

Comparing Argumentative Writing Features Across Two Timed Writing Tasks Among University EFL Learners

Reginald Gentry

Sophia University
7-1 Kioicho, Chiyoda-ku, Tokyo, 102-8554
r-gentry-8c7@sophia.ac.jp

Abstract

This study examined variation in argumentative writing features across two timed writing tasks completed by university English as a Foreign Language (EFL) learners in an academic writing course at a Japanese university. The final corpus comprised 154 argumentative essays written by 77 learners across four intermediate- and advanced-level course sections. Students completed two independent timed essays approximately six weeks apart under comparable conditions, with each requiring an introduction, body paragraphs, a counterargument, and a conclusion. Analysis focused on lexical density, syntactic complexity, cohesive device use, and argument structure markers, comparing variation across tasks and level groupings. Because students selected their own essay topics, topic variation was treated as a characteristic of authentic classroom writing rather than as an experimentally controlled variable. Lexical density increased significantly from the first to the second task across all four sections, whereas syntactic, cohesive, and argumentative features showed more limited or section-specific variation. These findings suggest that changes across repeated timed writing performances may differ by linguistic and rhetorical feature. The study contributes to learner corpus research by providing within-learner evidence from intermediate- and advanced-level EFL writers in an authentic university instructional context.

Keywords: learner corpus, argumentative writing, EFL writing

1. Introduction and Literature Review

Learner corpus research has extensively examined how textual features vary across level groupings, genres, and instructional contexts, but comparatively few studies have investigated how the same learners' argumentative writing varies across two independent timed writing performances completed within a single instructional context. This gap is notable given that timed, in-class writing is a common assessment format in EFL academic writing courses, and understanding how learners' textual features shift between such performances has direct implications for how instructors interpret single-sample writing assessments.

Lexical and syntactic complexity have been widely used as indicators of second language (L2) writing development and proficiency (Ai and Lu, 2013). Lexical density in particular has received less attention than lexical diversity in this literature, despite its relevance to how densely content is packed into learner prose. Syntactic complexity research has similarly documented developmental variation in L2 argumentative writing, including within-genre effects tied to topic and task conditions (Atak and Saricaoglu, 2021).

Cohesion has also been established as a meaningful dimension of L2 writing quality. Cohesive devices, following the foundational framework of Halliday and Hasan (2014), operate at both local and global levels of text organization, and research has shown that L2 writers' use of cohesive devices develops over the course of instruction and relates to human judgments of essay quality (Crossley, Kyle, and McNamara, 2016). However, this line of research has focused primarily on development across a semester or longer, rather than on variation between two independent timed writing performances completed weeks apart.

Argument structure, including the presence of claims, counterarguments, concessions, and rebuttals, is a defining feature of argumentative writing and a frequent focus of

instruction in academic writing courses. Research applying Toulmin's model of argumentation to L2 assessed writing has found that both the presence and quality of structural elements such as counterarguments and rebuttals are associated with differences in writing proficiency, with these elements appearing more consistently in higher-scoring essays (Chuang and Yan, 2022). Despite its instructional centrality, however, argument structure marker frequency has been comparatively underexamined as a textual feature in learner corpus studies of timed writing variation, and in particular has rarely been tracked across two performances by the same learners.

A corpus-based approach was adopted for this study because it allows systematic, replicable measurement of textual features across a large number of authentic writing samples, rather than relying on a small number of texts examined through close reading or holistic judgment alone. Corpus methods make it possible to quantify features such as lexical density, syntactic complexity, cohesive device use, and argument structure markers consistently across every essay in the dataset, allowing patterns to be compared systematically across both writing tasks and level groupings (Sanni and Wirza, 2025). Because the essays were produced under normal classroom testing conditions, this approach also preserves the ecological validity of the writing samples, capturing what learners actually produce under authentic timed-writing conditions rather than performance elicited in an artificial or laboratory setting, a methodological priority increasingly emphasized in recent learner corpus design work (Gablasova et al., 2024).

The present study addresses this gap by examining variation in argumentative writing features, specifically lexical density, syntactic complexity, cohesive device use, and argument structure markers, across two independent timed writing tasks completed by the same cohort of university EFL learners approximately six weeks apart, and across four sections representing intermediate- and advanced-level groupings. Because students generated their own essay topics for each task, this design differs from

studies employing identical task repetition and instead captures variation under authentic, topic-varying instructional conditions. This approach aligns with the broader task-based literature on procedural task repetition, in which learners complete structurally similar but not identical tasks. A recent meta-analysis of task repetition in L2 written performance found that repeating tasks procedurally, rather than identically, is associated with gains across complexity, accuracy, and fluency measures, though effects vary by the specific dimension of performance examined and the interval between repetitions (Abdi Tabari, Zhuang, and Farahanynia, 2025). This body of work provides a relevant precedent for the present design, in which the two writing tasks shared structure and genre but not topic.

