# A Multi-Dialectal, Longitudinal Corpus of Human-AI Hybrid Language Production

**Qiao Gan**[1], **Jonathan Dunn**[2], **Andrea Nini**[3], **Benjamin Adams**[1,4]
[1]New Zealand Institute of Language, Brain and Behaviour, University of Canterbury, New Zealand
[2]Department of Linguistics, University of Illinois Urbana-Champaign, United States
[3]Department of Linguistics and English Language, University of Manchester, United Kingdom
[4]Department of Computer Science and Software Engineering, University of Canterbury, New Zealand
qiao.gan@canterbury.ac.nz, jedunn@illinois.edu
andrea.nini@manchester.ac.uk, benjamin.adams@canterbury.ac.nz

**Abstract**

This paper presents a multi-dialectal, longitudinal corpus of human-AI hybrid language production, comprising purely human-written texts, purely LLM-generated texts, and hybrid texts produced under different LLM-assistance modes (e.g., stylistic suggestions, short continuations, partial essay generation). The corpus includes 693 participants from five national English dialects, with natural and hybrid samples paired within individuals over a four-week period. This design enables investigation of both short- and longer-term effects of LLM assistance on language use across geographic and social contexts. To illustrate the corpus's utility, we analyze linguistic features across three dimensions: lexical diversity, syntactic complexity, and stylistic variation. The results show that LLM assistance enhances lexical diversity without a corresponding increase in syntactic complexity, revealing distinct effects across linguistic dimensions. Overall, this corpus offers a valuable resource for studying human-AI interaction, dialectal variation, and the influence of AI assistance on written language.

**Keywords:** synthetic corpora, human computer interaction, linguistic variation

## 1. Introduction

Large language models (LLMs) such as ChatGPT have rapidly reshaped writing practices, offering fluent suggestions and initial drafts that foster a hybrid human-computer production environment (Lee et al., 2022; Lehmann and Buschek, 2025; Prakash et al., 2025; Swanson et al., 2021). This hybrid production environment raises the question of how such models impact linguistic diversity, especially for less dominant dialects and less prototypical registers. Some studies suggest that LLMs promote a uniform "default" style (Guo et al., 2024; Paschalidis, 2025), favoring certain lexical choices (e.g., *delve*, *pivotal*) and stylistic patterns such as nominalization (e.g., *decide* vs. *decision*). A model with such preferences would serve to reduce expressive variation across individuals (Agarwal et al., 2025; Dentella et al., 2025). Other studies, however, report benefits for linguistically diverse users, indicating that LLM support can improve grammar, fluency, and accessibility for non-native speakers and children with disabilities (Utepbayeva et al., 2024; Shin and Lee, 2025; Zhang et al., 2025).

These divergent findings sharpen this paper's central question: does LLM-assisted writing homogenize language or can it coexist with lexical and structural diversity? To help answer that question, we present the first multi-dialectal, longitudinal corpus of hybrid human-computer production. This corpus contains purely natural human-produced samples, purely synthetic LLM-produced samples, and a range of hybrid samples produced under different conditions. Importantly, speakers of five national dialects of English are included, with natural and hybrid samples paired for the same individuals.

Existing essay collections such as the ASAP dataset (Chen and He, 2013) and the PERSUADE corpus (1.0 and 2.0; Crossley et al. 2022, 2024) focus exclusively on human-written American English norms. Some corpora have begun to explore hybrid human-LLM collaboration (e.g., Padmakumar and He 2024; Swanson et al. 2021), such as the CoAuthor corpus, which records GPT-3–assisted writing sessions from 63 writers (Lee et al., 2022). However, CoAuthor's participants are not drawn from diverse English-speaking populations and the dataset covers only single-task interactions. To our knowledge, no existing public dataset combines essays from multiple English dialects under varied AI-assistance conditions in a longitudinal design.

In this work, we introduce a novel dataset designed to fill this gap. It comprises essays from 693 participants representing five English varieties (American, Canadian, British, Australian, New Zealand English) and spanning a range of social backgrounds. Each participant produced three essays in a within-subject design: a baseline essay written without LLMs, an LLM-assisted essay under one assigned condition, and a follow-up essay three weeks later (again without LLMs). The hy-



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 11286–11299
11–16 May 2026. 

brid human-computer conditions (between-subject design) offer distinct forms of assistance: (1) stylistic suggestions, (2) incremental continuations (3-5 words), or (3) partial essay generation (first half of the essay). We also collect detailed meta-data and a survey on participants' AI-usage practices. The key contributions of this dataset are:

- **Multiple English Dialects:** Written production from 693 participants across five English varieties, enabling comparative studies across geographic and social groupings.
- **Longitudinal Design:** A within-subject setup where each participant produces three essays on different topics: a baseline essay, an LLM-assisted essay, and a post-test essay written three weeks later.
- **Multiple Hybrid Conditions:** Participants experience one of four conditions: stylistic suggestions, short continuations, half-essay generation, or natural writing. This between-subject design isolates the influence of LLM assistance on writing outcomes.
- **Rich Meta-Data:** Each essay is accompanied by demographic information like age, gender, occupation, education, ethnicity, city of residence, and attitudes towards AI.

Taken together, this corpus enables investigation of our central question: Does LLM-assisted writing reduce the linguistic diversity found in human-generated text? By analyzing linguistic features across both hybrid conditions and dialects, we can assess whether LLM-assistance promotes convergence toward a uniform style or preserves diverse forms of expression. In this paper, we describe the data collection process and the resulting corpus. We then illustrate its research potential through analyses of lexical diversity, syntactic complexity, and stylistic variation. We aim for this dataset to serve as a foundation for research on the interplay between LLM assistance and linguistic diversity.

## 2. Data Collection

**Design.** After refining the study design through two pilot studies, one on topic selection and essay length (Appendix A), and the other on the human-LLM interaction interface, we conducted the formal data collection using GPT-4.1 for the LLM-assisted conditions (specific parameter settings are given in Appendix B). As illustrated in Figure 1, this study employs a mixed-methods design, with a within-subject longitudinal factor (pretest vs. AI-assisted vs. posttest) and a between-subject factor (four writing conditions; each participant experiences only one). This allows us to investigate the effects of different types of LLM writing assistance over a four-week period, capturing both immediate and longer-term impacts and providing a baseline of natural variation within individuals.

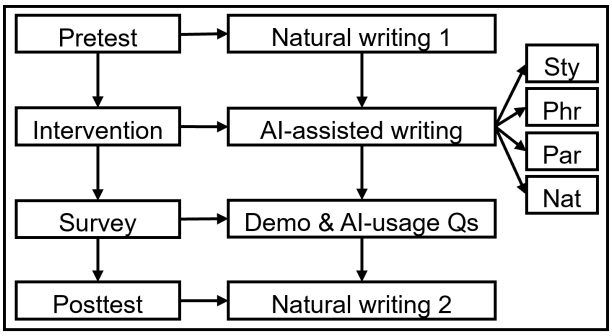


Figure 1: The experimental design (Sty = Stylistic, Phr = Phrasal, Par = Paragraph, Nat = Natural).

To enhance representational diversity (e.g., Dunn et al. 2024), we recruited 720 native speakers of five English varieties via Prolific, including American, British, Canadian, Australian, and New Zealand English. Participants were randomly assigned to one of four conditions, resulting in roughly equal group sizes. The number of samples for each condition varies due to differences in adherence to study instructions (see Table 1).

| Condition | US | CAN | UK | AUS | NZ | Total |
|---|---|---|---|---|---|---|
| Natural | 31 | 30 | 30 | 30 | 31 | 152 |
| Stylistic | 36 | 36 | 38 | 36 | 35 | 181 |
| Phrasal | 36 | 37 | 37 | 36 | 35 | 181 |
| Paragraph | 37 | 35 | 36 | 36 | 35 | 179 |

Table 1: Participant distribution by Condition and Dialect.

**Conditions.** Participants were scheduled to complete three writing sessions, each lasting approximately 30 minutes and producing a 500-word essay (Figure 1). Sessions 1 and 3 involved natural writing on two different topics. Session 1 served as an individual baseline. Session 3 was administered as a posttest; however, its primary purpose was not to examine behavioural change over time, but to provide a statistical control for between-session variability. The posttest retention rate after three weeks was 86% (598/693). Five topics, selected from the first pilot study, were evenly distributed across sessions to control for topic effects.

