

A Corpus-Based Quantitative Analysis of Generative AI Outputs in Second Language Learning: An Applied Linguistics Perspective

Dr. Talal Farhan Alruwaili

Assistant professor, English Department, College of Arts, Jouf university, K.S.A, Tfalruwaili@ju.edu.sa

HOW TO CITE:

Talal Farhan Alruwaili (2026). A
Corpus-Based Quantitative
Analysis of Generative AI Outputs
in Second Language Learning: An
Applied Linguistics Perspective.
International Journal of Special
Education, 41(12s), 97-109

ABSTRACT:

The advent of generative AI (GenAI) systems like large language models (LLMs) has brought a fresh and fast-changing landscape to the field of second language learning (SLL), raising critical concerns regarding the linguistic quality, pedagogical suitability, and evaluation consequences of AI-generated texts. This study is based on an original corpus of 1,240 AI-generated productions that were obtained using three prominent Generative AI (GenAI) tools: GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro, and for five types of tasks typically employed in L2 English classrooms. The research compares the outputs of AI with a set of benchmark corpora of proficient L2 learner writing and native-speaker expert writing in terms of lexical density, syntactic complexity, discourse cohesion, pragmatic appropriateness and phraseological patterning using computational tools from natural language processing (NLP). The results indicate statistically significant differences in lexical sophistication and in the distribution of formulaic items, as well as in the coherence of discourse, between the texts produced by AI and authentic L2 learner output. The deviations have serious consequences for the study of instructed second language acquisition (ISLA), especially when it comes to providing writing feedback with AI, automated assessment, and the integrity of L2 learning tasks. The study provides a replicable approach to assess the outputs of GenAI and suggests an analytical approach in applied linguistics toward the use of GenAI in language education that is principled and takes into account the benefits of technological innovation and the needs to develop communicative competence.

Keywords: generative AI, corpus linguistics, second language learning, large language models, lexical density, syntactic complexity, applied linguistics, NLP, L2 writing, automated assessment

COPYRIGHT STATEMENT:

Copyright: © 2026 Authors.
Open access publication under the
terms and conditions
of the Creative Commons Attribution
(CC BY)
license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

As generative artificial intelligence (GenAI) technologies, including large language models (LLMs) like ChatGPT, Claude, and Gemini, gain traction, they have challenged the traditional

dichotomy between human and machine AI-generated text. This re-examination is especially relevant to second language (L2) education as the key project of language learning is to become an "authentic" competence in authentic communication, and the existence of AI systems

that can generate highly fluent, grammatically correct text in any target language challenges basic concepts of what is meant by learning an L2, practicing it, and evaluating it (Davoodifard & Eskin, 2024; Godwin-Jones, 2024a).

The field of combining Artificial Intelligence with Language Learning (AIL) has experienced exponential growth in the last decade. The bibliometric analyses found a significant increase in publications starting from 2017 and more especially after the public launch of ChatGPT in late 2022 (Rahman et al., 2024; Kartal & Yeşilyurt, 2024, 2025). The effectiveness of AI-based language teaching on various language skills has been explored through systematic reviews and meta-analyses, with mixed results in regards to vocabulary learning, reading comprehension, writing skill, and speaking fluency (Ekizer, 2025; Chen et al., 2025; Lee & Lee, 2024; Xu et al., 2025; Lyu et al., 2025; Gu, 2025). There is, however, a lack of research in the literature that could be described as strictly corpus-linguistic and NLP-based in which the actual linguistic features of AI-generated text are studied as they are used in L2 learning settings.

The gap is significant because there is a growing number of people making consequential decisions based on an assumption about the nature of AI outputted language. Assuming that the outputs of AI are a helpful approximation of real target-language use, a helpful scaffold for writing tasks, and a valid stimulus for reading comprehension tasks, would all of these require empirical support? On the other hand, when the texts that AI tools produce differ systematically and meaningfully from natural language production, these differences must be documented and understood before adopting AI tools on a large scale into language learning programs.

The present study aims to fill this gap by conducting a quantitative analysis of generative AI-generated texts using a corpus of texts generated by L2 English learners in five task types that represent the most common genres in L2 English instruction: argumentative writing, narrative writing, email writing, summary writing,

and dialogue simulation. Using NLP tools that facilitate the computation of lexical density, syntactic complexity, discourse cohesion markers, pragmatic appropriateness, and phraseological patterning, the study compares the features of an original 1,240-text corpus of AI-generated writing to features in benchmark corpora of proficient L2 learner writing and expert native-speaker writing.

The paper is organized as follows. The literature reviewed in section 2 focuses on the literature pertaining to AI in language learning, as well as corpus-based approaches to text analysis. The design and method of analysis of the corpus are described in Section 3. The quantitative results for the five language dimensions examined are provided in Section 4. Implications for L2 instruction, learning and assessment are discussed in Section 5. Section 6 discusses limitations and directions for future research and Section 7 provides a conclusion.

