Dil Dergisi/ 2026:177(1): 61–77
DOI: 10.33690/dilder.1859378
Araştırma Makalesi/Research Article

What linguistic factors affect writing quality in L2 English writing? A multidimensional approach

İkinci dil (D2) İngilizce metinlerde yazma kalitesini etkileyen dilbilimsel faktörler nelerdir? Kapsamlı bir yaklaşım

Hakan CANGIR[a], Kutay UZUN[b], Taner CAN[c], Kurtuluş KÜLLÜ[d], Enis OĞUZ[e] Ömer Faruk KAYA[f]

[a]Ankara University, [b]Trakya University, [c]TED University, [d]University of Central Florida, [e]Middle East Technical University, [f]Bursa Technical University.

ABSTRACT

Current trends in learner corpus research indicate that integrating corpora with natural language processing (NLP) techniques can yield pedagogically more convenient results. Studies in this line of research to date have investigated writing quality in L2 English from different angles and scrutinized various linguistic features. However, only few studies have analyzed different linguistic features multi- dimensionally and no study has explored the effect of various linguistic features on L2 English writing quality in the Turkish context. Building on earlier findings and considering the niche in the Turkish context, the current study uses a specialized learner corpus and NLP tools to extract linguistic features and attempts to investigate the multi-faceted linguistic factors influencing writing quality in L2. The findings suggest that certain linguistic and textual features (e.g., word2vec, academic word frequency, phrasal frequency, and grammar errors) can play an important role in writing quality in L2 English. The mixed effects model results are discussed considering previous research and usage-based linguistic models.

Keywords: *Learner Corpus; NLP; Linguistic Features; Writing Quality; Writing in L2*

ÖZET

Öğrenci derlemi araştırmalarındaki güncel eğilimler, derlemlerin doğal dil işleme (DDİ) teknikleriyle bütünleştirilmesinin pedagojik açıdan daha elverişli sonuçlar verebileceğini göstermektedir. Bu araştırma alanındaki çalışmalar, bugüne kadar İkinci Dil (D2) İngilizce yazma kalitesini farklı açılardan incelemiş ve çeşitli dilbilimsel özellikleri titizlikle ele almıştır. Ancak, yalnızca sınırlı sayıda çalışma farklı dilbilimsel özellikleri çok boyutlu olarak analiz etmiş ve hiçbir çalışma çeşitli dilbilimsel özelliklerin Türkiye bağlamındaki D2 İngilizce yazma kalitesi üzerindeki etkisini araştırmamıştır. Önceki bulgulardan hareketle ve Türkiye bağlamındaki bu boşluğu göz önünde bulundurarak hazırlanan bu çalışma; dilbilimsel özellikleri belirlemek için özelleşmiş bir öğrenci derlemi ile DDİ araçlarını kullanmakta ve D2 yazma kalitesini etkileyen çok yönlü dilbilimsel faktörleri incelemeyi amaçlamaktadır. Bulgular; belirli dilbilimsel ve metinsel özelliklerin (örneğin; word2vec, akademik sözcük sıklığı, öbeksel sıklık ve dilbilgisi hataları) D2 İngilizce yazma kalitesinde önemli bir rol oynayabileceğini göstermektedir. Karma etkiler modeli sonuçları, önceki araştırmalar ve kullanım temelli dilbilimsel modeller ışığında tartışılmaktadır.

Anahtar sözcükler: *Öğrenci Derlemi; DDİ (Doğal Dil İşleme); Dilbilimsel Özellikler; Yazma Kalitesi; D2'de Yazma*

✉ **Sorumlu yazar/Corresponding author:** hcangir@ankara.edu.tr

Gönderildiği tarih/Received: 09.01.2026, **Kabul tarihi/Accepted:** 27.06.2026, **Yayımlanma tarihi/Publication Date:** 30.06.2026

Atıf/To cite: Cangır, H., Uzun, K., Can, T., Küllü, K., Oğuz, E., & Kaya, Ö. F. (2026). What linguistic factors affect writing quality in L2 English writing? A multidimensional approach. *Dil Dergisi*, 177(1). 61–77. 10.33690/dilder.1859378

©2026 Dil Dergisi, Ankara Üniversitesi.
Creative Commons Alıntı-GayriTicari-Türetilemez 4.0 Uluslararası (CC BY-NC-ND 4. 0) lisansı ile lisanslanmıştır.

Introduction

The emergence of learner corpus research has fundamentally shifted the landscape of second language acquisition (SLA) and language assessment. Since the creation of early reference corpora (Francis & Kučera, 1964), corpus-based methods have grown from descriptive analyses of native-speaker language to sophisticated investigations of learner production spanning a wide range of tasks, proficiency levels, and language backgrounds (Granger, 2009; Myles, 2021; Paquot & Plonsky, 2017). The increasing availability of large-scale learner corpora, such as EFCAMDAT (Geertzen et al., 2014) and the Cambridge English Corpus, has created new opportunities to identify the linguistic features that distinguish proficiency levels and to translate these findings into evidence-based teaching and assessment materials (Barker et al., 2015; Du et al., 2022; Gilquin, 2022). Crucially, the scale of modern learner

corpora makes manual analysis difficult, rendering natural language processing (NLP) tools indispensable for extracting and quantifying linguistic indices at the level of granularity that contemporary research demands.

At the applied end of this development lies automated essay scoring (AES), the use of computational algorithms to evaluate written text based on a range of linguistic properties, including lexical sophistication, syntactic complexity, and cohesion. AES systems replicate and, in some documented cases, exceed the consistency of human raters (Attali & Burstein, 2006), and are now embedded in high-stakes assessment contexts such as the TOEFL iBT, GRE, and GMAT. Beyond scoring efficiency, AES research has deepened our theoretical understanding of which linguistic features drive rater judgements of quality; a question with direct implications for language pedagogy, rubric design, and the development of automated feedback systems. User-friendly NLP tools such as TAALES (Kyle & Crossley, 2015) and TAASSC (Kyle, 2016) have made this line of research increasingly accessible, allowing researchers without extensive programming expertise to compute hundreds of indices and build robust prediction models from learner corpora.

Despite this progress, the field faces two persistent limitations. First, most NLP-based writing studies examine individual feature families (lexical, syntactic, or cohesive) in isolation, using fixed-effects regression models that treat essays as independent observations and fail to account for variance attributable to individual writers and task conditions. Second, the Turkish EFL context remains almost entirely absent from this literature: large-scale, feature-rich investigations of what predicts English writing quality among Turkish university students have yet to be conducted, leaving a significant empirical and practical gap. The present study addresses both limitations. Drawing on four validated NLP instruments (TAALES, TAACO, TAASSC, and GAMET), it simultaneously models lexical, syntactic, cohesive, and grammatical features within a mixed-effects framework, treating learner identity and task type as random effects to yield generalizable, construct-valid findings. In doing so, it offers insights that are directly applicable to the development of context-sensitive automated scoring and feedback systems in Turkish state higher education.