Session 2 contained the LLM-assisted writing conditions. Here a *Natural Writing* group serves as a control, capturing variation within individuals that is influenced only by repetition of the production task. Three other conditions with LLM-assistance are contrasted with this control group: First, the *Stylistic Suggestion* condition provides grammar and style suggestions. Second, the *Phrasal Con-*

| Factor | Participants ($N$ = 693) |
|---|---|
| Age | 19-35 (49%), 36-55 (39%), 56+ (11%) |
| Gender | Women (51%), Men (48%), other (1%) |
| Ethnicity | African (10%), Asian (16%), White (66%), other (8%) |
| Education | High School (30%), Bachelor (48%), Master (18%), other (4%) |

Table 2: Participant demographics.

*tinuation* condition provides 3-5 word continuations based on the current sentence. In both of these conditions, participants are required to query the LLM at least five times per essay. Each query generates three alternative suggestions which participants can accept, revise, or reject. Third, the *Paragraph Continuation* condition generates a synthetic paragraph (around 250 words) to initiate the writing process. Three alternative paragraphs are provided which participants can accept, revise, or reject. While each participant completes the same tasks for Sessions 1 and 3, the four conditions in Session 2 allow us to observe the impact of LLM-assistance on participants' written production.

We used an adapted *CoAuthor* interface (Lee et al., 2022), evaluated in the second pilot study and integrated with Qualtrics, to collect both natural and LLM-assisted writing, including keystrokes. This setup allows us to track participant responses to LLM suggestions as well as to monitor general writing behaviors such as revisions and pausing. We also collected demographic information: age, gender, ethnicity, occupation, education, current and previous residence, attitudes toward AI, and frequency and purpose of AI usage (Appendix C). As shown in Table 2, participant information on age, gender, ethnicity, and education supports detailed analyses along those dimensions. Raw occupation and residence data are shared so other researchers can apply their own classification schemes.

We also used the same five topics to prompt GPT-4.1 to produce 108 essays for each of the five varieties. Instructions specified, for example, "as if you are a native speaker of American English who uses American English spelling, vocabulary, syntax, and style". These essays were then evenly assigned across Sessions 1-3 to create a synthetic group comparable to the other participant groups.

## 3. Corpus Description

**Corpus Size.** Table 3 presents the total and mean word counts for essays across conditions and dialects. The number of participants who completed each essay, as well as those retained for the post-test, are also reported. Session 1 is a natural writing condition for all participants; Session 2 involves random assignment to different LLM-assisted conditions; and Session 3 is a follow-up natural writing session. Mean essay lengths are broadly comparable across conditions, and differences across dialects are small, indicating similar levels of engagement and task compliance among participants.

**Human-LLM Contribution Patterns.** Table 4 shows a word-level breakdown of LLM usage, including the mean number of queries, the acceptance rate of AI suggestions, and the proportion of human-written text across conditions and dialects.

Overall, participants in the Stylistic condition relied less on GPT-4.1, with moderate query counts and high acceptance rates, whereas those in the Phrasal condition submitted many more queries, but acceptance rates were lower. The Paragraph condition shows the lowest query counts. Thus, the proportion of text written by participants varied across conditions and varieties, suggesting that different types of LLM-assistance influence how much content participants generate themselves.

**Human-LLM Collaboration.** Following Storch (2002) and Lee et al. (2022), we use *mutuality* to evaluate human-LLM interaction in Session 2. Mutuality quantifies the degree of collaboration between the human and GPT-4.1 (1.0 = fully collaborative; 0.0 = fully independent), based on an action $e$, such as multiple queries, selecting, reopening, and navigating suggestions. We define:

$$\text{Mutuality} = \frac{\sum_i [e_i \in \mathcal{I}]}{\sum_i [e_i \in \mathcal{I}] + \sum_i [e_i \in \mathcal{A}]}$$

where $\mathcal{I} = \{\text{insert, choose, reopen, navigate}\}$ denotes the set of human-LLM interaction event blocks, and $\mathcal{A} = \{\text{dismiss, insert, delete}\}$ denotes the set of writing-alone event blocks.

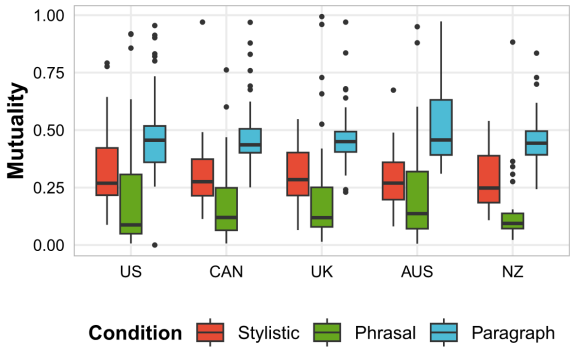


Figure 2: Mutuality scores by Condition and Dialect.

As shown in Figure 2, mutuality varies substantially across writing conditions, but less so across dialects. Specifically, the Paragraph condition consistently facilitated human-LLM co-creation across all varieties, with median mutuality scores between 0.44 and 0.46. This high degree of balance suggests that this condition is most effective at pro-

| Condition | Dialect | Session 1 | | Session 2 | | Session 3 | |
|---|---|---|---|---|---|---|---|
| | | *Samples* | *Essay Length* | *Samples* | *Essay Length* | *Samples* | *Essay Length* |
| Natural | US | 31 | 17,188 (554 avg) | 31 | 17,224 (556 avg) | 23 | 12,700 (552 avg) |
| | CA | 30 | 16,263 (542 avg) | 30 | 16,885 (563 avg) | 26 | 14,765 (568 avg) |
| | UK | 30 | 16,086 (536 avg) | 30 | 15,928 (531 avg) | 28 | 15,487 (553 avg) |
| | AU | 30 | 16,532 (551 avg) | 30 | 16,470 (549 avg) | 28 | 15,208 (543 avg) |
| | NZ | 31 | 17,288 (558 avg) | 31 | 16,985 (548 avg) | 29 | 15,419 (532 avg) |
| Stylistic | US | 36 | 18,946 (526 avg) | 36 | 18,760 (521 avg) | 30 | 16,473 (549 avg) |
| | CA | 36 | 19,334 (537 avg) | 36 | 19,384 (538 avg) | 29 | 16,056 (554 avg) |
| | UK | 38 | 20,807 (548 avg) | 38 | 19,971 (526 avg) | 34 | 18,630 (548 avg) |
| | AU | 36 | 18,876 (524 avg) | 36 | 19,474 (541 avg) | 31 | 16,646 (537 avg) |
| | NZ | 35 | 18,502 (529 avg) | 35 | 18,917 (540 avg) | 25 | 13,660 (546 avg) |
| Phrasal | US | 36 | 19,502 (542 avg) | 36 | 19,756 (549 avg) | 32 | 17,213 (538 avg) |
| | CA | 37 | 20,040 (542 avg) | 37 | 21,227 (574 avg) | 32 | 18,397 (575 avg) |
| | UK | 37 | 20,329 (549 avg) | 37 | 20,296 (549 avg) | 35 | 19,205 (549 avg) |
| | AU | 36 | 18,943 (526 avg) | 36 | 19,370 (538 avg) | 27 | 14,575 (540 avg) |
| | NZ | 35 | 19,575 (559 avg) | 35 | 19,165 (548 avg) | 29 | 15,806 (545 avg) |
| Paragraph | US | 37 | 19,621 (530 avg) | 37 | 20,474 (553 avg) | 31 | 16,860 (544 avg) |
| | CA | 35 | 19,283 (551 avg) | 35 | 19,901 (569 avg) | 34 | 20,352 (599 avg) |
| | UK | 36 | 20,013 (556 avg) | 36 | 20,442 (568 avg) | 35 | 19,045 (544 avg) |
| | AU | 36 | 19,323 (537 avg) | 36 | 20,943 (582 avg) | 32 | 17,200 (538 avg) |
| | NZ | 35 | 18,682 (534 avg) | 35 | 19,327 (552 avg) | 28 | 14,655 (523 avg) |
| GPT 4.1 | US | 36 | 18,750 (521 avg) | 36 | 18,633 (518 avg) | 36 | 18,804 (522 avg) |
| | CA | 36 | 18,875 (524 avg) | 36 | 18,972 (527 avg) | 36 | 18,953 (526 avg) |
| | UK | 36 | 19,041 (529 avg) | 36 | 19,012 (528 avg) | 36 | 18,966 (527 avg) |
| | AU | 36 | 19,221 (534 avg) | 36 | 19,406 (539 avg) | 36 | 19,361 (538 avg) |
| | NZ | 36 | 19,322 (537 avg) | 36 | 19,607 (545 avg) | 36 | 19,367 (538 avg) |