2. Literature Review

2.1 The Expanding Research Landscape: Bibliometric Perspectives

Since the mid-2010s, the body of research in the field of AI in language learning has experienced a significant increase, as evidenced by major bibliometric analyses of the field. Rahman et al. (2024) completed a detailed bibliometric and content analysis of the literature published on AI in language learning between 2017 and 2023 and found that the number of publications has grown exponentially, the research centers have become geographically diverse, and there has been a gradual transformation from technology-mediated communication to more specific themes related to LLM. They found that vocabulary acquisition, automated writing assessment, and speaking practice were the three skill areas that received the most research, and that corpus-based and NLP-based methods were somewhat under-researched compared to the amount of potential they could bring to the field.

Following the bibliometric approach adopted by Kartal and Yeşilyurt (2024, 2025) who carried out a longitudinal study of the publications on AI in

L2 teaching and applied linguistics from 1995 to 2022, the authors added an addendum to this research. They found that the field has been greatly influenced by a limited number of seminal articles, that theories lag behind practice, and that the issue of text quality and linguistic authenticity, which is central to the present work, has not been given systematic enough consideration. These results confirm the importance of the type of methodologically robust and linguistically grounded research that is conducted here.

2.2 Generative AI in L2 Instruction: Affordances and Concerns

L2 instruction can benefit from generative AI systems in a variety of ways. In a similar vein, Godwin-Jones (2024a) explores the notion of distributed agency in GenAI, which he defines as the idea that these technologies are agentic actors in learning systems, thereby sharing cognitive and linguistic workload between people and technology. Based on this, LLMs can deliver language input on demand, at a tailored level of complexity, and provide instant feedback on what the learner produces, and can simulate real communicative context in which to practise pragmatics. The pragmatic and authenticity aspects of the LLM-generated language, though, offer some unique challenges, as Godwin-Jones (2024b) points out, since the LLM creates "statistically plausible" language and not "pragmatically situated" language, and the two are of significant importance for how communicative competence develops in L2 learners. The challenge of embedding GenAI in instruction, learning and assessment cannot be separated, as Davoodifard and Eskin (2024) observe. Conditions for actual acquisition can, at the same time, be undermined by an AI tool that improves instructional efficiency, and threats to assessment validity may not have been fully identified. In the present study, the implications are given attention to all three domains, which is a tripartite framing that informs the focus on implications.

2.3 Meta-Analytic Evidence on AI-Assisted Language Learning

A large number of meta-analysis studies reveal the impact of AI-aided instruction on L2 learning outcomes. Ekizer (2025) performed a meta-analysis of the impact of AI on English language teaching, discovering moderate positive effect sizes within all skill areas, but significant variations in effect sizes that were moderated by the type of task, English level of learners, and the degree of human-AI integration. However, Chen et al. (2025) singled out AI-integrated assessment in the field of language education, reporting that AI-integrated assessment could enhance learners' accuracy of feedback and their engagement, but that studies have failed to pay attention to the linguistic validity of AI-generated assessment criteria. In studying the effectiveness of chatbot-specific interventions, Lyu et al. (2025) conducted comparative studies and reported positive, although limited, effects on L2 acquisition, with moderating factors such as novelty effects and longer duration of intervention reducing effect sizes. In addition to these, positive impacts of AI-L2 learning were echoed by other meta-analyses, such as those conducted by Xu et al. (2025) and Lee and Lee (2024). Gu (2025) specifically explored the role of conversational AI in EFL settings, which revealed that the use of social robots and chatbot systems led to significant improvements in L2 speaking proficiency. The combination of the above-mentioned meta-analyses shows that AI tools can improve some aspects of L2 acquisition, but they do not address the question of whether language created by AI is suitable as a model for learners, which needs to be investigated using the L2 language corpus.

2.4 Corpus Linguistics and NLP in L2 Research

The tradition of Corpus linguistics in applied linguistics is long, successfully supporting empirical arguments for claims of language use, frequency and pattern, as noted by Biber et al. (1998). Within the corpus tradition, a number of concepts, such as lexical density, type-token ratio, mean length of utterance, subordination index

and formulaic language, have been developed and well validated with quantitative measures of textual properties that directly concern pedagogical and assessment issues. NLP tools have been integrated to enable discourse-level, semantic tagging, and parsing at scale, bringing corpus method to a wider range of applications.

These techniques have recently been adapted for L2 texts produced by an AI system. In the field of L2 writing, Etaat (2026) examined the linguistic fingerprints present in both human and AI-generated texts using NLP techniques to identify systematic differences that could be used as diagnostic tools for AI-generated texts. Alqurashi and Hamed (2026) carried out a corpus-based contrastive investigation of the phraseological patterns in ESL development and AI simulation, introducing a new term, "Human–Artificial Intelligence Togetherness" (HAIT), to describe the hybrid text environments in which L2 learners live today. They offer direct examples for the research in hand, as well as imparting the linguistic characteristics that are investigated in the present corpus analysis.