This study asks two research questions: (1) How do lexical density, syntactic complexity, cohesive device use, and argument structure markers vary between two timed argumentative writing tasks completed by the same learners? (2) How do these features vary across four sections representing different level groupings on the second writing task?

2. Methods

Participants and Setting

Participants were initially drawn from 81 first-year university students enrolled in a required first-year academic writing course at a university located in central Japan, across four sections representing intermediate- and advanced-level classes (roster sizes 22, 20, 16, and 23). Inclusion in the analytic sample required submission of both Essay 1 and Essay 2; students who submitted only one essay or neither were excluded, as were repeater students (i.e., second-year students retaking the course), since their additional prior exposure to English-language coursework afforded more time to develop writing proficiency relative to first-year classmates. Four students were excluded on these grounds: one student each in Section A and Section B who did not submit both essays, one Section D student who submitted Essay 2 without a corresponding Essay 1 submission, and one additional Section D student who was excluded as a second-year repeater. This yielded a final matched analytic sample of $N = 77$ students (Section A $n = 21$, Section B $n = 19$, Section C $n = 16$, Section D $n = 21$). Students' declared majors spanned foreign language, science, economics, philosophy, and foreign literature.

Writing Tasks

Classes met twice weekly across a 15-week semester, with each class session lasting 90 minutes. Students completed two independent timed argumentative writing tasks approximately six weeks apart under comparable in-class testing conditions, with 80 of the 90-minute session allotted to the writing task itself. Each task required a multi-paragraph argumentative essay of 150-400 words, comprising an introduction, body paragraphs with supporting reasons, a counterargument, and a conclusion. Students composed their essays on their own personal electronic devices (laptop or tablet computer) rather than on paper or an instructor-supplied device.

Each writing task used an instructor-created Word document template. Students downloaded the template from the course's Blackboard site at the start of the writing

period and uploaded their completed file to the same site when time ended. The template required students to enter identifying information (name and student ID) and their chosen topic before beginning the essay itself. The instructor reviewed each submission for completeness immediately after class; students whose submissions were missing required information were asked to resubmit. During the class session preceding each writing task, students were given a sample essay and the scoring rubric used to evaluate their essays, allowing them to understand the task requirements and ask clarifying questions in advance of the timed writing session itself.

Students generated their own essay topics from a shared set of prompts rather than writing on an instructor-assigned topic. This design choice reflects research linking topic choice to increased learner autonomy and engagement in L2 writing: when learners are given control over topic selection, their intrinsic motivation and investment in the writing task increase (Alzubi and Nazim, 2024), consistent with self-determination theory's account of autonomy as a basic driver of intrinsic motivation and engagement (Ryan and Deci, 2000). Topic variation across the corpus is therefore treated as a feature of the dataset consistent with authentic academic writing conditions rather than as a controlled variable to be minimized.

Corpus Size

The final corpus comprises 154 essays (77 students * 2 essays each): Section A $n = 21$, Section B $n = 19$, Section C $n = 16$, Section D $n = 21$, per essay task. Essay counts were verified through direct review of submission documents rather than gradebook records alone, since several essays lacked standard identifying headers and required manual confirmation of submission boundaries. This verification step was necessary because the section-level sample sizes reported here directly determine the degrees of freedom and statistical power of the comparisons reported in Results; given the modest size of each section, an undercount or overcount of even one or two essays could shift a result across the significance threshold, so counts were confirmed against the essays themselves rather than assumed from administrative records.

Analysis

Essays were analyzed for four categories of textual feature: lexical density, syntactic complexity, cohesive device use, and argument structure markers, selected because together they span the lexical, syntactic, and rhetorical dimensions of argumentative writing most consistently associated with L2 writing development and quality in the literature reviewed above (see Limitations for a description of the analysis method and its constraints). Within-class comparisons (Essay 1 vs. Essay 2, same section) were evaluated using Welch's independent-samples *t*-test rather than the standard Student's *t*-test because Welch's test does not assume equal variances between the two samples being compared, which is a safer assumption given the modest and unequal section sizes in this corpus. Across-class comparisons on Essay 2 were evaluated using one-way analysis of variance (ANOVA) rather than a series of pairwise *t*-tests because ANOVA allows all four sections to be compared simultaneously on a single outcome without

inflating the overall risk of a false positive that repeated pairwise testing would introduce.