Table 3: Total word counts (with averages) by Session, Condition, and Dialect, including number of samples (86% of samples retained in Session 3).

| Condition | Variety | Query | Acceptance | Human |
|---|---|---|---|---|
| Stylistic | US | 7.9 | 92% | 48% |
| | CAN | 9.4 | 93% | 49% |
| | UK | 8.8 | 91% | 46% |
| | AUS | 9.6 | 92% | 44% |
| | NZ | 8.4 | 92% | 45% |
| Phrasal | US | 29.1 | 83% | 56% |
| | CAN | 27.9 | 82% | 50% |
| | UK | 33.3 | 87% | 38% |
| | AUS | 34.8 | 77% | 44% |
| | NZ | 29.1 | 72% | 50% |
| Paragraph | US | 2.7 | 63% | 46% |
| | CAN | 3.1 | 75% | 47% |
| | UK | 3.7 | 76% | 50% |
| | AUS | 3.9 | 61% | 45% |
| | NZ | 2.0 | 79% | 50% |

Table 4: Mean number of LLM queries, acceptance rates of LLM suggestions, and proportion of human-written text by Condition and Dialect.

moting collaboration while preventing LLM dominance. In contrast, the Phrasal condition produced the most imbalanced outputs, with median mutuality scores between 0.09 and 0.14, indicating a strong skew in contributions. The Stylistic condition resulted in moderately balanced work, with median mutuality scores around 0.25-0.28. While this pattern generally held across all varieties, the Phrasal condition exhibits its lowest mutuality, and therefore its most extreme imbalance, in New Zealand English. Collectively, these results highlight the importance of carefully designing LLM-assistance paradigms to ensure that the LLM serves as a true partner rather than a primary author.

## 4. Corpus Analysis

To illustrate the research potential of this dataset, we present a set of analyses focusing on lexical diversity, syntactic complexity, and stylistic variation. These analyses are intended as demonstrations of the types of questions the corpus enables rather than as exhaustive theoretical accounts. Specifically, we examine two questions: (1) whether LLM-assisted writing is associated with changes in lexical diversity, syntactic complexity, and stylistic variation, and whether such changes persist from Session 2 into Session 3; and (2) whether different forms of LLM assistance correspond to distinct patterns across dialects and over time.

**Lexical Diversity** is operationalized using the Measure of Textual Lexical Diversity (MTLD; Jarvis (2013); McCarthy and Jarvis (2010); Nasseri and Thompson (2021)). MTLD was developed to address the sensitivity of traditional type-token ratio (TTR) measures to text length. It estimates the average length of word sequences that maintain a minimum TTR threshold, set at 0.72 here following McCarthy and Jarvis (2010). Specifically, the algorithm processes the text word by word, tracking the ratio of unique words (types) to total words (tokens). When the TTR falls below 0.72, one MTLD "factor" is completed and the count restarts. The final MTLD score represents the mean length of these factors, with higher values indicating greater lexical diversity. Let a text consist of a sequence of tokens $s = w_1, w_2, \ldots, w_N$. The TTR at position $i$ is:

$$TTR_i = \frac{V(w_1, w_2, \ldots, w_i)}{i} \quad (1)$$

where $V(s)$ is the number of types in $s$. Let the TTR threshold be $\tau = 0.72$. The algorithm progresses through the text, incrementing $i$ until $TTR_i < \tau$; when this occurs, one factor is completed. If the text yields $F$ complete factors and a partial remainder proportion $R$, the MTLD is computed as:

$$MTLD = \frac{N}{F + R} \quad (2)$$

where $N$ is the total number of tokens.

To investigate whether LLM assistance increases lexical diversity and whether this effect persists over time, we fitted a linear mixed-effects regression model (Bates et al., 2015) for the three LLM-assisted conditions with Session as the predictor (1st vs. 2nd vs. 3rd) and MLTD as the dependent variable and Participant and Topic as random effects. This set-up accounts for the repeated-measures design (Baayen et al., 2008), in which each participant wrote three essays on three topics.

As shown in Table 5, the model reveals a significant effect of Session on MLTD, indicating that LLM-assisted writing influenced lexical diversity. As shown in Figure 3, post-hoc pairwise comparisons using Tukey adjustment (Lenth, 2025) indicate that the LLM-assisted essays had significantly higher MLTD than both the first natural essays ($\beta = -20.0$, $p < .001$) and the third natural essays, written three weeks later ($\beta = 15.5$, $p < .001$).Consistently, the first and third natural essays did not differ significantly in MLTD ($\beta = -4.5$, $p = .299$).

| Predictors | Estimate ($\beta$) | $t$ value | $p$ |
|---|---|---|---|
| (Intercept) | 84.08 | 54.38 | <0.001 |
| Session [2nd] | 20.00 | 16.40 | <0.001 |
| Session [3rd] | 4.53 | 2.01 | =0.098 |

Table 5: Model output for MLTD by Session, combining the three LLM-assisted conditions. Random effects: Participant ($SD = 12.5$), Topic ($SD = 2.0$). Reference = 1st.

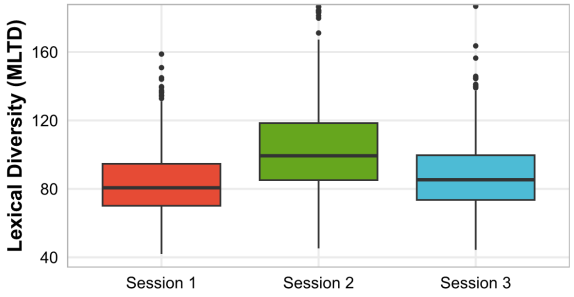


Figure 3: Changes in MLTD across sessions, combining the three LLM-assisted conditions.

A natural follow-up question is whether different types of LLM assistance influence lexical diversity differently across dialects and sessions. To examine this, we fitted a new mixed-effects model with MLTD as the outcome variable and the interaction among Condition (GPT, natural, stylistic, phrasal, paragraph), Dialect (US, CAN, UK, AUS, NZ), and Session as fixed effects, with random intercepts for Participant and Topic. The analysis reveals a significant three-way interaction between these factors on MLTD ($p < .05$) (the full model output is summarized in Appendix D). Post-hoc comparisons (visualized in Figure 4) indicate that in the purely synthetic and purely natural conditions, the two baselines, lexical diversity did not differ significantly across sessions for any variety. In the stylistic condition, Session 2 showed significantly higher lexical diversity than Session 1 for British, Australian, and New Zealand English (all $p < .05$), but not for the two North American varieties, suggesting region-specific sensitivity to stylistic assistance. In the phrasal condition, lexical diversity increased significantly from Session 1 to Session 2 in American and British English (both $p < .05$) but remained

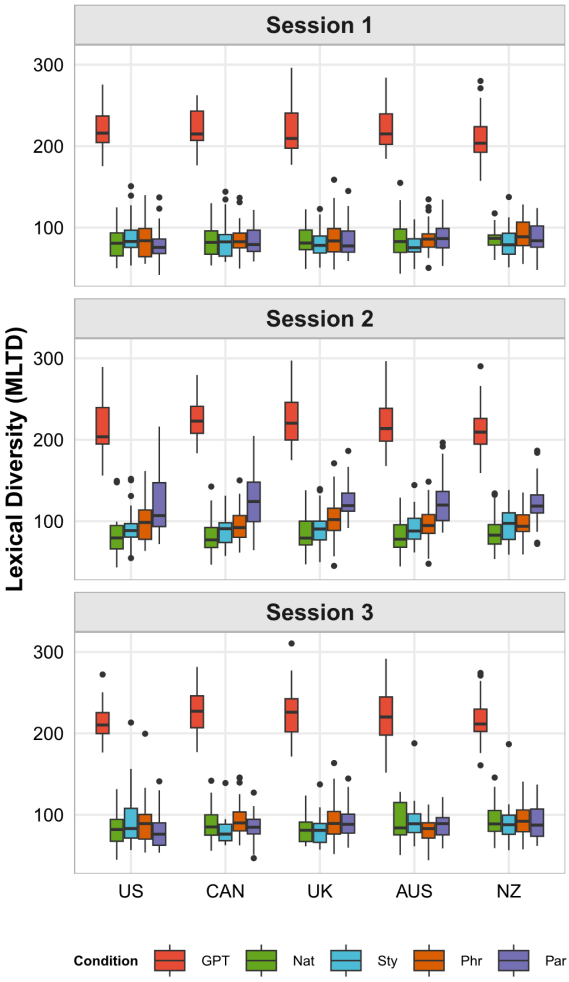


Figure 4: MLTD by Session, Dialect and Condition. Higher MLTD indicates greater lexical diversity.

stable for the other three varieties. The paragraph condition displayed the most pronounced changes: Session 1 vs. Session 2 and Session 2 vs. Session 3 contrasts were significant across all varieties (all $p < .001$).