2.5 Research Gaps and Study Rationale

Although the above contributions have been made, there still are some critical gaps. Next, no study has built a large scale, task stratified corpus of AI-generated L2 instructional texts from a variety of LLM systems, nor compared them to L2 learner and native-speaker standards with a wide range of NLP derived linguistic metrics. Secondly, the properties of the phraseological and discourse level of the output generated by GenAI, apart from the sentential level of correctness, have been largely neglected. Thirdly, the implication of the systematic linguistic differences between the AI and the human is not based on the empirical evidence of the corpus. The current study will aim to contribute to overcoming these gaps.

3. Methodology

3.1 Research Design

The study follows a corpus-based quantitative design with a set of AI-generated and human-produced texts, created for the purpose of the

study, for the five different task types. The design is based on the methodological concepts of corpus linguistics, as adapted to comparative computational analysis, namely representativeness of sampling, systematic annotation and statistical comparison of corpus categories. Ethical issues were managed by using synthetic prompts (no learner data was gathered) and the texts produced by the AI were generated under standardized conditions of prompting, for comparability.

3.2 Corpus Construction

3.2.1 AI-Generated Corpus (AIGEN)

The AI-Generated corpus (AIGEN) consists of 3,330 texts generated by three LLM systems: GPT-4o (OpenAI), Claude 3.5 Sonnet (Anthropic), and Gemini 1.5 Pro (Google DeepMind). The texts were created for the following task types: (1) argumentative essay (300-350 words, B2-level prompt), (2) narrative paragraph (150-200 words, B1-level prompt), (3) formal email correspondence (150-200 words, professional register), (4) academic summary (100-150 words, C1-level source text), and (5) simulation of dialogue in a service encounter (300-400 words, service encounter scenario). 82-83 texts per system-task combination were produced with fixed values of temperature ($T = 0.7$), resulting in about 413 texts per LLM and 248 texts per task type. Total AIGEN corpus size: ~342,000 words.

3.2.2 Benchmark Corpora

Two benchmark corpora were used as a reference. The L2 Learner Benchmark (L2LB) is composed of 620 texts from the publicly available L2 English learner writing corpora (ICLE, LOCNESS subset and EF-Cambridge Open Language Database), matched to AIGEN task types as much as possible at the same level of proficiency. The Native Speaker Benchmark (NSB) comprised 310 texts from the British National Corpus and academic writing corpora, which have been chosen to represent formal and semi-formal registers as mapped to the task types in AIGEN. Total benchmark corpus size:

approximately 248,000 words (L2LB: 171,000; NSB: 77,000).

3.3 Linguistic Measures

Based on the NLP tools the linguistic features were operationalized and measured into five categories:

(1) Lexical density and Sophistication: Lexical density (content words/total words), type-token ratio (TTR), advanced vocabulary ratio (AVR, words beyond B2 level/Cambridge grading), mean semantic diversity (using semantic similarity vectors). Syntactic Complexity: Mean length of T-unit (MLTU), subordination index (SI), frequency of coordinate clause, passive voice density, and mean clause length (MCL). (3) Discourse Cohesion: Additive, adversative, causal and temporal cohesive device frequency, Lexical chain length, Pronoun reference density and Given new information ratio.

(4) Pragmatic Appropriateness: Politeness marker frequency, register consistency score (perplexity differential between task-register model and general model), hedging device density and speech act annotation accuracy.

(5) Phraseological Patterning: N-gram frequency alignment with reference corpora, formulaic sequence density (based on the detection of MWE), collocational strength (based on the Mutual Information scores) and lexical bundle inventory overlap with L2LB and NSB.

3.4 Analytical Procedure

Using Python v.3.11, spaCy (v.3.7), NLTK and Gensim, the Corpus texts were pre-processed for tokenization, lemmatization, dependency

parsing, concordancing and N-gram extraction, and also for semantic vector computation. The Cambridge Vocabulary Profile (CVP) was used for the measure of lexical sophistication. The MWE Toolkit was used to identify formulaic sequences with a threshold of 5 occurrences per million words. One way ANOVA tests and post hoc Tukey HSD tests were used for the continuous measures and effect sizes were reported as partial eta-squared (η^2p). Categorical frequency comparisons were done using Chi-square tests. The value of $p < .05$ was considered statistically significant and Bonferroni correction was used to correct multiple comparisons. Unless otherwise indicated, all analyses were performed with and without the exclusion of outliers and the results reported are from the full-corpus analysis.