Theoretical Background and Literature Review

L2 Writing Quality as a Multidimensional Construct

Writing quality in a second language is not a unitary trait but a multidimensional construct shaped by the interaction of lexical, syntactic, cohesive, and rhetorical competencies (Bachman & Palmer, 1996; Norris & Ortega, 2009). Empirical research consistently identifies lexical sophistication (i.e., the use of low-frequency, contextually appropriate vocabulary) as the most robust predictor of holistic writing scores across tasks and proficiency levels (Kyle & Crossley, 2016; Zhang, 2022). Syntactic complexity, particularly fine-grained phrasal indices and usage-based measures of syntactic sophistication, also demonstrates reliable predictive power, often outperforming traditional clausal metrics (Kim & Crossley, 2018). Cohesion, the degree to which a text exhibits explicit and implicit connectivity, contributes a further predictive dimension, though its relationship to quality is not always linear. The overuse of explicit connectives can signal lower proficiency relative to the implicit cohesive strategies preferred in proficient academic prose (Crossley et al., 2016; Tywoniw & Crossley, 2019). Despite this accumulated evidence, prior studies have tended to examine these feature families in isolation using fixed-effects models that treat all essays as independent observations, obscuring the systematic variance attributable to individual writers and task types.

Usage-Based Perspectives on L2 Writing Production

Usage-based linguistics holds that language knowledge is fundamentally probabilistic, emerging from accumulated experience with form-function pairings in meaningful context (Ellis, 2002; Ellis & Wulff, 2020). Constructions encountered frequently and in reliable mappings become entrenched in long-term memory and are processed more fluently, whereas infrequent or functionally redundant structures remain fragile and prone to error. For L2 writing, this framework makes specific, testable predictions. The writers with deeper exposure to the frequency profiles and collocational norms of academic English are expected to produce texts that are more lexically sophisticated, more syntactically fluent, and more cohesively integrated (Römer, Skalicky, & Ellis, 2020). Formulaic sequences, which are recurrent multi-word units that function as holistic processing chunks, play a particularly important role, freeing cognitive resources for higher-order rhetorical decisions and marking the writer as a member of the academic discourse community (Ellis, Simpson-Vlach, & Maynard, 2008; Murakami & Ellis, 2022). These theoretical claims map directly onto the NLP indices employed in the present study: COCA-derived word frequency norms operationalize depth of genre-specific exposure; verb-argument construction indices in TAASSC capture native-like syntactic entrenchment; and phraseological measures index the internalization of conventionalized academic multi-word sequences. Usage-based theory further predicts that grammatical accuracy reflects the reliability of form-function mappings in the developing L2 system, motivating the inclusion of error indices (GAMET) as an independent predictor dimension (Ellis, 2022).

NLP and Corpus-Based Approaches to Writing Assessment

The operationalization of multidimensional writing constructs has been transformed by the development of validated automated text analysis tools. The present study draws on four instruments from Crossley and Kyle's Suite of Automatic Linguistic Analysis Tools (SALAT). TAALES (Kyle & Crossley, 2015) computes approximately 400 lexical and phrasal indices, including COCA-based word frequency, academic formula use, and word association norms, and has consistently identified lexical sophistication as a dominant predictor of writing quality. TAASSC (Kyle, 2016) measures both traditional clausal complexity and usage-based indices of syntactic

sophistication derived from native-speaker corpora, with fine-grained and VAC-based indices proving particularly informative. TAACO (Crossley et al., 2016) calculates around 150 local, global, and text-level cohesion indices, capturing connectivity through lexical overlap, semantic similarity, and referential chains. GAMET (Crossley et al., 2019) provides grammar and mechanics error counts, indexing the accuracy dimension of the construct. Together, these instruments enable a principled, theory-grounded mapping of how discrete linguistic features relate to human judgements of writing quality, moving the field beyond holistic score prediction towards genuine construct validation (Crossley, 2020; Senel, 2025).

Automated Essay Scoring and the Turkish EFL Context

Automated essay scoring (AES) has a history extending from Page's (1968) early work to contemporary transformer-based systems capable of matching trained human raters with correlations above .80 in large, standardized corpora (Lohmann et al., 2024). However, the validity of globally developed AES systems cannot be assumed in specific EFL contexts where learner populations, task formats, and rubric conventions differ substantially from development conditions (Weigle, 2013). Recent work has extended NLP-based writing assessment to the Iranian EFL context, where computational lexical and cohesion indices have been shown to align meaningfully with human rater judgements (Janebi Enayat, 2025), and to the Korean L2 context, where usage-based linguistic features have been integrated with large language model outputs to improve proficiency prediction (Hur & Ji, 2025). Research specifically addressing English writing quality among Turkish EFL learners at scale remains, to our knowledge, absent from the literature. Existing automated scoring work in Turkey has explored scoring Turkish as a second language using transformer-based models (Yavuz, 2025; Uysal & Doğan, 2021), but stops short of the fine-grained linguistic feature analysis required to understand which aspects of writing these models respond to. The present study fills this gap by providing a construct-validation account of L2 English writing quality in Turkish higher education, one with direct implications for the development of context-sensitive automated scoring and feedback systems in state institutions.

The Present Study and Research Questions

Guided by the multidimensional nature of L2 proficiency (Norris & Ortega, 2009) and the accumulated body of evidence on linguistic features that predict writing quality, the present study moves beyond score prediction to investigate construct validity through a multidimensional linguistic lens. Rather than relying on the linear models common in NLP-based research, which frequently overlook learner and task variability, it employs a mixed-effects modelling approach that treats individual differences and task types as random effects (Crossley, 2020), enabling a more precise mapping of how lexical, syntactic, and cohesive features interact to shape human judgements of writing quality. Crucially, this methodological choice is not merely technical but theoretically motivated. It repositions writing quality as a dynamic construct rather than a static output, rendering the study both confirmatory and theory-driven. While established frameworks such as Complexity, Accuracy, and Fluency (CAF; Norris & Ortega, 2009) offer valuable lenses for understanding L2 development, the present study adopts a construct-validation approach that prioritizes indices with established predictive power in automated scoring contexts, examining specifically how features calibrated to academic register collectively inform human judgements of writing quality. Drawing on TAALES, TAACO, TAASSC, and GAMET, and situating its findings within the Turkish EFL context, the study is, to our knowledge, the first large-scale investigation in Turkish higher education to examine multidimensional linguistic features, spanning lexical sophistication, syntactic complexity, coherence, and grammatical accuracy, as simultaneous predictors of writing quality, with direct implications for the development of context-sensitive automated scoring and feedback systems in Turkish state institutions.