Moving from lexical to syntactic complexity, we operationalize eight measures derived from dependency parses generated by spaCy's *en_core_web_sm* model (v3.3.3; Honnibal et al. 2020), which employs a transition-based parser trained on the Universal Dependencies English Web Treebank. These measures reflect both local syntactic choices and global structural properties.

**Phrasal Complexity.** To measure phrasal complexity, we use Mean Noun Phrase Length (MNPL), which quantifies the average number of tokens within noun phrases, calculated across all noun chunks identified by the parser. While MNPL captures overall phrase size, NP Complexity (NPC) specifically measures the internal structural sophistication of noun phrases by counting the number of syntactic modifiers within each NP (Lu, 2010). We operationalize this as the average count of adjectival modifiers (*amod*), nominal compounds (*compound*), prepositional phrase attachments (*prep*), relative clauses (*relcl*), adjectival clauses (*acl*), and numerical modifiers (*nummod*) per noun phrase. We also compute VP Complexity (VPC) (Kyle and Crossley, 2017), which captures the size and elaboration of verb phrases by measuring the average number of tokens within the syntactic sub-tree headed by each verb. This includes the main verb along with all its dependents such as auxiliaries, objects, adverbials, and clausal complements.

Additionally, Prepositional Phrase Density (PPD) measures the average number of prepositional phrases per sentence, calculated by counting all tokens tagged as adpositions (ADP) and normalizing by sentence count. Prepositional phrases serve multiple functions in English, including indicating spatial and temporal relationships, expressing abstract relations, and post-modifying nouns. Higher PP density reflects greater use of these complex relationships and is characteristic of more formal, academic registers (Biber et al., 2011).

Finally, Nominalization Density (ND) measures the frequency of deverbal and deadjectival nouns per 100 words; these are nouns morphologically derived from verbs or adjectives. Nominalizations are identified through morphological patterns, specifically nouns ending in suffixes such as *-tion, -sion, -ment, -ness, -ity, -ance, -ence, -ancy, -ency, -ship, -ism, -acy, -ure, -al, -age, -ery*, and *-ry*. Nominalization is a key feature of formal registers, allowing writers to package processes and qualities as entities and thereby increase informational density and promote more abstract discourse (Halliday, 2004).

**Clausal and Dependency Complexity.** We included three measures in this category. Mean Clause Length (MCL) captures the average number of words per clause, where clauses include both main clauses (sentences) and subordinate clauses. Subordinate clauses were identified through dependency relations such as clausal complements (*ccomp*), open clausal complements (*xcomp*), adverbial clauses (*advcl*), adjectival clauses (*acl*), and relative clauses (*relcl*). The total word count was divided by the sum of main and subordinate clauses to compute this measure, capturing the density of information packaging within clausal units.

Average Dependency Length (ADL) quantifies the mean linear distance between syntactic heads and their dependents, measured in word positions (Futrell et al., 2015; Liu, 2008). For each token, we calculated the absolute difference between its position and its head's position, then averaged across all non-root dependencies in the text.

Dependency Type Entropy (DTE) captures the diversity of syntactic relations used in a text, calculated as the Shannon entropy (Shannon, 1948) of the distribution of dependency relation types:

$$H = -\sum_{i} p(i) \cdot \log_2 p(i) \tag{3}$$

where $p(i)$ is the proportion of each dependency type in the text. Higher entropy indicates more varied and balanced use of syntactic constructions, while lower entropy suggests reliance on a limited set of constructions (Shimorina et al., 2021). DTE complements frequency-based metrics by capturing structural diversity rather than simple counts.

To address the two research questions, we first ran Spearman correlation analyses to explore the relationships among these syntactic measures. Correlations varied considerably: for example, Spearman $\rho$ was $0.83$ for NPC–MNPL, $0.76$ for VPC–PPD, and $0.60$ for ADL–PPD, whereas MCL–DTE ($\rho = -0.09$) and ADL–ND ($\rho = -0.04$) were largely uncorrelated.

To capture a composite measure, we conducted Principal Component Analysis (PCA) on all measures. The first principal component (PC1) accounted for $39\%$ of the total variance, with the highest positive loadings observed for PPD ($0.43$), MCL ($0.42$), VPC ($0.41$), MNPL ($0.39$), ADL ($0.38$), and NPC ($0.37$), suggesting that PC1 reflects a general dimension of syntactic complexity. Higher PC1 scores indicate texts that are denser, longer, and more syntactically complex.

We then fitted a mixed-effects regression model with PC1 as the dependent variable, Session as the fixed effect, and Participant and Topic as random intercepts. Session had a significant effect on PC1 ($p < .001$). As shown in Figure 5, post-hoc pairwise comparisons revealed that LLM-assisted essays in Session 2 were syntactically simpler than those in

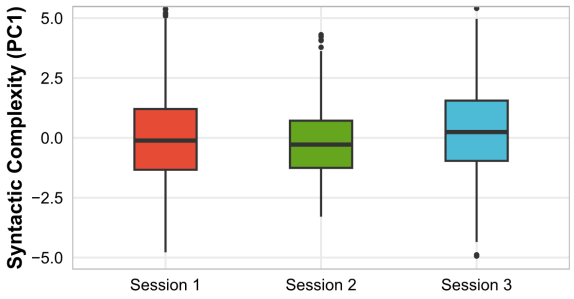


Figure 5: Changes in PC1 scores across sessions, combining the three LLM-assisted conditions.

Session 1 ($\beta = 0.3$, $p < .001$). However, no significant differences were found between Sessions 1 and 3 or between Sessions 2 and 3.

To examine whether different types of LLM assistance influenced syntactic complexity differently across dialects and sessions, we fitted a new model for the interaction among Condition (GPT, natural, stylistic, phrasal, paragraph), Dialect, and Session, with random intercepts for Participant and Topic. The analysis revealed a significant three-way interaction among these factors on PC1 ($p < .05$) (the full model is summarized in Appendix D).

Post-hoc comparisons (visualized in Figure 6) revealed distinct patterns of syntactic complexity across conditions and varieties. In the GPT condition, which served as the LLM baseline, syntactic complexity did not differ significantly across sessions for any variety. This consistency suggests that the LLM persona of a dialect speaker maintains stable syntactic structures across repeated generations, independent of sessions or topics. In the natural writing condition, no significant differences were found across sessions for American, Canadian, British, or Australian English. However, New Zealand English writers produced a more complex third essay compared to the second ($\beta = -0.98$, $p < .05$).