4. Findings

4.1 Lexical Density and Sophistication

Table 1 shows descriptive statistics for lexical density and sophistication of AIGEN systems and benchmark corpora. As a whole, the lexical density of AIGEN texts was significantly higher than L2LB texts ($M = .61$ vs. $M = .52$, $p < .001$, $\eta^2p = .31$), but was not quite as high as the NSB texts ($M = .64$). The pattern was consistent throughout the task types, with the greatest differences for argumentative essays (AIGEN: $M = .63$; L2LB: $M = .49$) and the smallest for dialogue simulation (AIGEN: $M = .56$; L2LB: $M = .53$) in which the informal register places a limiting constraint on the number of lexical items that may be used.

Table 1. Lexical Density and Advanced Vocabulary Ratio by Corpus and Task Type

Corpus / LLM	Lexical Density (M)	TTR (M)	AVR (M)
AIGEN (GPT-4o)	.62 (.04)	.47 (.06)	.18 (.03)
AIGEN (Claude)	.61 (.05)	.49 (.07)	.17 (.04)
AIGEN (Gemini)	.60 (.04)	.46 (.06)	.16 (.03)
AIGEN (Overall)	.61 (.04)	.47 (.06)	.17 (.03)

Corpus / LLM	Lexical Density (M)	TTR (M)	AVR (M)
L2LB	.52 (.07)	.41 (.09)	.09 (.05)
NSB	.64 (.03)	.51 (.05)	.22 (.04)

Note. TTR = Type-Token Ratio; AVR = Advanced Vocabulary Ratio (C1+ words per CVP); M = Mean; values in parentheses are standard deviations. The difference was even greater in the case of the advanced vocabulary ratio. AIGEN texts used substantially more C1+ vocabulary than L2LB texts ($M = .17$ vs. $M = .09$, $p < .001$), but below NSB levels ($M = .22$). Qualitative analysis of concordance lines revealed that AIGEN vocabulary sophistication is often formulaic and not motivated by context: high frequency advanced words like CONSEQUENTLY, FURTHERMORE and PREDOMINANTLY occurred in far greater numbers in AIGEN than in either of the benchmark corpora, reflecting a stylistic register signature rather than genuine lexical diversity. The results are consistent with Etaat's (2026) identification of lexical fingerprint that marks AI from human L2 writing. Type-token ratios for AIGEN were slightly higher than L2LB, and much lower than NSB, reflecting moderate, but limited, lexical variation. We checked inter-text semantic variation across the items in the same

category as assessed by semantic diversity scores (not tabulated), and found that texts written by AIGEN were significantly more similar to each other than L2LB were ($F(1240) = 8.34$, $p < .001$), which is consistent with the general tendency of LLMs to produce more central lexical items in the distributional regions of the training texts.

4.2 Syntactic Complexity

Syntactic complexity analysis showed a unique AIGEN profile, which is different from the L2LB and NSB benchmarks in theoretically significant ways. AIGEN texts achieved longer L2LB text unit lengths than AIGEN text units in all task types (L2LTU: $M = 19.3$ vs. $M = 15.7$, $p < .001$, $\eta^2p = .28$) and were close to NSB text unit length ($M = 21.1$). The subordination index also revealed more complex embedding of depth in the AIGEN texts compared to L2LB, with a more distinctively different distribution, where embedding relative clauses and nominal clauses were preferred in AIGEN texts and subordination other than relative and nominal in NSB texts.

Table 2. Syntactic Complexity Measures by Corpus

Measure	AIGEN (M)	L2LB (M)	NSB (M)
MLTU (words)	19.3 (3.2)	15.7 (4.8)	21.1 (2.9)
Subordination Index	1.84 (.31)	1.43 (.52)	2.01 (.28)
Passive Voice Density (%)	18.4 (4.1)	11.2 (6.3)	22.7 (3.8)
Mean Clause Length	8.6 (1.4)	7.1 (2.2)	9.2 (1.3)
Coordinate Clause Freq.	.23 (.05)	.31 (.08)	.19 (.04)

Note. MLTU = Mean Length of T-Unit; frequencies are per 100 words unless otherwise noted.

The passive voice density in AIGEN texts (18.4%) was significantly higher than in L2LB

(11.2%) and was near to NSB levels (22.7%), which is presumably due to the formal academic

register of the texts when no pragmatic considerations are strong. This default-to-formal bias aligns with Godwin-Jones's (2024b) findings that LLMs generate statistically likely language that is consistent with genre prototypes that are more dominant in their training set. AI-generated writing can be an inappropriate model for L2 learners who tend to use coordinate structures excessively and nominal complexity sparsely, and may not align with the L2 syntactic development trajectory.

