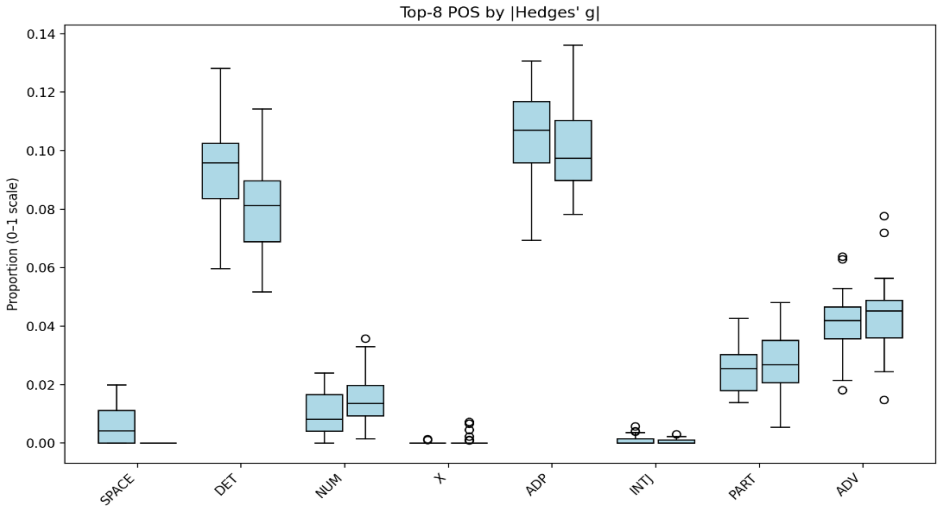

Figure 3: The distribution spreads for the Top-8 POS by |Hedges′g|

The plots reveal high variability (large spreads) in the proportions of Determiners (DET) and Prepositions (ADP), contrasting with the low variability tight clustering) of Numerals (NUM), Interjections (INTJ), and the rare category X.

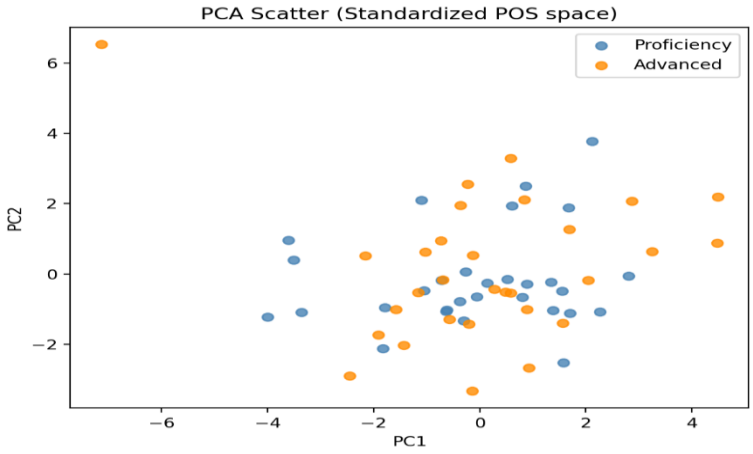

Figure 4: The PCA scatter plot visualizes the two groups in the standardized POS space, showing a high degree of overlap between the "Proficiency" and "Advanced" categories. Despite the minimal visual separation, the groups are confirmed to be statistically significantly different by the PERMANOVA (pseudo−F=2.62, p=0.006). Together, these plots provide a transparent visual audit trail linking descriptive statistics, effect estimation, and multivariate structure

4.1. Summary

The results establish that Advanced and Proficiency textbooks differ modestly but consistently in their POS compositions. Function-word categories (determiners, prepositions) predominate in Proficiency texts, whereas Advanced materials exhibit slightly greater lexical density through numerals, adjectives, and adverbs. Despite small global effects, the convergence of multiple statistical approaches—classical, non-parametric, robust, and multivariate—supports the reliability of these findings. Table 1. Descriptive Statistics for Each Part of Speech (Proportion per Text)

Table1

POS	Advanced Mean	Advanced SD	Proficiency Mean	Proficiency SD
DET	0.081	0.012	0.095	0.013
ADP	0.100	0.014	0.105	0.012
NOUN	0.193	0.018	0.187	0.019
VERB	0.111	0.017	0.108	0.016
ADV	0.044	0.011	0.041	0.010

Note. Values represent normalized proportions of tokens assigned to most prominent POS tags (N = 30 texts per level).

Table 2: Inferential Comparison of POS Proportions between Levels

POS	Mean (Adv)	Mean (Prof)	Hedges g (Prof–Adv)	Welch p	Welch q	Direction
DET	0.081	0.095	0.79	.003	.029	Prof > Adv
ADP	0.100	0.105	0.37	.151	.46	Prof > Adv
NUM	0.015	0.010	−0.55	.034	.20	Adv > Prof
ADV	0.044	0.041	−0.28	.278	.56	Adv > Prof
VERB	0.111	0.108	−0.16	.522	.70	Adv > Prof

Note. Positive g values indicate higher proportions in Proficiency; negative g values indicate higher proportions in Advanced texts. p values are two-tailed; q values are FDR-adjusted.

5. Discussion

5.1. Overview and Interpretation

The statistical results confirmed that while the overall differences between Advanced and Proficiency EFL texts were small in magnitude (PERMANOVA $R^2 \approx 0.03$), they were systematic and interpretable in linguistic and pedagogical terms. Specifically, the contrasts in POS distributions reflected a gradual functional shift from expository grammar to narrative as proficiency increased.