In the stylistic condition, syntactic complexity showed modest variability. For American English, essays in Session 2 were significantly simpler than those in Session 3 ($\beta = -1.23$, $p < .01$), while the difference between Sessions 1 and 2 was not significant ($\beta = 0.63$, $p = .09$). A similar trend was observed for Canadian English, where Session 2 essays were the simplest overall, differing significantly from Session 3 ($\beta = -1.27$, $p < .01$). No significant session effects were found for British or Australian English, whereas New Zealand English again showed a steady increase in syntactic complexity over time, with Session 3 essays significantly more complex than both Session 1 ($\beta = -1.36$, $p < .01$) and Session 2 ($\beta = -1.62$, $p < .01$). In the

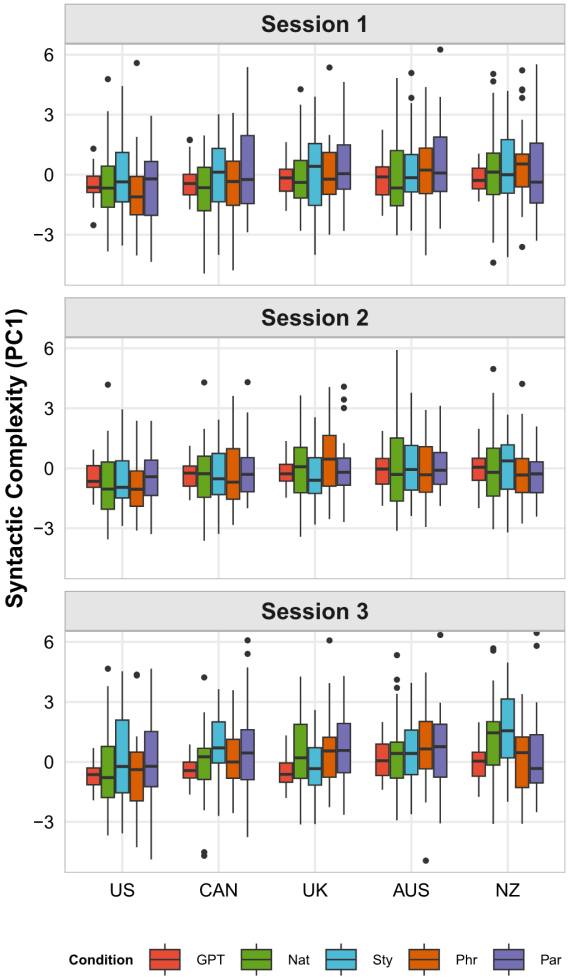


Figure 6: Syntactic complexity (PC1 scores) by Session, Dialect and Condition. Higher PC1 values indicate greater syntactic complexity.

phrasal condition, syntactic complexity remained stable across sessions for American, Canadian, British, and Australian English. In contrast, New Zealand English exhibited a significant reduction in syntactic complexity in Session 2 compared to Session 1 ($\beta = 0.78$, $p < .05$), reinforcing the pattern that New Zealanders produced simpler syntax under LLM assistance than in natural writing.

Finally, in the paragraph condition, syntactic complexity did not differ significantly across sessions for American, Canadian, Australian, or New Zealand English. However, British writers produced significantly simpler LLM-assisted essays than their third natural essays ($\beta = -0.95$, $p < .05$).

Taken together, these results could indicate that LLM-generated texts remain relatively stable across varieties and sessions, whereas human writers show greater variability in syntactic complexity. In particular, patterns observed for New Zealand English suggest that dialect-specific trajectories may interact with LLM assistance in ways that warrant

further sociolinguistic investigation.

**Stylistic Variation via Nominalizations.** Building on the analyzes of lexical and syntactic complexity, we next examine stylistic variation through patterns of nominalization, operationalized as preferences for nominal versus verbal realizations of lexemes. Nominalization is a well-established marker of register and stylistic formality, associated with increased abstraction, informational density, and impersonal stance in academic and institutional discourse (Halliday, 1994). By recasting processes as entities, nominalization enables clause compression and complex noun-phrase packaging, making it particularly sensitive to differences in discourse style.

Here, we use nominalization patterns to test whether stylistic preferences differentiate human writing from LLM-assisted production, focusing on the Session 2 dataset, where degrees of human–LLM interaction vary systematically. Compared to other derivational processes, which do not consistently index register variation, nominalization provides a theoretically motivated and empirically tractable indicator of stylistic differences between human and synthetic authorship.

Recent work suggests that synthetic production differs in the characteristics of its noun phrases (Dentella et al., 2025). We operationalize this hypothesis in two ways: first, that texts with a greater degree of synthetic authorship contain more nouns; and second, that they show a preference for nominal rather than verbal forms of specific lexemes.

The first question was examined using a regression model with Noun Frequency as the outcome variable and Condition as the predictor. The results, visualized in Figure 7, showed that compared to LLM-generated texts, both natural writing and LLM-assisted paragraph-continuation texts used nouns more frequently (both $p < .001$). In contrast, the stylistic and phrasal conditions, which involved greater human-LLM interaction as described in Section 3, did not differ from LLM texts in noun frequency. Thus, in our dataset, synthetic texts do not contain more nouns than natural or hybrid texts as observed in previous studies.

Given that synthetic texts do not contain more nouns overall, it remains possible that they differ in the types of nominalizations they employ. Many lexemes can appear as nominal or verbal forms (e.g., *acceptance* vs. *accept*). To establish a baseline for contemporary usage, we used large background corpora (Wikipedia, tweets, news comments, and blogs; Dunn 2020) to determine, for each lexeme, the proportion of occurrences in nominal form. For example, *adore* appears nominally in $60\%$ of cases, whereas *aim* appears nominally in only $30\%$. Based on these baseline probabilities, we categorized lexemes into three nominalization classes for the Session 2 dataset: low ($< 20\%$ nominal usage), mid ($20$-$40\%$), high ($40$-$60\%$). Previous work suggests that synthetic texts should show higher rates of nominalization in low-nominal lexemes, reflecting a tendency to encode information in nouns rather than verbs (Dentella et al., 2025).

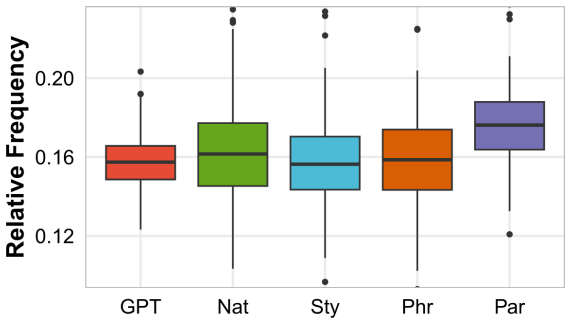


Figure 7: Noun frequency by Condition in Session 2, combining all dialects within each condition.

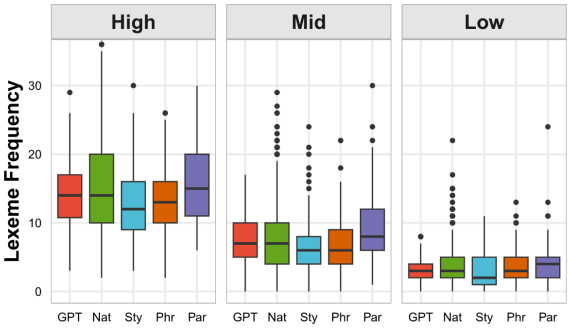


Figure 8: Lexeme frequency by Category in Session 2, combining all dialects within each condition. High indicates lexemes usually used as nouns and Low indicates lexemes usually not used as nouns.

To examine nominalization patterns across lexeme categories, we fitted a regression model with Lexeme Frequency as the outcome variable and the interaction between Condition and Category as predictors. The results revealed a significant interaction, prompting post-hoc comparisons to identify specific contrasts among LLM-generated, natural, and hybrid texts. As shown in Figure 8, for the high nominalization class, nominalization frequency in LLM texts was lower than in natural writing and paragraph-continuation essays (both $p < .001$), but did not differ from essays in the stylistic and phrasal conditions, consistent with the patterns for overall noun frequency. In the mid nominalization class, LLM texts showed significantly lower nominalization frequency than paragraph-continuation essays ($\beta = -2.04$, $p < .001$), while remaining comparable to the other groups (all $p > .05$). For the low nominalization class, nominalization frequency in LLM texts was significantly lower than in natural writing

($\beta = -0.92$, $p < .05$), but comparable to the other three conditions. Overall, nominalization frequency was generally higher in less synthetic texts, contrary to the hypothesized pattern. This suggests that either previous work has overestimated the tendency of synthetic texts to use nominalizations, or that this effect interacts with register. Moreover, purely natural texts exhibited greater variability across lexemes, indicating a wider range of nominalization patterns than either partially or fully synthetic texts.

## 5. Conclusion

In this paper, we introduced a multi-dialectal, longitudinal corpus of human–AI hybrid writing across assistance conditions. The analyses presented here illustrate the types of linguistic questions this dataset enables, demonstrating how interaction strength, dialect, and time jointly shape hybrid writing outcomes. Across measures, LLM-assisted writing was associated with increased lexical diversity and a lack of increase in syntactic complexity, although these patterns varied by condition and dialect. These findings suggest that the impact of LLM assistance on linguistic diversity is not uniform but mediated by interaction mode and sociolinguistic background.