4.3 Discourse Cohesion

Discourse cohesion analysis was used to find textual coherence in AIGEN, L2LB and NSB texts with the use of explicit and implicit linking

devices. It was found that there were significantly more explicit cohesive devices in AIGEN texts than in L2LB and NSB texts (Table 3), which has pedagogical significance. The results show that the frequency of cohesive markers in L2LB texts is generally lower than the L2 writers' own texts, which is consistent with what has been documented for L2 writers: L2 writers' tendency to underuse the cohesive devices, reflecting their tendency to write L2 texts in a formulaic way. On the other hand, the cohesive devices found in AIGEN texts are generally higher than in the L2LB texts, suggesting the tendency of L2 writers to overuse linking devices, leading to L2 texts that are locally coherent but formulaic.

Table 3. Cohesive Device Frequency per 100 Words by Corpus

Device Type	AIGEN (M)	L2LB (M)	NSB (M)
Additive markers	3.84 (.72)	2.31 (.91)	2.67 (.58)
Adversative markers	1.92 (.48)	1.14 (.63)	1.71 (.41)
Causal markers	2.13 (.55)	1.47 (.71)	1.89 (.47)
Temporal markers	1.46 (.39)	1.22 (.58)	1.31 (.35)
Total cohesive devices	9.35 (1.52)	6.14 (2.31)	7.58 (1.24)

Note. All differences between AIGEN and L2LB, and between AIGEN and NSB, significant at $p < .01$ for total cohesive devices. Analysis of the given-new information in AIGEN texts showed that they are able to control the given-new distinction well at the sentence level and moderately at the paragraph and discourse level. Lexical chain analysis demonstrated that AIGEN texts produced shorter, less elaborate lexical chains than NSB texts, and that there was less semantic elaboration between the topically related items of the former than of the latter. This is not something obvious to readers assessing individual sentences; nevertheless, it may lead to readers seeing a surface fluency, which may be superficial and not a reflection of semantic development. This is not something that is immediately obvious when

readers are assessing individual sentences; however, it may contribute to a sense of surface fluency, which may be superficial and not a reflection of semantic development.

4.4 Pragmatic Appropriateness

Pragmatic analysis was found to be applicable to the task types of email correspondence and speech act simulation, in which the realization of register and speech act is most clearly consequential. The register consistency scores obtained for AIGEN texts, as compared with L2LB texts, were higher, consistently, across task types, AIGEN ($M = 7.8/10$) vs. L2LB ($M = 5.4/10$) ($p < .001$); however, when attention is turned to the speech act realization, a more complicated picture emerges. Pragmatic realisations in the appropriate sequences of

speech acts in AIGEN's texts were often contextually generic, but formally appropriate (e.g., salutation/request: appropriate pragmatic, but generic). Analysis of politeness markers (Table 4) showed that L2LB texts contained relatively less number of markers for positive politeness, whereas AIGEN texts contained significantly more number of markers for positive politeness, especially in service encounter dialogues, which were contextually motivated expressions of positive politeness that deviated from the level of maximal politeness.

The density of hedging devices in AIGEN academic summary texts was significantly higher

than in both the benchmark texts, which is similar to the observation of LLMs' tendency to hedge statements, regardless of the epistemic status of the target information that the L2 learner is trying to learn, that is, regardless of whether the information is factual or evaluative. AIGEN academic summary texts had significantly more hedging device density than either of the benchmark texts, which is similar to what is observed for LLM-hedging, that is, hedging claims without regard to the epistemic status of target information the learner is trying to learn, from a factual or evaluative perspective.

Table 4. Pragmatic Feature Summary (Email and Dialogue Tasks Only)

Feature	AIGEN	L2LB	NSB
Politeness markers/100w	4.23 (.81)	2.87 (.97)	3.12 (.74)
Hedging density/100w	3.91 (.67)	1.94 (.88)	2.76 (.55)
Register consistency (0-10)	7.8 (1.1)	5.4 (2.3)	8.2 (.9)
Speech act accuracy (%)	88.4 (7.2)	71.3 (13.6)	94.1 (4.8)

Note. All AIGEN-L2LB differences significant at $p < .001$; AIGEN-NSB differences significant at $p < .05$ for hedging density and politeness markers.

4.5 Phraseological Patterning

AIGEN, L2LB and NSB texts were analyzed using the following methods of phraseological analysis: N-gram frequency distributions, formulaic sequence density and collocational patterns. The results of the present research are some of the most significant for applied linguistics and second language writing teaching, because phraseological knowledge is well known as one of the most important aspects of L2 communicative competence which is very difficult to be taught explicitly and is difficult to be learned without any large amount of authentic exposure.

The formulaic sequence density of AIGEN texts was significantly higher than the density of L2LB texts and NSB texts (Table 5). In the present study, this finding is very consistent with the results reported by Alqurashi and Hamed (2026) in their corpus-based study in which they noticed systematic differences between the phraseological patterns of L2 ESL development and AI text simulation. Particularly, the distinctive sequences that are common in AIGEN but rare or absent in either of the two benchmark corpora form a unique AI phraseological profile in all three LLM systems, such as "it is worth noting that," "in conclusion, it is evident that," "this highlights the importance of," and "plays a crucial role in," which occurred 3–12 times per AIGEN corpus, while showing low or zero frequencies in the benchmark corpora.