To address the gaps identified in the foregoing review, the study is guided by the following research questions:

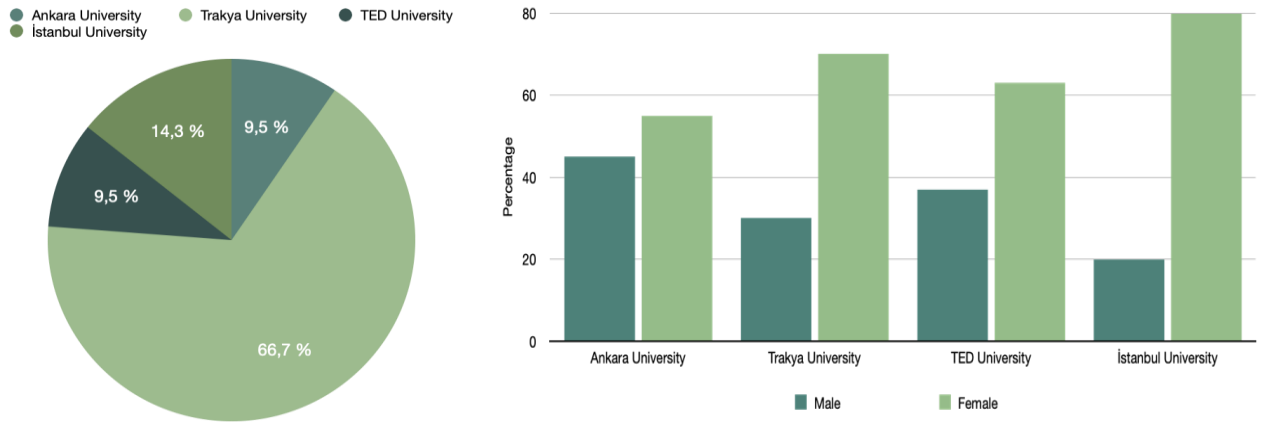
1. To what extent do multidimensional linguistic features (lexical, syntactic, cohesive, and grammatical) predict human-rated writing quality in a Turkish EFL context?
2. Which specific linguistic indices emerge as the most robust indicators of writing quality when systematic variance attributable to individual learners and task types is controlled?

Methodology

Participants

231 participants from four universities in Türkiye participated in the study. Their ages range from 19 to 35. They study in the departments of English Language Teaching and English Language and Literature (76,2%). Their B1+ proficiency level is supported by their placement in these programs via the national foreign language entrance exam, which serves as a competitive selection mechanism for language-based degrees in Türkiye. The remaining 23.8% of the sample consisted of students from intensive English preparatory programs. To maintain consistency across the group, these participants were recruited from the final and highest level of their respective programs. In the Turkish higher education system, reaching this highest level indicates that students have met the internal requirements necessary to begin departmental coursework in English. Therefore, their status in these academic tracks provides a reliable institutional basis for the assumed B1 threshold.

Figure 1. Dispersion of participants across schools and genders



Overall, the participants were around the age of 20, with few exceptions. The majority were studying at Trakya University (66,7%) and the rest of them were dispersed across the other three universities. There were more female participants than male participants in each university (Figure 1).

Each participant in our study was asked to complete the Language History Questionnaire [LHQ3] (Li et al., 2020) and LexTALE (Lemhöfer & Broersma, 2012). LHQ3 provided language background data and language scores based on the participant answers (e.g., L2 Domination and L2 Immersion) while LexTALE (*Mean* = 68,3) gave estimates of L2 language proficiency. The raw dataset consisted of average total scores of the writings, task numbers, participant numbers, language scores and language background information based on LHQ3 (Language History Questionnaire; Li et al., 2020) and LexTALE (Lemhöfer & Broersma, 2012) results collected from the participants. One problem emerged during data collection was missing data; nearly one third of our participants left the questionnaire and LexTALE incomplete. Nevertheless, since the final mixed model analyses did not include LexTALE scores or questionnaire indices, they were merely used to give overall information regarding the participants. Table 2 shows the mean and median scores for the variables *age*, *LexTale*, and *language background* (e.g., L2 proficiency) along with the missing data percentages.

Table 2. Descriptive statistics for the participants

	n=231
Age	
Mean (SD)	19.9 (2.44)
Median [Min, Max]	19 [18, 35]
Missing	55 (23.8%)
L2 Proficiency	
Mean (SD)	0.70 (0.15)
Median [Min, Max]	0.71 [0.14, 1]
Missing	56 (24.24%)
L2 Immersion	
Mean (SD)	0.47 (0.35)
Median [Min, Max]	0.50 [0.06, 2]
Missing	56 (24.24%)
L2 Dominance	
Mean (SD)	0.42 (0.11)
Median [Min, Max]	0.42 [0.07, 1.11]
Missing	56 (24.24%)
L2 to L1 Dominance Ratio	
Mean (SD)	0.87 (0.25)
Median [Min, Max]	0.85 [0.21, 1.69]
Missing	56 (24.24%)
LexTale	
Mean (SD)	68.3 (11.8)
Median [Min, Max]	67.5 [46.25, 100]
Missing	76 (32.9%)

Learner Corpus

The corpus was developed as part of a collaborative research initiative (Project No. 220K289), supported by funding from the TÜBİTAK 3501 scheme (The Scientific and Technological Research Council of Türkiye). Our learner corpus consisted of 690 texts written by 231 participants from different departments at state universities. The participants wrote one or more (up to 4) opinion essays on the given prompts below:

1. Cars should be banned from city centers to reduce traffic problems in big cities. Do you agree or disagree? To what extent do you agree? Explain your reasons using examples.
2. Universities should adopt a hybrid education model instead of online education. Do you agree or disagree? To what extent do you agree? Explain your reasons with detailed examples.
3. Empathy is considered to be one of the essential personal/social skills in the 21st century. Do you agree or disagree/To what extent do you agree? Explain your reasons using examples.
4. The Internet has caused people to be isolated from their real lives. Do you agree or disagree? To what extent do you agree? Explain your reasons using examples.

The selected prompts were sourced from a pool of standardized argumentative questions at Ankara University. These items maintain a consistent level of task complexity according to Robinson's Triadic Componential Framework (Robinson & Gilabert, 2007). By aligning resource-directing variables (such as reasoning demands) and resource-dispersing variables (such as prior knowledge and planning time), the tasks imposed a comparable cognitive load on the participants, also eliminating potential prompt effects. This structural

similarity allowed the prompts to be analyzed within a single mixed-effects model without confounding the results. Argumentative essays were chosen as they represent the most widely investigated genre in L2 writing, offering robust findings for score prediction models (e.g., Kyle & Crossley, 2016).