Importantly, the corpus provides controlled longitudinal evidence that complements prior single-session studies and offers a resource for examining longer-term adaptation in human–AI writing practices. Future research can build on this dataset to disentangle the roles of topic choice, learning effects, demographic factors, and sustained LLM exposure in shaping language production over time.

## 6. Ethics Statement and Limitations

This study was reviewed and approved by the Human Ethics Committee at the University of Canterbury, New Zealand (HREC 2024/170/LR-PS). All participants provided informed consent prior to participation and were compensated at a standard hourly rate approved by the Committee and consistent with Prolific's fair payment policy. All data were anonymized to protect participant privacy.

The present design captures the effects of a single episode of LLM-assisted writing followed by a three-week interval without continued exposure, and therefore may underestimate longer-term stylistic adaptation from sustained use. Some findings, particularly the relative complexity of LLM-generated versus human-written texts, differ from prior reports, potentially reflecting differences in methodology, corpus composition, or recent LLM fine-tuning.

Another limitation concerns demographic factors in explaining dialectal differences. While the dataset includes detailed participant metadata (e.g., education, age, occupation), these variables were not incorporated into the regression models to focus on cross-condition and dialect effects. This choice allowed us to minimize model complexity in an initial corpus study, but some observed differences, such as patterns in New Zealand English, may partially reflect demographic variation rather than purely linguistic factors. Future work will integrate demographic predictors and examine repeated interactions over extended periods to disentangle sociolinguistic influences and assess persistent effects of LLM assistance.

## 7. Acknowledgements

This research was supported by a Royal Society of New Zealand Marsden Research Grant (23-UOC-048).

## 8. Data and Code Availability

Our corpus, along with the Python and R scripts and associated datasets, is available at https://github.com/GanQiaoUC/LREC_MulDiLoC-HAP. The International Standard Language Resource Number (ISLRN) of our corpus is 119-185-989-055-2.

## 9. Bibliographical References

Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2025. AI suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

R. Harald Baayen, David J. Davidson, and Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67:1–48.

Douglas Biber, Bethany Gray, and Kornwipa Poonpon. 2011. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, 45(1):5–35.

Hongbo Chen and Ben He. 2013. Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1741–1752.

Scott A Crossley, Perpetual Baffour, Yu Tian, Aigner Picou, Meg Benner, and Ulrich Boser. 2022. The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (PERSUADE) corpus 1.0. *Assessing Writing*, 54:100667.

Scott A Crossley, Yu Tian, Perpetual Baffour, Abigail Franklin, Meg Benner, and Ulrich Boser. 2024. A large-scale corpus for assessing written argumentation: Persuade 2.0. *Assessing Writing*, 61:100865.

Vittoria Dentella, Weihang Huang, Silvia Angela Mansi, Jack Grieve, and Evelina Leivada. 2025. ChatGPT-generated texts show authorship traits that identify them as non-human. *arXiv preprint arXiv:2508.16385*.

Jonathan Dunn. 2020. Mapping languages: The corpus of global language use. *Language Resources and Evaluation*, 54(4):999–1018.

Jonathan Dunn, Benjamin Adams, and Harish Tayyar Madabushi. 2024. Pre-trained language models represent some geographic populations better than others. In *Joint 30th International Conference on Computational Linguistics and 14th International Conference on Language Resources and Evaluation, LREC-COLING 2024*, pages 12966–12976. European Language Resources Association (ELRA).

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences of the United States of America*, 112(33):10336–10341.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. The curious decline of linguistic diversity: Training language models on synthetic text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, Mexico City, Mexico. Association for Computational Linguistics.

M. A. K. Halliday. 1994. *An Introduction to Functional Grammar*, 2 edition. Edward Arnold, London.

M. A. K. Halliday. 2004. *The Language of Science*. Continuum, London, England.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. 2020. spaCy: Industrial-strength natural language processing in Python.

Scott Jarvis. 2013. Capturing the diversity in lexical diversity. *Language Learning*, 63:87–106.

Kristopher Kyle and Scott Crossley. 2017. Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, 34(4):513–535.

Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a human-AI collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Florian Lehmann and Daniel Buschek. 2025. StudyAlign: A software system for conducting web-based user studies with functional interactive prototypes. *Proceedings of the ACM on Human-Computer Interaction*, 9(4):1–26.

Russell V. Lenth. 2025. emmeans: Estimated marginal means, aka least-squares means. R package version 1.11.2-8. https://doi.org/10.32614/CRAN.package.emmeans.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

Philip M McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Maryam Nasseri and Paul Thompson. 2021. Lexical density and diversity in dissertation abstracts: Revisiting English L1 vs. L2 text differences. *Assessing Writing*, 47:100511.

Vishakh Padmakumar and He He. 2024. Does writing with language models reduce content diversity? In *Proceedings of the 2024 International Conference on Learning Representations (ICLR)*, Vienna, Austria.

Aristotelis Ioannis Paschalidis. 2025. AI and the great linguistic flattening. Available at: https://www.unesco.org/en/articles/ai-and-great-linguistic-flattening.

Arjun Prakash, Shruti Aggarwal, Jeevan John Varghese, and Joel John Varghese. 2025. Writing without borders: AI and cross-cultural convergence in academic writing quality. *Humanities*

*and Social Sciences Communications*, 12(1):1–11.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Anastasia Shimorina, Yannick Parmentier, and Claire Gardent. 2021. An error analysis framework for shallow surface realization. *Transactions of the Association for Computational Linguistics*, 9:429–446.

Dongkwang Shin and Jang Ho Lee. 2025. Leveraging LLM-based chatbots for interactional grammar feedback in L2 writing: opportunities and challenges. *Innovation in Language Learning and Teaching*, pages 1–13.

Neomy Storch. 2002. Patterns of interaction in ESL pair work. *Language Learning*, 52(1):119–158.

Ben Swanson, Kory Mathewson, Ben Pietrzak, Sherol Chen, and Monica Dinalescu. 2021. Story Centaur: Large language model few shot learning as a creative writing tool. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 244–256, Online. Association for Computational Linguistics.

Aigerim Utepbayeva, Nadezhda Zhiyenbayeva, Leila Assylbekova, and Olga Tapalova. 2024. Artificial intelligence applications (Fluency SIS, Articulation Station Pro, and Apraxia Farm) in the psycholinguistic development of preschool children with speech disorders. *International Journal of Information and Education Technology*, 14(7):927–935.

Xinli Zhang, Ruiting Huang, Ruihua Zhang, Mingyi Li, Yun-Fang Tu, Yuchen Chen, Lailin Hu, and Gwo-Jen Hwang. 2025. AI-facilitated article revisions for primary school students with writing difficulties: Effects of a large language model-based PDRPE approach on writing performance, attitude and anxiety. *Journal of Computer Assisted Learning*, 41(5):e70131.

# Appendices

## A. Topics and Their Rating Scores

To refine the formal study design, we first ran a pilot to test (i) which topics would engage participants and (ii) a feasible writing length in 30 minutes.

Ten prompts were selected from the Reddit/Writing forum, as shown below. Then, 60 American English speakers were recruited via Prolific (U.S. residence, L1 English) to complete two tasks in Qualtrics. First, they rated their willingness to write on each topic (1-10 scale). Then, each wrote two essays on randomly assigned topics: 30 participants were asked to write around 300 words per essay, and 30 wrote around 500 words. These word limits were chosen based on prior essay corpora (Lee et al., 2022; Padmakumar and He, 2024). The results showed that mean topic ratings ranged from 4.8 to 7.0, and we selected the five most preferred topics. Writing times indicated that two 500-word essays ($M = 549$ words/essay, $SD = 52$) were achievable in about an hour ($M = 62$ minutes, $SD = 20.0$) (two 300-word essays: $M = 339$ words/essay, $SD = 40$; $M = 56$ minutes, $SD = 28.4$), so we adopted 500 words as the target length in the formal study.