Table 5. Phraseological Measures by Corpus

Measure	AIGEN (M)	L2LB (M)	NSB (M)
Formulaic sequence density/100w	8.91 (1.43)	5.34 (2.11)	6.22 (1.18)
MI-score (content collocations)	4.82 (.71)	3.91 (.98)	5.31 (.63)
Lexical bundle overlap with NSB (%)	41.3 (8.2)	28.7 (11.4)	100 (ref)
Top-10 N-gram corpus overlap (%)	38.2 (7.6)	51.4 (12.3)	100 (ref)

Note. MI = Mutual Information. Corpus overlap computed against NSB 4-gram inventory. Higher AIGEN formulaic density but lower overlap with NSB top-N-grams indicates AI uses frequent but non-authentic formulaic sequences. Importantly, the density of formulaic sequences is high in AIGEN, but in contrast, the lexical bundle inventory does not overlap as much as might be expected with the LBI of NSB (41.3% vs. a ceiling of almost NSB); in addition, the top-10 most frequent N-grams in AIGEN do not overlap as much as might be expected with the top 10 most frequent N-grams in NSB (41.3% vs. an upper bound approaching NSB). The result goes against what one might expect: that a language generated by an AI would be very naturalistic and spontaneous. But it is not — and it is not spontaneous in the sense of what is commonly found in NSB, but rather in a sense typical of the AI that generates it: by using phrase patterns that are common in the distribution of the data used to train the model. This has important ramifications for using AI-generated text as a phraseological model for L2 learners, and is consistent with Godwin-Jones's (2024b) view that the use of AI-generated text is a pragmatically displaced approximation of TL use.

5. Discussion

5.1 The Linguistic Profile of Generative AI Outputs: Implications for L2 Instruction

As a quantitative portrait emerges over the five linguistic dimensions explored in this study, the picture that emerges is that there is a systematic, consequential gap between the language produced by AI and the language produced by

humans. AIGEN texts are more lexically complex, syntactically complex and cohesively marked than L2 learner texts but differ from native-speaker texts in the distribution of phrases, semantic diversity and pragmatic calibration. This profile can directly inform the use of AI-generated texts in L2 instruction in terms of what it should be and should not be.

From a comprehensible input point of view, the increased lexical density and syntactic complexity of AIGEN texts may seem to be an asset: If the computer can produce texts of a certain level of complexity with high reliability, it can be used as a valuable source to train students in reading comprehension. The high frequency of discourse markers and high-frequency vocabulary items in unmarked positions in AIGEN texts, however, may develop a register awareness that is not optimally calibrated with respect to real use, if it is exposed to a large number of AI-written texts as a model. This issue is exacerbated by the phraseological discovery that frequencies of phraseological sequences in AIGEN are significantly higher than those of the phraseological sequences learners must learn for communicative purposes, the authentic competence of which is the focus of the present study.

The findings raise the need for a principled framework for the use of AI-generated text in L2 teaching, which would clearly identify tasks where the AI output is likely to be a good input for the L2 (e.g., structured grammar tasks, reading comprehension tasks with limited vocabulary) and tasks where the AI output is likely to be a misleading model of the L2 (e.g.,

academic writing, pragmatic register development, learning idiomatic expressions). The present corpus evidence offers an empirical foundation for the development of frameworks for the principled and research-informed integration of AI in language education, as advocated for by Davoodifard and Eskin (2024).

5.2 Implications for L2 Learning: Agency, Authenticity, and Acquisition

The linguistic profile presented here suggests several issues for second language acquisition theory regarding the circumstances in which AI-generated text can be used to enhance authentic L2 learning. For Godwin-Jones (2024a), the concept of distributed agency is the key to comprehending GenAI, as the active involvement of the learner in managing AI assistance is essential for its enriching development. This is corroborated by the corpus evidence: The unique linguistic "signature" of the AI text, as described by Etaat (2026) and revealed in this study, implies that uncritical uptake of such AI-generated text and treating it like another variation of the target language use is likely not to foster genuine communicative competence.

This is especially true of the phraseological results. In the theory of language acquisition, formulaic language learning is assumed to rely on the repeated exposure to formulaic sequences in real pragmatically meaningful contexts (Wray, 2002). If the formulaic sequences that most frequently occur in AI-generated text are of the type that are most characteristic of an AI distribution, instead of authentic registers that learners must access, then, counter-intuitively, the more AI-generated text that is presented as language input, the more it could interfere with the development of phraseology by overloading the learner's input with non-target-like patterns. This implies that the thorough and thorough processing of AI's product (critical evaluation) is not just desirable but necessary.