3. Results

Within-Class Variation (Essay 1 to Essay 2)

Lexical density increased significantly from Essay 1 to Essay 2 in all four sections (Section A: 53.17% to 60.45%, $t(37.06) = -4.90, p < .001$; Section B: 55.91% to 60.60%, $t(35.7) = -3.39, p = .002$; Section C: 55.53% to 60.08%, $t(29.97) = -2.78, p = .010$; Section D: 54.92% to 61.51%, $t(38.85) = -4.14, p < .001$), indicating that students produced denser, more content-word-heavy prose on the second timed writing task regardless of level grouping. Mean sentence length also increased in all four sections, reaching significance in Section D (17.33 to 20.46 words, $t(38.99) = -3.03, p = .005$) and approaching significance in Section A and Section C. Subordination frequency showed no consistent direction of change and no section reached significance, suggesting that increased lexical density was not accompanied by a parallel increase in clausal embedding. Cohesive device frequency and argument-structure marker frequency both declined slightly across most sections, though none of these differences reached significance. Table 1 presents the lexical density and mean sentence length figures for both essays by section, together with the results of the Welch's t-tests comparing them.

Sec.	N	Metric	Ess.1	Ess.2	p
A	21	Lex. density %	53.17	60.45	<.001
A	21	Sent. length	16.18	18.41	.075
B	19	Lex. density %	55.91	60.60	.002
B	19	Sent. length	19.57	20.17	.600
C	16	Lex. density %	55.53	60.08	.010
C	16	Sent. length	15.34	17.06	.060
D	21	Lex. density %	54.92	61.51	<.001
D	21	Sent. length	17.33	20.46	.005

Table 1: Lexical density and mean sentence length by section, Essay 1 vs. Essay 2 (Welch's t-test).

Across-Class Variation (Essay 2)

A one-way ANOVA comparing the four sections on Essay 2 found a significant difference in mean sentence length ($F(3, 73) = 3.53, p = .019$), with Sections B and D producing longer sentences on average than Sections A and C. No other measure showed a significant across-section difference on Essay 2 (lexical density: $p = .846$; subordination: $p = .615$; cohesive devices: $p = .737$; argument markers: $p = .839$), suggesting that once students had completed one prior timed writing task in the course, the four sections converged on broadly similar profiles for most of the measures examined, with sentence length as the one dimension distinguishing them.

Tables 2 and 3 break the cohesive device and argument structure marker findings down by category rather than as the single aggregated figure reported above, to show which specific types of device or marker are driving the section-level patterns. Table 2 presents cohesive device frequency by category, and Table 3 presents argument structure marker frequency by category, both for Essay 2.

Sec.	Add.	Adv.	Caus.	Temp.
A	3.28	0.81	0.94	0.82
B	3.71	0.93	0.88	0.59
C	2.60	0.94	0.90	1.06
D	3.51	0.73	0.83	0.67

Table 2: Cohesive device use by category, per 100 words, Essay 2 (Additive, Adversative, Causal, Temporal/Sequential).

Sec.	Claim	Ctr.	Conc.	Reb.	Concl.
A	0.90	0.19	0.00	1.62	0.86
B	1.16	0.26	0.00	1.84	0.79
C	0.81	0.06	0.12	1.94	0.88
D	0.95	0.24	0.10	1.29	1.05

Table 3: Argument structure marker use by category, average count per essay, Essay 2 (Claim, Counterargument-intro, Concession, Rebuttal, Conclusion).

Across both tables, rebuttal markers and additive cohesive devices are consistently the most frequent category in every section, while concession markers are rare or absent throughout. Taken together, these results indicate that lexical density is the most robust and consistent marker of change between the two timed writing tasks, while syntactic, cohesive, and argumentative features show more limited, category-specific, or section-specific patterns of variation.

4. Discussion

The most robust finding of this study is the significant, consistent increase in lexical density from the first to the second timed writing task across all four sections. This pattern held regardless of section or level grouping, suggesting that increased lexical density between timed writing performances may be a general feature of repeated timed writing under these instructional conditions rather than a level-specific effect. This finding extends prior lexical complexity research, which has more often examined lexical diversity than lexical density (Ai and Lu, 2013), by showing that density itself shifts measurably even across a relatively short interval of approximately six weeks and without identical task repetition.