1. Ever lie awake at night thinking about that big dream you're chasing? What's that goal that keeps you motivated, and what are you doing to get there? Share your journey, the ups and downs, and what that dream means to you like you're posting on a forum. (score: 7.0)

2. Think about the first time you got hooked on a new gadget or app – it felt awesome, right? But over time, maybe it started messing with your routine or even your relationships. How has tech changed your everyday life, for better or worse? Share your story like you're posting on a forum. (score: 6.8)

3. Ever get frustrated by people assuming things about you because of your job or hobby? Maybe someone judged you based on a stereotype that just isn't true. How did that make you feel, and what's the real story behind your passion? Share your thoughts like you're posting on a forum. (score: 6.6)

4. Ever find yourself disagreeing with a widely accepted opinion just because you see things differently? Think about that time when your perspective clashed with the crowd. How did you handle it, and what's your take on the issue? Share your story like you're posting on a forum. (score: 6.5)

5. Imagine you've achieved financial independence through your job, but now you're dealing with personal challenges and aging parents, and you're thinking about leaving the workforce. You've got hobbies like making video games, but you're also wondering if jumping back into the industry later would be a hassle. If you were in this situation, how would you decide—focusing on your own well-being, taking care of family, or staying active in your career? Share your thoughts like you're posting on a forum. (score: 6.1)

6. Think back to a time when one moment flipped your world around – a decision or event that changed how you see everything. How did it make you feel at the time, and what did you learn from it? Share your story like you're posting on a forum. (score: 5.9)

7. Imagine you could jump in a time machine and witness a historical event firsthand. Which moment would you choose, and what draws you to that piece of history? Share your time-travel daydream in a forum-style post. (score: 5.6)

8. Picture this: You see a woman who always walks her kid to school, and every time she passes a young person, she ends up snapping at them. One day, she goes off on a teenager just chilling to his music, and you catch the whole thing. What would you do if you were there? Do you think people should be held more accountable for how they treat strangers? Share your thoughts like you're posting on a forum. (score: 5.6)

9. Imagine you're driving late at night and you see a woman in distress with an unresponsive body nearby. You get scared, drive past, and call emergency services afterward—but the guilt of not stopping still lingers. Have you ever been in a situation where you didn't act the way you wished? What would you do if you found yourself there? Share your thoughts like you're posting on a forum. (score: 5.0)

10. Imagine you're home sick when a recruiter calls out of the blue. Caught off guard, you panic and blurt out an honest but not-so-great answer about why you're job hunting. The call ends, and you realize you might've left a bad impression. Would you reach out to clear things up or just move on and take it as a lesson? How would you bounce back from a mistake like this? Share your thoughts like you're posting on a forum. (score: 4.8)

## B. System Settings

**Server**

Following the CoAuthor interface (Lee et al. 2022), our server was written in Python using Flask and adapted to include additional AI-assistance conditions (https://github.com/AI-Assisted-Writing-Impact/coauthor-interface). It is used to request suggestions and record events during a writing session. The system is deployed on our institution's infrastructure.

**Decoding parameters**

The following decoding parameters for GPT-4.1 were used to generate suggestions:

- temperature = 0.75
- presence_penalty = 0.3
- frequency_penalty = 0.5
- Top P = 1

## C. Survey Questions

**Demographic questions**

1. What is your age?
2. What is your gender?
3. What is the highest level of education you have completed?
4. What is your current occupation?
5. How would you describe your ethnicity?
6. Which variety of English do you primarily speak (e.g., American, British, Australian English)?
7. What other languages can you comfortably hold a conversation in?
8. In which city and country are you currently located?
9. Please list any cities and countries where you have lived for at least 5 years, if they are at least 100km away from your current location.

**AI-usage and attitude questions**

1. How familiar are you with artificial intelligence (AI)? (Single choice)
   - Not familiar at all
   - Slightly familiar
   - Moderately familiar
   - Very familiar
   - Extremely familiar
2. Which AI tools have you used in the past 3 months? (Multiple choice)
   - ChatGPT
   - Google Bard
   - Claude
   - Llama
   - DALL-E
   - Copilot
   - DeepSeek
   - Others (please specify)

- I haven't used any AI tools

3. How frequently do you use AI tools? (Single choice)
   - Multiple times per day
   - Daily
   - Several times a week
   - Weekly
   - Monthly
   - Rarely
   - Never
4. What are your primary purposes for using AI? (Multiple choice)
   - Writing assistance
   - Coding help
   - Research/Information gathering
   - Creative projects
   - Problem-solving
   - Learning/studying
   - Entertainment
   - Work tasks
   - Other (please specify)
5. How would you rate your overall opinion of AI? (Single choice)
   - Very negative
   - Negative
   - Neutral
   - Positive
   - Very positive

## D. Regression Model Outputs

Table 6: Linear mixed-effects model predicting MLTD from the interaction of Condition, Variety and Phase, with random intercepts for PID and topics. Reference levels: GPT for Condition, US for Variety, and 1st for Phase. Variance inflation factors were all below 5, indicating no multicollinearity.

| **Predictors** | **Estimate ($\beta$)** | $p$ |
|---|---|---|
| (Intercept) | 222.32 | <0.001 |
| ConditionNat | −141.22 | <0.001 |
| ConditionSty | −134.63 | <0.001 |
| ConditionPhr | −135.34 | <0.001 |
| ConditionPar | −140.97 | <0.001 |
| VarietyCAN | −1.23 | 0.820 |
| VarietyUK | −0.96 | 0.860 |
| VarietyAUS | −0.29 | 0.958 |
| VarietyNZ | −13.73 | <0.05 |
| Phase2nd | −6.09 | 0.197 |
| Phase3rd | −9.24 | <0.05 |
| ConditionNat:VarietyCAN | 3.75 | 0.640 |
| ConditionSty:VarietyCAN | −4.25 | 0.580 |
| ConditionPhr:VarietyCAN | −1.94 | 0.800 |
| ConditionPar:VarietyCAN | 3.22 | 0.674 |
| ConditionNat:VarietyUK | 2.77 | 0.729 |
| ConditionSty:VarietyUK | −7.35 | 0.335 |
| ConditionPhr:VarietyUK | −0.63 | 0.934 |
| ConditionPar:VarietyUK | 2.46 | 0.748 |
| ConditionNat:VarietyAUS | 5.71 | 0.475 |
| ConditionSty:VarietyAUS | −10.26 | 0.181 |
| ConditionPhr:VarietyAUS | −1.18 | 0.877 |
| ConditionPar:VarietyAUS | 6.49 | 0.395 |
| ConditionNat:VarietyNZ | 16.69 | <0.05 |
| ConditionSty:VarietyNZ | 7.87 | 0.307 |
| ConditionPhr:VarietyNZ | 18.06 | <0.05 |
| ConditionPar:VarietyNZ | 17.80 | <0.05 |
| ConditionNat:Phase2nd | 6.96 | 0.315 |
| ConditionSty:Phase2nd | 9.41 | 0.158 |
| ConditionPhr:Phase2nd | 19.08 | <0.01 |
| ConditionPar:Phase2nd | 45.37 | <0.001 |
| ConditionNat:Phase3rd | 9.88 | 0.207 |
| ConditionSty:Phase3rd | 14.79 | <0.05 |
| ConditionPhr:Phase3rd | 13.38 | 0.069 |
| ConditionPar:Phase3rd | 7.77 | 0.291 |
| VarietyCAN:Phase2nd | 11.00 | 0.099 |
| VarietyUK:Phase2nd | 10.45 | 0.117 |
| VarietyAUS:Phase2nd | 3.61 | 0.588 |
| VarietyNZ:Phase2nd | 9.72 | 0.145 |
| VarietyCAN:Phase3rd | 16.95 | <0.05 |
| VarietyUK:Phase3rd | 13.74 | <0.05 |
| VarietyAUS:Phase3rd | 9.99 | 0.134 |
| VarietyNZ:Phase3rd | 16.40 | <0.05 |
| ConditionNat:VarietyCAN:Phase2nd | −14.00 | 0.155 |
| ConditionSty:VarietyCAN:Phase2nd | −6.66 | 0.480 |
| ConditionPhr:VarietyCAN:Phase2nd | −13.77 | 0.143 |
| ConditionPar:VarietyCAN:Phase2nd | −7.31 | 0.438 |
| ConditionNat:VarietyUK:Phase2nd | −9.48 | 0.336 |
| ConditionSty:VarietyUK:Phase2nd | −1.99 | 0.832 |
| ConditionPhr:VarietyUK:Phase2nd | −8.01 | 0.394 |
| ConditionPar:VarietyUK:Phase2nd | −8.54 | 0.363 |
| ConditionNat:VarietyAUS:Phase2nd | −8.97 | 0.362 |
| ConditionSty:VarietyAUS:Phase2nd | 6.77 | 0.473 |
| ConditionPhr:VarietyAUS:Phase2nd | −6.27 | 0.506 |
| ConditionPar:VarietyAUS:Phase2nd | −5.48 | 0.559 |
| ConditionNat:VarietyNZ:Phase2nd | −7.84 | 0.424 |
| ConditionSty:VarietyNZ:Phase2nd | 0.20 | 0.983 |
| ConditionPhr:VarietyNZ:Phase2nd | −18.34 | 0.053 |
| ConditionPar:VarietyNZ:Phase2nd | −12.77 | 0.176 |
| ConditionNat:VarietyCAN:Phase3rd | −11.76 | 0.251 |
| ConditionSty:VarietyCAN:Phase3rd | −23.55 | <0.05 |
| ConditionPhr:VarietyCAN:Phase3rd | −10.12 | 0.292 |
| ConditionPar:VarietyCAN:Phase3rd | −13.86 | 0.148 |
| ConditionNat:VarietyUK:Phase3rd | −15.32 | 0.132 |
| ConditionSty:VarietyUK:Phase3rd | −17.20 | 0.073 |
| ConditionPhr:VarietyUK:Phase3rd | −8.98 | 0.347 |
| ConditionPar:VarietyUK:Phase3rd | −3.93 | 0.681 |
| ConditionNat:VarietyAUS:Phase3rd | −6.90 | 0.497 |
| ConditionSty:VarietyAUS:Phase3rd | −0.84 | 0.931 |
| ConditionPhr:VarietyAUS:Phase3rd | −18.06 | 0.070 |
| ConditionPar:VarietyAUS:Phase3rd | −6.92 | 0.472 |
| ConditionNat:VarietyNZ:Phase3rd | −10.12 | 0.317 |
| ConditionSty:VarietyNZ:Phase3rd | −13.07 | 0.186 |
| ConditionPhr:VarietyNZ:Phase3rd | −17.81 | 0.070 |
| ConditionPar:VarietyNZ:Phase3rd | −9.94 | 0.309 |
| **Random effects:** | | |
| PID: Variance = 129.07; SD = 11.36 | | |
| Topic: Variance = 9.08; SD = 3.01 | | |