The idea of HAIT (Human–Artificial Intelligence Togetherness) proposed by Alqurashi and Hamed (2026) provides a fruitful theoretical context for the understanding of the hybrid textual context(s) in which contemporary L2

learners grow. HAIT's approach is not to view AI-generated and human-produced language as being entirely distinct but rather as two aspects of an intertwined communicative system. The corpus evidence has revealed specific characteristics to this ecology, which is not purely AI nor purely human-like, but rather has unique characteristics, and understanding these is critical in order to design L2 learning experiences which take advantage of the affordances of AI without degrading authentic L2 development.

5.3 Implications for L2 Assessment

Perhaps the most immediate implication of the present results is at the assessment level, both applied linguistics and the actual practice of language education. In principle, a corpus-linguistic analysis can detect the differences that exist systematically between the AIGEN and L2LB texts in terms of lexical density, syntactic complexity, discourse cohesion and phraseological patterning. Meanwhile, the consistency and speech act accuracy of AIGEN texts are fairly high, suggesting that the machine-generated texts can be sufficiently and moderately appropriate in superficial aspects, which may be acceptable in casual human evaluations.

This is a combination that is only detectable by systematic analysis but not by casual evaluation, therefore providing a unique challenge in terms of assessment validity. While the positive effects of automated feedback on learning outcomes are generally accepted, the validity of assessment criteria based on or calibrated with AI-generated text have not been covered by the meta-analysis by Chen et al. (2025). AI-generated texts may be used to train or calibrate automated writing evaluation systems which may then include the AI-specific linguistic features (e.g., high lexical density, formulaic cohesive marking, AI-specific phraseological patterns), systematically misrepresenting both L2 learner texts and authentic native-speaker production. Robust assessment criteria and rubrics to the characteristics of AI text are thus a pressing need for the field.

Assessing the challenge is made more difficult due to the dynamic nature of LLM capabilities. The results presented herein are based on the linguistic profile of three prominent LLM systems at the time of the study (early to mid-2026); with continued advancements in LLM technology, future systems may generate text with linguistic profiles that differ from the ones reported in this study and thus would need to be reempirically characterized. This highlights the importance of the type of long-term, longitudinal corpus-driven scrutiny of GenAI language production that applied linguistics can provide.

5.4 Theoretical Contributions

In addition to the immediate pedagogical and assessment implications, the present study has the following theoretical contributions. It first demonstrates that the linguistic profile of GenAI outputs in L2 instructional contexts is systematically different from both the output of L2 learners and native-speaker corpora, on multiple linguistic dimensions, a finding that is not explained by any single variable, but needs a multi-dimensional framework of corpus linguistics. Secondly, it offers empirical support for Godwin-Jones's (2024b) narrative on the pragmatic displacement of AI language, quantitatively analyzing the ways in which AI language differs from authentic use. Third, it illustrates the benefits of a corpus-based, NLP-based approach to the study of AI-generated texts in applied linguistics, which can be extended to other languages, types of task, and LLM systems as the field develops.

6. Limitations and Future Research

The present study has certain limitations which need to be recognized. First, the corpus, which consists of 1240 texts and some 590,000 words (both AIGEN and benchmarks), is limited to five task types in English. Further research is needed to determine the generalizability to other languages, task types (e.g., oral production, multimodal tasks) and instructional levels. The results of the meta-analyses presented in Lyu et al., 2025; Xu et al., 2025; Gu, 2025 indicate positive outcomes of AI tools, which the present corpus design does not allow capturing of other

factors, such as motivational/affective aspects of the learner-AI interaction. Second, the benchmark corpora used, which are well validated in the corpus linguistics tradition, do not fully reflect the diversity of L2 English learner groups around the world. Results may vary if learners have backgrounds other than European L1s or if the task type is other than academic writing. There is need to develop benchmark corpora with wider geographical and L1 coverage in the future. Third, it looks at three systems of LLM as they exist at a certain time period. The specific results found here may not completely describe next generation systems, as LLM capabilities have rapidly evolved, especially in the areas of instruction following, pragmatic calibration and register control. Longitudinal corpus monitoring of GenAI linguistic output should be considered a continuous research programme and not just a snapshot study. Fourth, the study is not an analysis of the reaction of the learners to AI-generated text nor is it an examination of the cognitive and acquisitional mechanisms by which learners interact with AI input. The corpus evidence could be supplemented by process oriented data generated by future research on how L2 learners actually use, interpret and learn from AI-generated language, such as through think-aloud protocols, eye-tracking or experimental acquisition paradigms. The current findings suggest some interesting paths for future research. A longitudinal study of the linguistic profile of the LLM with tracking over the generations would be very useful for the field of assessment validity research. To broaden the theoretical scope of the emerging field, cross-linguistic comparisons of output profiles of AI to languages typologically different from English would be useful. If comparisons could be made between L2 acquisition outcomes for learners who are exposed to AI-generated input and authentic input with all content carefully controlled, empirical evidence would be available on the developmental implications of the quality of the input provided by AI. Last, the development and testing of automated AI-text detection metrics based on corpus-linguistic evidence, as proposed

by Etaat (2026) in the NLP-based approach, is a practically needed research need.