After completing the informed consent forms, the participants submitted their essays on an online platform without any spelling or grammar support. We intentionally preferred a controlled environment for essay collection because the state-funded language teaching institutions in Türkiye conduct writing exams in the same manner. We intended our results to inform the language teaching experts in those institutions and enlighten the assessment practices in the country. The learner corpus consists of 190,668 tokens, 7735 types and 6663 lemmas. Produced by the versatile Voyant tools (Sinclair & Rockwell, 2016), the numbers in Table 3 provide further textual details about the essays, which could tentatively indicate their textual quality.

Table 3. Textual overview of the learner corpus

Parameters	Highest	Lowest	Average
Document length	685	92	276.70
Vocabulary density*	0.670	0.256	0.449
Average words per sentence**	57.20	7.20	18.10
Readability Index***	16.735	5.201	10.517

* The ratio of the number of words to the number of unique words in the document

** 3 participants had 81 words or more for average words per sentence, and they were considered as outliers.

*** Coleman–Liau index

Tools for language background, writing evaluation and linguistic analysis

Language History Questionnaire[1] (LHQ3)

LHQ version 3 (Li et al., 2020) is a versatile tool to discern the linguistic background (e.g., L2 immersion, dominance, code-switching frequency etc.) and language proficiency of second language learners. It builds on the fundamental questions asked in earlier studies and provides a web-based interface with some sophisticated features, such as aggregated scores and visuals. We use the descriptive output of the questionnaire to provide preliminary information about the participants.

LexTALE[2]

LexTALE, which stands for Lexical Test for Advanced Learner of English, is a practical test for vocabulary knowledge commonly used by cognitive scientists, psycholinguists, and corpus linguists. It is a simple lexical decision task in which participants need to decide if the word on the screen is a valid word in English by pressing the corresponding button. Commonly used as a proxy for vocabulary knowledge and general English proficiency (e.g., Öksüz et al., 2021), LexTALE is free to administer, and it takes less than five minutes to complete (Lemhöfer & Broersma, 2012).

Writing Evaluation Rubric

To standardize the essay evaluation procedure, the raters had online calibration meetings. They did some preliminary evaluation on randomly selected texts and sought to develop collective writing evaluation criteria considering the reference rubrics in the literature (e.g., IELTS), and raters' expectations of writing quality and the Turkish context. The Delphi technique (Dalkey & Helmer, 1963) was used to design the writing evaluation rubric. In the Delphi technique, several rounds of questionnaires are conducted to gather expert opinions on a particular subject. The research team participated in three rounds of questionnaires in the evaluation rubric development phase. In each of these questionnaires, they have been asked questions about their expectations from foreign language users in terms of their language use, vocabulary sophistication, textual coherence, and such. Based on the agreed points among the raters, the writing evaluation rubric (See Appendix) has been finalized and used for essay evaluation. The writing evaluation rubric has five sections: vocabulary (2 pts.), language use (2 pts.), task completion (2 pts.), coherence-cohesion (2 pts.), and spelling-punctuation (2 pts). We believe that the cooperatively developed evaluation rubric is a (semi) authentic tool that meets the goals of our research, addresses the linguistic features under investigation, and mirrors the experiences of the expert raters as well as the earlier counterparts.

Linguistic Analysis[3]

For linguistic analysis, TAALES, TAACO, TAASSC, and GAMET were used to extract numerical data regarding lexical sophistication, textual cohesion, grammar-mechanical errors, and syntactic complexity. TAALES is a Python-based tool that can extract +400 indices of lexical sophistication, and it is used to build predictive models for second language writing and speaking proficiency. TAACO is a Python-based tool which measures 150 indices of textual cohesion, such as semantic association, word2vec, and type-token ratio indices. TAASSC, on the other hand, measures indices regarding syntactic development, such as syntactic complexity and phrasal/clausal complexity. Finally, GAMET calculates indices for structural and mechanical errors, such as grammar and spelling.

We believe lexical sophistication and semantic associations complement one another in that textual cohesion has a lot to say about lexical richness in L2 writing. As research indicates (e.g., Monteiro et al., 2023), in addition to lexical frequency, semantic context is probably the strongest force structuring the mental lexicon (Landauer & Dumais, 1997; Lund & Burgess, 1996). Therefore, indices indicating semantic association can be particularly useful to complement the assessment of L2 writing, which can help us understand how human raters evaluate lexical sophistication. Taking grammatical and mechanical errors into account is a rational decision because L2 writing is full of these types of errors, which

[1] https://lhq-blclab.org

[2] https://www.lextale.com/index.html

[3] Software can be accessed at https://www.linguisticanalysistools.org

needs to be controlled for or explored further in studies exploring writing performance in L2 (e.g., Crossley et al., 2019a). Finally, when grammar accuracy is taken into account, we believe syntactic sophistication also needs to be considered, as grammar accuracy requires language users to have control over syntactic rules. Therefore, we think grammar and syntactic errors go hand in hand with syntactic sophistication and complexity, or in other words, they are inextricably interwoven parts of a whole. The rubric categories (Vocabulary, Grammar, Task Completion, Coherence-Cohesion, and Spelling-Punctuation) were analyzed separately to identify specific areas of linguistic difficulty, as well as collectively as a composite score to assess overall writing proficiency. This approach allows for more fine-grained examination of how different linguistic features relate to writing quality. This approach is also parallel with the CAF measures (complexity, accuracy and fluency), which is commonly used to detect language proficiency. As writing is regarded as a multi-faceted construct (Kuiken & Vedder, 2019; Kyle & Crossley, 2018; Lu & Ai, 2015) it should be assessed considering various linguistic and textual features.

Score Rating Procedure

Six different raters evaluated the student writings based on the pre-set criteria. Each text was rated by two raters. For the rating phase, the scores were classified as low (0-4), low-mid (4-6), high-mid (6-8) and high (8-10) to control for rater reliability. If the first and second raters' writing performance scores fell under the same class, their scores were averaged. If there was a difference in the classes, a third rater stepped in and scored the same text. The two scores (among the three) within closer proximity were averaged. In each rating pair, there was an experienced researcher and a graduate student researcher. The raters are academics (with at least 10 years of EFL experience) and doctoral students from English Language Teaching, English Language and Literature and Linguistics departments in Türkiye.

Table 4. Intraclass Correlation

Type	Point Estimate	Lower 95% CI	Upper 95% CI
ICC3,1	0.714	0.684	0.741

Note. ICC type as referenced by Shrout & Fleiss (1979).

Following the steps summarized above, the team had an intraclass correlation of 0.74 (Cronbach alpha) in the writing evaluation phase (Table 4), which is considered a moderate reliability in Koo and Li (2016). Table 5 summarizes the dispersion of average scores across the writing tasks in the dataset.