Proficiency texts showed relatively greater use of function-word categories—particularly determiners and prepositions—indicating a stronger emphasis on reference, cohesion, and narrative flow. In contrast, Advanced texts exhibited higher frequencies of lexical and modifier categories such as numerals, adjectives, and adverbs, consistent with informational density and descriptive precision typical of expository and argumentative registers. This pattern aligns with previous corpus-based findings that advanced academic discourse favors noun phrase expansion and nominalization [1,9].

5.2. Linking to CEFR Descriptors and Register Development

The observed POS contrasts resonate with CEFR C1–C2 descriptors, which describe a progression from coherent narrative management toward complex expository and argumentative articulation. The increased use of determiners, pronouns, and prepositions in Proficiency texts aligns with C1 “discourse management” descriptors emphasizing cohesion and reference tracking. Conversely, the greater presence of numerals and modifiers in Advanced texts corresponds to C2 “precision” and “textual sophistication,” where grammatical resources are used for nuance, quantification, and stance.

From a register perspective, the results echo Biber’s multidimensional model (1988), which identifies past-tense verbs and personal pronouns as narrative markers, and attributive adjectives, nouns, and comparatives as features of informational/expository discourse. The present corpus, though limited to pedagogical materials, mirrors this developmental continuum, suggesting that the Advanced Masterclass series deliberately exposes learners to more informational grammatical profiles.

5.3. Pedagogical Implications

These findings hold direct implications for text selection and grammar focus in advanced EFL instruction. Teachers designing grammar, reading, or writing tasks can strategically use Proficiency texts to highlight narrative grammar—for instance, determiners, pronouns, and prepositions that promote cohesive storytelling. Advanced texts, by contrast, can be used to practice expository structures such as quantifiers, comparative adjectives, and participial modifiers.

The quantitative differences also underscore the pedagogical value of data-informed syllabus design. Rather than relying on intuition, instructors can use POS profiles to align classroom materials with target grammar areas. For example, if a course emphasizes argumentative writing, exposure to Advanced-level passages with higher nominal density may scaffold learners toward complex syntactic packaging. Similarly, Proficiency texts can support remedial work on reference and cohesion before tackling C2-level discourse.

Finally, this study demonstrates that POS profiling, when applied transparently, can serve as a diagnostic tool for evaluating textbook progression. Publishers and educators may adapt this method to verify whether their materials truly embody the intended CEFR level distinctions.

5.4. Limitations and Future Directions

Several limitations temper these conclusions. First, the dataset was restricted to a single publisher series, limiting generalizability across pedagogical traditions. Second, POS-level analysis, though precise, abstracts away from deeper syntactic structures such as clause subordination, phrasal complexity, or dependency patterns that might further differentiate proficiency levels. Third, the sample size (30 texts per group) affords limited power to detect medium-sized effects after multiple-comparison correction.

Future research should therefore extend this approach to multi-publisher corpora and incorporate syntactic and lexical complexity indices (e.g., T-unit ratios, dependency distance). Cross-validation with learner production corpora could also verify whether textbook grammatical exposure aligns with actual learner usage. Such work would advance both corpus-based pedagogy and textbook evaluation methodology.

5.5. Summary

In summary, the Discussion interprets the results as evidence of a subtle but systematic grammatical evolution across the C1–C2 interface: from informational precision toward cohesive narrative framing. These findings substantiate the pedagogical intuition that Proficiency materials consolidate grammatical cohesion, whereas Advanced materials expand lexical and descriptive sophistication. Importantly, this interpretation rests on a reproducible, code-based analytic framework that can be replicated for any future EFL corpus study.

6. Conclusion

This study quantitatively compared the distribution of part-of-speech categories across Advanced Masterclass and

Proficiency Masterclass EFL textbooks to examine whether grammatical composition changes systematically between CEFR levels C1 and C2. Despite modest overall differences (PERMANOVA $F = 2.624$, $p = .006$, $R^2 \approx .03$), the analyses revealed consistent functional contrasts: Proficiency texts displayed higher proportions of determiners and prepositions, whereas Advanced texts featured more numerals, adjectives, and adverbs. These shifts reflect a broader developmental trend from informational and expository elaboration toward cohesive narrative grammar at higher proficiency levels.

By linking quantitative corpus evidence with pedagogical interpretation, the study demonstrates how POS profiling can be used as an objective diagnostic framework for textbook evaluation and selection. The reproducible code-based pipeline—combining classical, non-parametric, robust, and multivariate tests—ensures full methodological transparency and provides a model for future corpus-driven studies of language-teaching materials. Although limited to one textbook series, the approach can be readily extended to multi-publisher corpora or learner-production data to trace grammar development across the advanced continuum.

Acknowledgements

The authors gratefully acknowledge the use of OpenAI's ChatGPT (GPT-5, 2025) and Google's Gemini 2.5 flash and pro for technical assistance in developing and documenting Python/Colab code for POS extraction, performing statistical analyses, and refining the academic writing style. Both systems were employed as language-modeling aids under the authors' supervision. All data interpretation, results validation, and final text approval were inspected and validated by the authors.

References

[1] Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511621024

[2] Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Pearson Education

[4] Oxford University Press. (2012). *Advanced Masterclass: Student's Book.* OUP.

[5] Oxford University Press. (2012). *Advanced Masterclass: Workbook.* OUP.

[6] Oxford University Press. (2015). *Proficiency Masterclass: Student's Book.* OUP.

[7] Oxford University Press. (2015). *Proficiency Masterclass: Workbook.* OUP.

[8] Römer, U. (2006). Pedagogical applications of corpora: Some reflections on the current scope and a wish list for future developments. *Zeitschrift für Angewandte Linguistik, 44*(2), 121-134.

[9] Römer, U., Cortes, V., & Friginal, E. (Eds.). (2020). *Advances in corpus-based research on academic writing: Effects of discipline, register, and writer expertise*. John Benjamins Publishing Company