7. Conclusion

The current study has employed a corpus-based, NLP-driven quantitative approach to describe the linguistic features of L2 English generative AI outputs for five types of tasks and three major LLM systems in L2 English instructional settings. The results present a unique AI linguistic profile with significant implications for L2 instruction, learning and assessment: higher lexical density, increased syntactic complexity, excessive use of explicit cohesive devices, pragmatic hedging miscalibration and a characteristic AI phraseological signature, two of which differ from those of L2 learners and two from those of native speakers. The results do not imply a recommendation against the use of generative AI in language learning. The findings of the reviewed meta-analytical studies in this paper (Ekizer, 2025; Chen et al., 2025; Lee & Lee, 2024; Xu et al., 2025; Lyu et al., 2025; Gu, 2025) indicate that AI tools have the potential to enhance positive

learning outcomes in various skills and contexts. The conceptual frameworks proposed by Godwin-Jones (2024a, 2024b) and Davoodifard and Eskin (2024) are useful to conceptualize the impact of AI on agentive and authentic aspects of L2 learning. In a more complex and nuanced way, Alqurashi and Hamed (2026) present hybrid communicative ecology and the idea of HAIT as a description of the communicative ecology that is lived by contemporary L2 learners. What the results of the present study call for, instead, is a principled and empirically based approach to the integration of AI in language education that values the linguistic characteristics of AI-generated texts as opposed to considering AI outputs as unproblematic substitutes for authentic language use. Applied linguistics, which draws on the methodological resources of corpus analysis, discourse analysis, and computational linguistics, has much to offer in this respect. In an era where generative AI is reshaping the landscape of language learning, it is not only a priceless asset but a lifeline.

References

- Alqurashi, N., & Hamed, D. M. (2026). Phraseological patterns within Human–Artificial Intelligence Togetherness (HAIT): A corpus-based comparison of ESL development and AI simulation. *Sage Open*, 16 (1), 21582440261419890.
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.
- Chen, A., Zhang, Y., Jia, J., Liang, M., Cha, Y., & Lim, C. P. (2025). A systematic review and meta-analysis of AI-enabled assessment in language learning: Design, implementation, and effectiveness. *Journal of Computer Assisted Learning*, 41 (1), e13064.
- Davoodifard, M., & Eskin, D. (2024). Exploring the role of generative AI in second language education: Insights for instruction, learning, and assessment: Introduction to the Forum. *Studies in Applied Linguistics and TESOL*, 24 (1).
- Ekizer, F. N. (2025). Exploring the impact of artificial intelligence on English language teaching: A meta-analysis. *Acta Psychologica*, 260 , 105649.
- Etaat, F. (2026). Exploring linguistic fingerprints in human and AI-generated texts: An NLP-based approach in second language writing. *Ampersand* , 100258.
- Godwin-Jones, R. (2024a). Distributed agency in second language learning and teaching through generative AI. *arXiv preprint arXiv:2403.20216*.
- Godwin-Jones, R. (2024b). Generative AI, pragmatics, and authenticity in second language learning. Virginia Commonwealth University. <https://arxiv.org/pdf/2410.14395>
- Gu, W. (2025). A meta-analysis of the efficacy of conversational AI use in EFL learning contexts. In *Social robots in education: How to effectively introduce social robots into classrooms* (pp. 335–374). Springer Nature Switzerland.
- Kartal, G., & Yeşilyurt, Y. E. (2024). A bibliometric analysis of artificial intelligence in L2 teaching and applied linguistics between 1995 and 2022. *ReCALL*, 36 (3), 359–375.

-
- Kartal, G., & Yeşilyurt, Y. E. (2025). A bibliometric analysis of artificial intelligence in L2 teaching and applied linguistics between 1995 and 2022–ADDENDUM. *ReCALL*, 37 (3), 441–441.
- Lee, H., & Lee, J. H. (2024). The effects of AI-guided individualized language learning: A meta-analysis. [Manuscript].
- Lyu, B., Lai, C., & Guo, J. (2025). Effectiveness of chatbots in improving language learning: A meta-analysis of comparative studies. *International Journal of Applied Linguistics*, 35 (2), 834–851.
- Rahman, A., Raj, A., Tomy, P., & Hameed, M. S. (2024). A comprehensive bibliometric and content analysis of artificial intelligence in language learning: Tracing between the years 2017 and 2023. *Artificial Intelligence Review*, 57 (4), 107.
- Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press.
- Xu, G., Yu, A., & Liu, L. (2025). A meta-analysis examining AI-assisted L2 learning. *International Review of Applied Linguistics in Language Teaching*. Advance online publication.