Table 5. Dispersion average scores across writing tasks

	Task	N	Mean	SD	Minimum	Maximum
Average Score	1	183	6.02	1.51	2.50	9.50
	2	176	6.46	1.49	1.75	9.50
	3	161	6.35	1.41	2.50	10.00
	4	170	6.41	1.32	2.00	10.00

Model Selection Procedure and Data Analysis

In the first step, we used NLP tools summarized in section 3.3.4. to explore a large collection of indices indicating a wide range of linguistic features. When these comprehensive indices were added to our initial dataset, we had over 900 potential factors to explain the variety of the average essay scores. Having these many variables in a mixed effects model might not be effective in most cases. First, we might face the problem of Freedman's paradox (Freedman, 1983), which argues that variables with no real relationship to the dependent variable might seem so when the number of variables is close to the number of data points. In our case, we have more variables than our total data points. Second, having more variables does not necessarily ensure having more predictive or explanatory power. Therefore, we applied a multi-step model selection procedure to come up with an ideal model to investigate the associations between writing quality in L2 and the linguistic features under scrutiny.

Building on our spark in the previous studies in the literature, we initially wanted to focus on the linguistic features which are claimed to affect writing quality in previous research and validate earlier findings. With that in mind, using earlier studies, we filtered the linguistic features with high predictive power and addressing different aspects of writing quality in L2. The linguistic features under initial scrutiny are listed in Table 6

Table 6. Linguistic features with strong relationship with writing quality

Category	Linguistic features	Reference
Lexical sophistication	- Lexical frequency	(Crossley & McNamara, 2012)
	- Bi-gram & tri-gram frequency	(Grant & Ginther, 2000)
	- Familiarity	(Reppen, 1994)
	- Meaningfulness	(Guo, Crossley, & McNamara, 2013)
	- Specific words	(Kyle & Crossley, 2016)
	- Imageability	(Crossley, et al., 2014)
	- Concreteness	(Granger & Bestgen, 2014)

Syntactic complexity	- T-units - Clausal subordination - Passive structures - Phrasal complexity - Mean length of noun phrases	Lu (2011) Grant & Ginther (2000) Taguchi, Crawford, & Wetzel (2013) Kyle and Crossley (2018) Ortega (2003)
Textual Cohesion	- Lexical and semantic overlap across sentences - Greater diversity of words (less lexical overlap) - Word2vec all sentences - Word2vec all paragraphs - Connectives	Engber (1995) Grant & Ginther (2000) Berman & Verhoeven (2002) Jarvis (2002) Crossley & McNamara (2016)
Grammar	- Errors per 100 words - Misspelling per 100 words - Grammatical errors per 100 words	Uzun (2023) Eckes (2008) Crossley et al. (2019a)
*These linguistic features indicate the main headings of the category.		

After filtering the variables with strong predictive power from earlier research, we used these linguistic features as the potential predictor variables ($N = 90$) in our statistical analysis and used them as main effects.

We were aware of the possible overfitting problem and decided that we shouldn't have more than 70 independent variables considering the size of our dataset (690 student essays). Therefore, to avoid overfitting, we had to lower the independent variables in our dataset. We eliminated some variables which are likely to account for similar features of writing quality. For example, frequency profiles from only one reference corpus (COCA – Corpus of Contemporary American English) were used because as Durrant et al. (2019) also suggest frequency of lexis from different reference corpora like BNC (British National Corpus) and COCA shows almost the same tendencies with perfect correlations. Furthermore, because our participants were supposed to write essays in an academic style, we decided to focus only on the linguistic features indicating academic style for lexical sophistication and syntactic complexity. For example, only the collexeme and construction headings with academic features were used in the analysis. Additionally, rather than exploring the frequency profiles of the content or function words in general, we scrutinized the frequency of academic words and academic formulas. After these purposeful eliminations, we had 40 independent variables to build the statistical model.

Our model fitting procedure started with a maximal model that included overall writing scores as the outcome variable and the target linguistic features (within the framework of lexical sophistication, syntactic complexity, textual cohesion and grammar errors) as the covariates. Prior to building a MEM, possible correlations between the independent variables were analyzed to avoid multicollinearity (i.e., strong inter-associations among the independent variables). Although there are different claims regarding an acceptable variance inflation factor (VIF) and tolerance value cut-offs, we decided to keep VIF scores < 2.50 and tolerance values < 1.0 based on the suggestions by (Myers, 1990) and (Menard, 2002). The VIF and tolerance scores of the final independent variables can be seen in Table 7. After eliminating the variables with a VIF score of more than 2.5 and a tolerance value of more than 0.95, we had 28 potential variables.

Table 7. Collinearity Statistics

	Linguistic Features	VIF	Tolerance
GAMET (Grammar & mechanical errors)	Error counts per 100 words	2.48	0.364
	Grammar errors per 100 words	1.41	0.711
	Misspelling errors per 100 words	1.84	0.542
	Typographical errors per 100 words	1.14	0.880
	VAC frequency and direct object components	1.40	0.713
	Verb-VAC frequency component	1.82	0.549
	Academic collexeme ratio	1.08	0.928
TAASSC (Syntactic complexity)	Academic construction (type-token ratio)	1.85	0.541
	Academic construction (attested)	1.33	0.751
	Mean length of T-unit	1.58	0.634
	T-units per sentence	1.42	0.702
	Passive agents per clause	1.08	0.929
	Passive clausal subjects per clause	1.07	0.938
	LSA 1 all sentences (semantic similarity between all adjacent sent.)	2.33	0.430
TAACO (Textual cohesion)	Word2vec 1 all sentences (similarity score between all adjacent sent.)	1.90	0.527
	Word2vec 1 all paragraphs (similarity score between all adjacent parag.)	1.28	0.784
	Conjunctions (number of conjunctions)	1.43	0.702
	All logical connectives	1.20	0.830
	MRC Concreteness (Academic words)	1.69	0.592
	COCA academic frequency (content words)	1.96	0.511

TAALES (Lexical sophistication)	COCA academic bigram (MI score)	1.96	0.592
	COCA academic bigram (T score)	2.13	0.469
	COCA academic bigram (Delta-P)	1.82	0.548
	COCA academic trigram (MI score)	1.67	0.599
	COCA academic trigram (T score)	1.83	0.545
	COCA academic trigram (Delta-P)	1.39	0.718
	All Academic Word List (normalized)	1.31	0.766
	All Academic Formulas List (normalized)	1.42	0.706

Once we had all the potential variables ready, we built our MEM. The best fitting model was established by removing, one by one, the predictor variable that had the least impact on the model outcome until only the variables that significantly improved the fit and leading to lowest possible AIC (Akaike information criterion) were kept. It should be noted that the stepwise AIC-based model selection procedure used in this study is exploratory in nature. While AIC-based selection balances model fit and complexity, stepwise procedures are known to be sensitive to the specific dataset and may not always converge on the theoretically optimal model (Burnham & Anderson, 2002). Findings should therefore be interpreted with appropriate caution, and replication with larger and more diverse samples is encouraged.