The parallel, though less uniformly significant, increase in mean sentence length suggests that learners were producing not only denser but somewhat longer sentences on the second task. Notably, this increase in surface complexity was not accompanied by a corresponding increase in subordination frequency, which showed no consistent direction of change across sections. This dissociation between lexical density and subordinate clause frequency is consistent with prior findings that lexical and syntactic complexity do not always develop in lockstep, and that gains in one dimension of complexity do not guarantee parallel gains in another.

The slight decline in cohesive device frequency and argument-structure marker frequency observed in most sections, though not statistically significant, is a pattern worth noting rather than dismissing. One possible interpretation is that as students packed more content words into their sentences, they relied comparatively less on explicit connective and argumentative signposting to organize that content, though this interpretation is speculative given the absence of significance and would benefit from qualitative follow-up examining whether cohesion was instead achieved through means not captured by the marker list used here, such as lexical or referential cohesion.

A brief excerpt from one student's matched pair of essays illustrates the kind of shift the aggregate figures represent. In Essay 1, the student's counterargument paragraph reads: "Opponents may claim that prioritizing academic background can create a hierarchical structure among universities and they also ask a question why ability-based evaluation is insufficient." In Essay 2, the corresponding counterargument paragraph incorporates a specific statistic to support the claim: "According to NewsGuard that researches online misinformation, the number of AI-generated fake news websites increased from 2,089 in October 2025 to 3,006 in March 2026, a rise of about 44% in less than five months, which shows how rapidly this technology is spreading online." This student's overall lexical density rose from 56.7% in Essay 1 to 60.9% in Essay 2, consistent with the section-wide pattern; the excerpt suggests that at least part of this increase may reflect a greater willingness to incorporate specific, source-attributed detail into argumentation on the second task, though a single case cannot establish this as a general mechanism and it is offered here only as an illustration, not as evidence, of the pattern reported above.

5. Limitations

Textual features were extracted using word-list and pattern-based methods rather than a part-of-speech tagger or syntactic parser, since the analysis environment did not permit installation of natural language processing tools such as spaCy or a dependency parser. Lexical density was calculated via stop word exclusion rather than content-word tagging, and syntactic complexity was approximated using subordinators and coordinators frequency and mean sentence length rather than T-unit-based measures such as clauses per T-unit. Cohesive device and argument-structure marker counts relied on a curated, finite marker list and may undercount devices expressed through phrasing not captured by the list. Future analyses of this corpus should reprocess the essays with a validated part-of-speech tagger

(e.g., spaCy or the Stanford tagger) to recompute lexical density from true content-word counts, and with an established L2 syntactic complexity analyzer that performs automatic T-unit segmentation, which would allow clause-level subordination indices rather than the surface marker counts used here. Reprocessing cohesion and argument structure with validated instruments such as Coh-Metrix, the Tool for the Automatic Analysis of Cohesion (TAACO), or the Tool for the Automatic Analysis of Lexical Sophistication (TAALES) would similarly allow the marker-list approach used here to be checked against, and potentially replaced by, indices with established validity evidence. Until that reprocessing is done, the reported differences should be interpreted as indicative of broad patterns rather than fine-grained syntactic distinctions.

The six-week interval between Essay 1 and Essay 2 may also be too short to capture the full extent of change in some of the features examined. Lexical density shifted significantly within this window, but subordination, cohesion, and argument marker frequency did not, and it is possible that these measures require a longer developmental interval, or more than two timepoints, to show detectable change. A design incorporating three or more timed writing samples spaced across a full semester or academic year would allow this question to be addressed directly, distinguishing a genuinely flat trajectory from one that simply had not yet become visible within six weeks.

Although the corpus included students from five declared majors (foreign language, science, economics, philosophy, and foreign literature), the analysis did not examine variation by major because per-major subsamples within each section were too small to support reliable comparison. Disciplinary background is a plausible predictor of argumentative writing style independent of course level; for example, science majors might be expected to favor a more data-driven, analytical style of argumentation, while majors in fields such as philosophy or foreign literature might favor more discursive or interpretive elaboration. Testing this possibility would require a sampling design that deliberately recruits sufficient numbers of students from each major, which the present study, drawn from intact course sections rather than a major-stratified sample, was not positioned to do.