Table 7: Linear mixed-effects model predicting Syntactic Complexity from the interaction of Condition, Variety and Phase, with random intercepts for PID and topics. References: GPT for Condition, US for Variety, and 1st for Phase. Variance inflation factors were all below 5, indicating no multicollinearity.

| **Predictors** | **Estimate ($\beta$)** | $p$ |
|---|---|---|
| (Intercept) | −0.56 | 0.062 |
| Phase2nd | 0.07 | 0.822 |
| Phase3rd | −0.14 | 0.640 |
| ConditionNat | 0.12 | 0.774 |
| ConditionPar | 0.00 | 0.993 |
| ConditionPhr | −0.37 | 0.364 |
| ConditionSty | 0.61 | 0.133 |
| VarietyCAN | 0.13 | 0.744 |
| VarietyUK | 0.36 | 0.379 |
| VarietyAUS | 0.39 | 0.336 |
| VarietyNZ | 0.35 | 0.391 |
| Phase2nd:ConditionNat | −0.32 | 0.469 |
| Phase3rd:ConditionNat | −0.05 | 0.925 |
| Phase2nd:ConditionPar | −0.03 | 0.951 |
| Phase3rd:ConditionPar | 0.46 | 0.309 |
| Phase2nd:ConditionPhr | −0.07 | 0.865 |
| Phase3rd:ConditionPhr | 0.51 | 0.260 |
| Phase2nd:ConditionSty | −0.69 | 0.100 |
| Phase3rd:ConditionSty | 0.74 | 0.104 |
| Phase2nd:VarietyCAN | 0.04 | 0.927 |
| Phase3rd:VarietyCAN | 0.20 | 0.641 |
| Phase2nd:VarietyUK | −0.10 | 0.819 |
| Phase3rd:VarietyUK | −0.14 | 0.732 |
| Phase2nd:VarietyAUS | 0.02 | 0.965 |
| Phase3rd:VarietyAUS | 0.37 | 0.375 |
| Phase2nd:VarietyNZ | 0.11 | 0.803 |
| Phase3rd:VarietyNZ | 0.35 | 0.411 |
| ConditionNat:VarietyCAN | −0.62 | 0.300 |
| ConditionPar:VarietyCAN | 0.80 | 0.164 |
| ConditionPhr:VarietyCAN | 0.48 | 0.404 |
| ConditionSty:VarietyCAN | −0.09 | 0.877 |
| ConditionNat:VarietyUK | 0.01 | 0.990 |
| ConditionPar:VarietyUK | 0.62 | 0.278 |
| ConditionPhr:VarietyUK | 0.81 | 0.153 |
| ConditionSty:VarietyUK | −0.25 | 0.660 |
| ConditionNat:VarietyAUS | −0.01 | 0.990 |
| ConditionPar:VarietyAUS | 0.82 | 0.148 |
| ConditionPhr:VarietyAUS | 0.90 | 0.116 |
| ConditionSty:VarietyAUS | −0.12 | 0.838 |
| ConditionNat:VarietyNZ | 0.32 | 0.592 |
| ConditionPar:VarietyNZ | 0.31 | 0.583 |
| ConditionPhr:VarietyNZ | 1.20 | $<0.05$ |
| ConditionSty:VarietyNZ | −0.10 | 0.863 |
| Phase2nd:ConditionNat:VarietyCAN | 0.88 | 0.159 |
| Phase3rd:ConditionNat:VarietyCAN | 0.82 | 0.209 |
| Phase2nd:ConditionPar:VarietyCAN | −0.56 | 0.349 |
| Phase3rd:ConditionPar:VarietyCAN | −0.31 | 0.615 |
| Phase2nd:ConditionPhr:VarietyCAN | 0.05 | 0.937 |
| Phase3rd:ConditionPhr:VarietyCAN | −0.12 | 0.849 |
| Phase2nd:ConditionSty:VarietyCAN | 0.09 | 0.877 |
| Phase3rd:ConditionSty:VarietyCAN | −0.02 | 0.974 |
| Phase2nd:ConditionNat:VarietyUK | 0.50 | 0.423 |
| Phase3rd:ConditionNat:VarietyUK | 0.74 | 0.250 |
| Phase2nd:ConditionPar:VarietyUK | −0.34 | 0.571 |
| Phase3rd:ConditionPar:VarietyUK | 0.38 | 0.526 |
| Phase2nd:ConditionPhr:VarietyUK | 0.20 | 0.742 |
| Phase3rd:ConditionPhr:VarietyUK | 0.28 | 0.643 |
| Phase2nd:ConditionSty:VarietyUK | 0.23 | 0.700 |
| Phase3rd:ConditionSty:VarietyUK | −0.72 | 0.238 |
| Phase2nd:ConditionNat:VarietyAUS | 0.39 | 0.526 |
| Phase3rd:ConditionNat:VarietyAUS | 0.38 | 0.556 |
| Phase2nd:ConditionPar:VarietyAUS | −0.65 | 0.276 |
| Phase3rd:ConditionPar:VarietyAUS | −0.52 | 0.392 |
| Phase2nd:ConditionPhr:VarietyAUS | −0.43 | 0.476 |
| Phase3rd:ConditionPhr:VarietyAUS | −0.55 | 0.380 |
| Phase2nd:ConditionSty:VarietyAUS | 0.36 | 0.549 |
| Phase3rd:ConditionSty:VarietyAUS | −0.83 | 0.177 |
| Phase2nd:ConditionNat:VarietyNZ | −0.04 | 0.953 |
| Phase3rd:ConditionNat:VarietyNZ | 0.63 | 0.324 |
| Phase2nd:ConditionPar:VarietyNZ | −0.64 | 0.281 |
| Phase3rd:ConditionPar:VarietyNZ | −0.45 | 0.471 |
| Phase2nd:ConditionPhr:VarietyNZ | −0.88 | 0.141 |
| Phase3rd:ConditionPhr:VarietyNZ | −1.27 | $<0.05$ |
| Phase2nd:ConditionSty:VarietyNZ | 0.27 | 0.655 |
| Phase3rd:ConditionSty:VarietyNZ | 0.42 | 0.505 |
| **Random effects:** | | |
| PID: Variance = 1.35; SD = 1.16 | | |
| Topic: Variance = 0.02; SD = 0.14 | | |