To summarize the methodological steps we followed to narrow down the linguistic features, the initial pool of over 900 linguistic indices was reduced to a final set of predictors through a multi-stage filtering process. First, indices were selected based on their theoretical relevance to L2 writing quality and academic register. For example, following Durrant et al. (2019), we used frequency profiles exclusively from the COCA corpus, as multiple reference corpora often yield redundant, perfectly correlated data. Second, we addressed multicollinearity by scrutinizing inter-associations among the remaining 40 variables. Finally, we utilized an automated step-down selection process based on the Akaike Information Criterion (AIC). This ensured that only those indices which significantly improved the model fit and contributed unique variance to the writing quality construct were retained in the final analysis.

The final model included 11 independent variables (see Table 8 for the details), indicating different linguistic features. The covariates in our analysis were centred as suggested by Field (2016) to help combat multicollinearity between predictor variables. Participants and writing task types were used as random effects as we had multiple texts written by the same writers and four different task types which could affect the writing performance of the participants.

Table 8. Features in the final model

Covariates
Lexical sophistication
- MRC Imageability scores (academic words)
- COCA academic frequency (content words)
Syntactic complexity
- Academic collexeme ratio
- Academic constructions (TTR)
- Academic constructions (attested)
- Passive agents per clause
Textual cohesion
- Word2vec for all sentences
- All logical connectives
Grammar and mechanical errors
- Error count per 100 words
- Misspelling per 100 words
- Typographical errors per 100 words

The dataset was analyzed using a MEM instead of linear regression as opposed to earlier research (e.g., Kim et al., 2018). We claim that MEM can perform better at explaining the interrelations between the independent variables as well as their interactions with the dependent variable. The statistical analysis was computed usingjamovi project (2023), GAMLj (Gallucci, 2019), the model goodness of fit in GAMj (Gallucci, 2020).

Findings

Using the filtered linguistic features with potential to explain what makes up a quality academic writing in L2 and partly mirroring a standard writing evaluation rubric, we built a MEM to thoroughly investigate the influence of lexical sophistication, syntactic complexity, textual cohesion, and grammar-mechanical errors on L2 writing quality. Before the main analysis, all numerical variables were normalized (z-scores) using the scale function of R. This allowed a direct comparison across variables. Participant ID and Task ID were added as random effects to account for individual and task variability. Although each section in the rubric was scored separately, the final composite scores were used in the analyses, given that our primary goal was mainly related to overall writing quality. We used visual inspections to ensure the homoscedasticity and normality of the final model.

Table 9. Mixed Effects Modelling

Effect	*b*	SE	df	*t*	*p*	*Cohen's d*
(Intercept)	6.269	0.105	3.62	59.65	< .001	
- MRC Imageability (AW)	-0.331	0.063	67.848	-5.25	< .001	-1.27
- COCA academic frequency (CW)	-0.215	0.053	635.83	-4.05	< .001	-0.32
- AW collexeme ratio	0.078	0.043	584.88	1.80	0.074	0.15
- AW constructions (TTR)	-0.169	0.046	616.22	-3.66	< .001	-0.29
- AW constructions (attested)	-0.089	0.047	593.90	-1.90	0.058	-0.16
- Passive agents per clause	0.086	0.043	583.12	2.00	0.046	0.17
- All logical connectives	-0.105	0.049	592.43	-2.13	0.034	-0.18
- Word2vec (all sentences)	0.132	0.048	678.90	2.77	0.006	0.21
- Misspel. per 100 words	-0.212	0.048	666.46	-4.38	< .001	-0.34
- Grammar per 100 words	-0.244	0.046	648.74	-5.33	< .001	-0.42
- Typ. errors per 100 words	-0.157	0.047	663.48	-3.33	0.001	-0.26

Random effects	Variance	SD
Participant	0.64	0.80
Task	0.03	0.19
Goodness of fit:	R^2 conditional .540 | R^2 marginal .215	

As is seen in Table 9, we found some positive and negative significant associations between the indices under scrutiny and writing quality in L2, which explained around 21.5% of the variance in our data. When both fixed and random effects (i.e., participants and tasks) were considered, our data explained 54% of the variance. This difference reveals the potential influence of individual differences and the question type on the writing performance, which needs to be controlled in studies exploring writing quality in L2 (Crossley, 2020). The numbers highlighting the effect of lexical sophistication indicate a certain negative influence of lexical sophistication on the participants' writing performance. To be more precise, the student essays with more concrete words ($b = -0.331$, $SE = 0.063$, $t(67.85) = -5.25$, $p < .001$, $d = -1.27$) and with more frequent content words ($b = -0.215$, $SE = 0.053$, $t(635.83) = -4.05$, $p < .001$, $d = -0.32$) are associated with lower writing scores. In other words, more quality texts had more abstract words and less frequent content words covered in the academic part of COCA. In terms of syntactic complexity, there are three significant associations. Though indicating an insignificant association, texts with more academic collostructions (i.e., lexical company kept by a construction; [$b = 0.078$, $SE = 0.043$, $t(584.88) = 1.80$, $p = .074$, $d = 0.15$) were rated high, whereas texts with higher academic construction type-token ratio ($b = -0.169$, $SE = 0.046$, $t(616.22) = -3.66$, $p < .001$, $d = -0.29$) and covering more overlapping constructions with a reference corpus ($b = -0.089$, $SE = 0.047$, $t(593.90) = -1.90$, $p = .058$, $d = -0.16$) were scored lower, though indicating a statistically insignificant significance. Passive structures per clause contribute to the overall model and tentatively influence writing quality positively ($b = 0.086$, $SE = 0.043$, $t(583.12) = 2.00$, $p = .046$, $d = 0.17$). In addition, textual cohesion seemed to partly affect L2 writing quality. There is a positive association between higher word2vec (indicating semantic overlap between sentences; ($b = 0.132$, $SE = 0.048$, $t(678.90) = 2.77$, $p = .006$, $d = 0.21$) and higher writing scores. On the other hand, the number of logical connectives and L2 writing performance are negatively associated ($b = -0.105$, $SE = 0.049$, $t(592.43) = -2.13$, $p = .034$, $d = -0.18$). Another noteworthy finding was the strong negative association between grammar-mechanical errors and writing performance. Overall, grammar errors per 100 words ($b = -0.244$, $SE = 0.046$, $t(648.74) = -5.33$, $p < .001$, $d = -0.42$), misspelling per 100 words ($b = -0.212$, $SE = 0.048$, $t(666.46) = -4.38$, $p < .001$, $d = -0.34$), typographical errors per 100 words ($b = -0.157$, $SE = 0.047$, $t(663.48) = -3.33$, $p = .001$, $d = -0.26$) are negatively associated with writing scores. That is, participants with more grammatical-mechanical errors scored lower in the writing tasks.