Finally, the study sample was limited to intermediate- and advanced-level sections of the course, and the reported patterns, particularly the consistent lexical density increase, should not be assumed to generalize to beginner-level EFL writers. Learners with more limited linguistic resources may require more scaffolded support to produce a second timed writing performance denser than their first, or may show a different pattern of change entirely. Extending this design to beginner-level sections, with appropriate scaffolding built into the writing task, would clarify whether the pattern observed here holds only above a certain proficiency threshold.

This study also relied exclusively on quantitative textual analysis and did not incorporate any qualitative component, such as student interviews, think-aloud protocols, or retrospective reflections, that could help explain why lexical density increased so consistently while other features did not. Qualitative data of this kind could clarify whether the pattern reflects a deliberate strategy on

students' part, increased comfort with the timed-writing format on the second occasion, exposure to feedback on the first essay, or some other mechanism entirely. Future iterations of this design should incorporate a qualitative strand, for example brief post-task interviews or stimulated recall sessions with a subsample of students, to complement the corpus-based findings reported here.

Finally, no background information was collected on individual students' language learning histories, such as age of first exposure to English, years of prior formal instruction, or study-abroad experience, all of which could plausibly influence both baseline writing performance and the degree of change observed between the two tasks. Without this information, it is not possible to determine whether the consistency of the lexical density increase across sections reflects a genuinely uniform instructional effect or instead masks underlying variability tied to individual learners' language backgrounds. Future data collection alongside this kind of writing task should include a brief language background questionnaire to allow this possibility to be examined directly.

6. Conclusion

This study examined variation in lexical density, syntactic complexity, cohesive device use, and argument structure markers across two independent timed argumentative writing tasks completed by 77 matched university EFL learners across four sections of an academic writing course. The clearest finding was a significant increase in lexical density from the first to the second writing task across all four sections, accompanied by a less consistent but often significant increase in mean sentence length, while subordination frequency, cohesive device use, and argument marker frequency showed no clear directional pattern. Across sections, the four groups were largely similar on the second task, differing significantly only in mean sentence length.

These findings suggest that variation between two timed writing performances, even those completed under topic-varying, authentic instructional conditions rather than identical task repetition, is measurable and, for lexical density in particular, remarkably consistent across level groupings. This has practical implications for instructors and program administrators who rely on single-sample timed writing for placement or assessment purposes, as the present findings suggest that a learner's textual profile on an early timed writing task may not fully represent their profile on a later one, even within a short instructional interval. The directions for future research raised throughout this paper, extending the timeline beyond two timepoints, analyzing variation by declared major, testing generalizability to beginner-level learners, reprocessing the corpus with validated parser-based tools, incorporating a qualitative strand to explain the mechanisms behind the observed patterns, and collecting language background data on individual learners, together point toward a fuller account of how argumentative writing develops across a course than a single pair of timed samples can provide on its own.

Bibliographical References

- Abdi Tabari, M., Zhuang, J. and Farahanynia, M. (2025). Task repetition and L2 written performance: A meta-analysis. *Journal of Second Language Writing*, 70.
- Ai, H. and Lu, X. (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. In A. Díaz-Negrillo, N. Ballier, and P. Thompson (Eds.), *Automatic treatment and analysis of learner corpus data*. Amsterdam: John Benjamins, pp. 249–264.
- Alzubi, A. A. F. and Nazim, M. (2024). Students' intrinsic motivation in EFL academic writing: Topic-based interest in focus. *Heliyon*, 10(1):e24169.
- Atak, N. and Saricaoglu, A. (2021). Syntactic complexity in L2 learners' argumentative writing: Developmental stages and the within-genre topic effect. *Assessing Writing*, 47.
- Chuang, P. L. and Yan, X. (2022). An investigation of the relationship between argument structure and essay quality in assessed writing. *Journal of Second Language Writing*, 56:100892.
- Crossley, S., Kyle, K. and McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, 32:1–16.
- Gablasova, D., Harding, L., Bottini, R., Brezina, V., Ren, H., Iamartino, G., Li, Y., Liu, T., Poggesi, L., Savski, K., Toomaneejinda, A. and Zottola, A. (2024). Building a corpus of student academic writing in EMI contexts: Challenges in corpus design and data collection across international higher education settings. *Research Methods in Applied Linguistics*, 3(3):100140.
- Halliday, M. A. K. and Hasan, R. (2014). *Cohesion in English*. London: Routledge.
- Ryan, R. M. and Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68–78.
- Sanni, F. A. and Wirza, Y. (2025). An analysis of specialized vocabulary frequency, lexical density, and multimodality of students' work in an EMI class: A corpus-based study. *International Journal of Education*, 18(1):23–34.