Section 5 will discuss the detected negative and positive associations in more details and scrutinize the complex and interwoven factors affecting writing quality in L2 English by addressing the related literature and highlighting the differences and similarities with the current research.

Discussion

Lexical Sophistication and Writing Quality

One of the most conclusive findings in the literature is regarding the negative relationship between lexical frequency and writing quality. Crossley & McNamara (2012) detect lexical frequency as an important predictor of writing quality, which is in line with our finding. A further contribution of our study could be its focus on academic words only as we conclude that the use of more frequent content words available in the academic part of COCA is negatively associated with writing quality. In other words, the results indicate a tendency for texts with lower-frequency lexical items (i.e., more sophisticated or less familiar words) to be rated as higher quality. This pattern is consistent with usage-based accounts of L2 lexical development, which predict that writers with deeper exposure to academic register norms gradually internalise the lower-frequency, genre-specific vocabulary

that characterises expert academic prose — and that this internalisation is directly reflected in the sophistication of their written output (Ellis, 2002; Ellis & Wulff, 2020).

As for the effect of multi word units, though there are studies indicating a moderate effect of collocational frequency and association strength on writing performance in L2, we couldn't detect an association between lexical association strength measure (MI score in particular) and writing scores. A positive effect can be seen in earlier studies and deemed a rather conclusive finding. Bestgen and Granger (2014) find that MI scores based on occurrences in COCA correlate positively with overall writing quality. Bestgen (2017) observes a similar positive correlation between mean MI scores and the quality of writing in FCE (First Certificate in English) tasks. Likewise, Kim, et al. (2018) detect a parallel trend suggesting that there is a significant positive correlation between MI scores and writing proficiency scores. Garner et al. (2018) with a slightly different approach in their analysis conclude that as proficiency levels increase, L2 writers produce word combinations with higher MI scores (i.e., more unique word combinations, component parts of which almost always appear in the company of one another). The fact that we couldn't detect the effect of association strength measures (e.g., MI or Delta-P) on writing quality could be attributed to our participants' proficiency levels (B1 on average) in L2 English writing. We may speculate that more proficient writers like those participated in earlier research are likely to produce more unique word combinations or multi word units with higher MI scores, which in turn could influence the writing quality positively. Another tentative explanation could be the potential use of erroneous multi-word units. As our participants are not advanced L2 users, their attempts to produce unique and natural sounding word combinations might result in some errors, which could be associated with poor quality in writing. Finally, because we employed timed writing in data collection, our participants might have produced lexically less complex text. Therefore, an untimed writing task could trigger texts with different lexical features (Crossley, 2020) and is likely to indicate different results.

Another important aspect of lexical sophistication is the use of concrete and abstract lexical items in L2 writing. In our finding, lower imageability of academic words is associated with higher writing scores. In other words, the texts with more abstract academic words tend to be rated as higher quality. There are some conflicting results in the literature in this regard. For instance, Qin and Uccelli (2016) find significant positive correlation between the use of abstract words and the quality for argumentative texts. Guo, et al. (2013) detect a similar pattern indicating a positive correlation between the use of abstract words and quality in independent essays. Crossley and McNamara (2012), on the other hand, report a significant negative correlation between abstractness scores and essay grades. Our unique contribution could be the scrutiny particularly on the effect of imageability of academic words, which is lacking in earlier research with a more inclusive approach. Given the controversial findings, though; further research is needed for more conclusive results.

Textual Cohesion and Writing Quality

Text cohesion is an important element of academic writing that explains interrelations within and between sentences because it has potential to show lexical, semantic, and argumentative inter-connectivity in a text (Halliday & Hassan, 1976). We found two significant predictors of L2 writing quality: word2vec scores indicating semantic similarity across sentences and the use of logical connectives (e.g., actually, admittedly etc.).

Cohesion can be detected at the sentence level (i.e., local cohesion) and beyond sentence level, such as paragraph (i.e., global cohesion). Text cohesion has been widely used in assessing writing in L1. Studies have found that local cohesion markers indicate better writing quality in young L1 writers' texts (Struthers, Lapadat, & MacMillan, 2013). There are rather mixed findings for older writers. It was found that there are conflicting correlations between text quality and referential cohesion. Perin and Lauterbach (2016) claim they are negatively correlated, whereas MacArthur et al., (2019) assert the opposite. Although relatively rare, studies investigating the relationship between L2 writing quality and textual cohesion follow a similar trend as L1 writing. Overall, research investigating L2 writing to date has suggested that local cohesion and writing quality are not strongly associated (e.g., Crossley & McNamara, 2010, 2011), which is not aligned with our findings. There are a few studies suggesting a relationship between local cohesion and writing quality in L2 and our findings indicate similar trends. Both the indices contributing to our predictive model represent local cohesion features. Higher word2vec (at the sentence level) is associated with higher writing quality, reflecting more sentence level semantic overlap and more logical connectives result in less quality writing, similar to what Tywoniw and Crossley (2019) claim. Jarvis (2002) suggests that more proficient L2 users are inclined to write texts less cohesively, based on lexical and semantic overlap across sentences. In another study, Crossley and McNamara (2012) conclude that L2 texts with stronger performance contain fewer connectives, which is partly in line with our finding stating that more logical connectives indicate weaker writing performance.

On the other hand, research shows that there is a positive correlation between global cohesion and writing quality (e.g., McNamara et al., 2013). Exploring the effect of global cohesion, Crossley and McNamara (2016) contend that adjacent overlap across paragraphs correlates positively with text quality. We could not find the same positive relationship between a global cohesion aspect and writing scores, which could partly be explained by the language level of the participants in the current study (B1 on average). We can speculate that a countereffect might have been observed if the participants had been more advanced L2 users. Research exploring the development of L2 users in terms of text cohesion also shows that as proficiency increases, a movement from mastery in local cohesion to global cohesion can be observed (Crossley, 2020). In another study with a developmental perspective,

Crossley et al., (2016) state that there is a strong increase in noun overlap between paragraphs as proficiency increases, signaling an increasing strength in global cohesion across essays. An alternative explanation for the lack of a global cohesion effect could be the writing task type used in this study. For example, Abdi Tabari and Johnson (2023) conclude that global level cohesion appears to affect writing quality in L2 when participants are producing argumentative essays, whereas local cohesion markers seem to stand out as important factors contributing to narrative writing quality. Therefore, a further study with more proficient L2 English users (or users from various proficiency levels) and analyzing an alternative writing type could reveal a different picture.

Word2vec is the most significant positive association detected in our study, which is also the case for earlier studies (e.g., Crossley et al., 2019). The essays with more semantic overlap between sentences received higher scores by the raters. From a usage-based perspective, this finding is theoretically coherent: writers who have accumulated extensive exposure to academic discourse develop a stronger sensitivity to the semantic fields and collocational networks that hold academic arguments together, producing texts whose sentences are more topically and conceptually connected (Ellis, 2002; Römer, Skalicky, & Ellis, 2020). Word2vec (Mikolov et al., 2013) builds on a neural network model to represent words and phrases in a vector-space model. Each lexical item's embedding is computed through its surrounding context within the training dataset. "Words that co-occur in similar contexts are represented in each other's near vicinity, whereas those from unrelated contexts are represented distantly in different parts of the vector space" (Crossley et al., 2019, p. 17). Scrutinizing the effect of textual cohesion on writing quality, Crossley et al. (2016) state that L2 writers show increased cohesion in their writing over the course of a semester, which means as they gain proficiency, they produce higher-quality texts with stronger coherence. Crossley and McNamara (2011) report modest correlations between semantic similarity indices measured between paragraphs using LSA (average latent semantic analysis) and human ratings of text quality. Similarly, Crossley et al. (2019b) argue that semantic overlap at the paragraph level, as measured by LSA, word2vec, and LDA (average latent Dirichlet allocation divergence score), shows strong correlations with ratings of text coherence. We could not find a significant effect of semantic similarity across paragraphs (i.e., word2vec for all paragraphs), which could be due to the level of the essays written by our participants or due to the task types. Essays by more advanced L2 writers or that require a different organization stemming from the task (e.g., an integrated writing task) could show more semantic overlap across paragraphs and thus trigger a stronger effect on the writing performance.

Syntactic Complexity, Grammar Errors and Writing Quality

Several studies (e.g., Eckes, 2008) have attempted to explain the importance of grammatical and structural errors, and they contend that grammar plays a significant role in producing more advanced texts in L2. In our study, we find negative significant associations between GAMET indices (errors per 100 words, misspelling per 100 words, grammatical errors per 100 words and writing performance in L2) as in Uzun (2023) and Crossley et al. (2019a).

With respect to syntactic complexity, two statistically significant associations were identified. However, no relationship was observed between measures of association strength (e.g., MI scores) and writing quality; however, we found a positive effect of collexeme (collostruction) ratio on students' essay scores. Collexeme is a verb-construction contingency feature that could indicate an alternative association strength measure to MI scores, or Delta-P. Collexeme analysis (Stefanowitsch & Gries, 2003) measures the strength of verb-construction combinations rather than merely lexical combinations. Gries, Hampe, & Schönefeld (2005) similarly conclude that collostructions are strongly correlated with L2 production. To be more precise, the more academic collostructions the students use, the higher they score in writing. Other studies, like Taguchi, Crawford, & Wetzel (2013) also claim that increased writing proficiency is linked to greater phrasal complexity in written production. Kyle and Crossley (2017), in their study of the usage-based indices, also contend that essays with higher scores tend to contain strongly associated verb-construction combinations, which results in more native-like written production. Likewise, Hwang and Kim (2023) support the earlier remarks, stating that essays with higher scores are comprised of constructions with low frequency. These findings may imply that usage-based indices explored in earlier studies have the potential to enlighten syntactic complexity in L2 writing and usage-based learning of constructions also applies to L2 production (Ellis & Wulff, 2020).

As opposed to the positive collexeme ratio effect, we observe a negative influence of academic constructions on writing scores. Attested item indices (AW constructions in our case) comprise a rough measure of frequency (Kyle, 2016). We deduce that the frequency of the constructions extracted from the academic part of COCA affects writing quality negatively. Diversity and frequency are found to have a negative effect on writing scores, which means lower-scoring texts have more frequent constructions. Previous research also suggests that Verb-VACs (verb argument constructions) with higher frequency indicate lower scores in writing. This finding also partly correlates with earlier research (e.g., Bestgen, 2017) stating that more collocations (particularly the ones with higher MI scores) and a wider range of collostructions in writing (which can be considered formulaic language use) may indicate better production in L2 (Garner, 2022). From a broad perspective, we can claim that the observed associations in the MEM go hand in hand with usage-based models of language use (Tomasello, 2003). According to usage-based models, frequency of occurrence and associative learning (i.e., building lexical connections) relate to lexical acquisition

at the word and phrase level, which are likely to influence proficiency in L2 (Goldberg, 2006).

On the other hand, we find a negative influence of the type-token ratio of academic constructions on writing quality. Type-token ratio in this sense is the proportion of unique constructions to the total number of constructions, which indicates a greater diversity in terms of syntactic structures. This finding appears to contradict earlier findings exploring phrasal complexity (e.g., Kyle & Crossley, 2018). We may partly attribute this contradictory finding to our mere focus on academic constructions. A further analysis of constructions across the whole corpus could yield different findings. An alternative explanation could be that, as Wang (2014) also states, participants using a greater diversity of constructions may be prone to make more mistakes in their written production, which could correlate with weaker writing quality. Conversely, participants employing a smaller repertoire of constructions (particularly academic ones) might use them with greater control and produce texts free from mistakes, which results in better performance in writing. Therefore, the negative association between constructional diversity (as measured by type-token ratio) and writing quality can be theoretically explained through the lens of usage-based Construction Grammar and SLA research on formulaic language. From a usage-based perspective, language knowledge is rooted in frequency and entrenchment. That is, constructions that are encountered and rehearsed more frequently become more deeply entrenched in the learner's mental grammar (Bybee, 2010; Goldberg, 2006), making them more accessible and reliably deployable under the cognitive pressure of timed or high-stakes writing tasks. Higher-scoring writers may therefore strategically converge on a smaller set of well-entrenched, academically appropriate constructions rather than experimenting with novel or peripheral ones. This aligns with research on formulaic language in L2 writing, which consistently shows that proficient learners exploit recurrent multi-word sequences and phraseological patterns to achieve fluency and accuracy (Wray, 2002). Reduced constructional variety, in this view, is not a marker of impoverished repertoire but rather of selective, strategic deployment of high-frequency, target-like constructions, which can be considered a behavior that minimizes error risk and maximizes communicative precision.

Building on the claims by Crossley (2020) and in light of the findings of the current research, we believe investigations into writing quality should focus on multiple linguistic constructs and how they interact, in that writing quality cannot be assessed through isolated linguistic features. Studies in the literature also provide evidence for this interconnected effect, particularly highlighting the significant influence of lexical sophistication and textual cohesion on writing quality (Guo et al., 2013). Jarvis et al. (2003) also state that linguistic features of different sorts can interact with each other in L2 writing, and essays with higher scores should not be defined by a fixed set of linguistic features.

We hope attempts like this study, albeit its limitations, can help enlighten the intricate concept of L2 writing quality and illuminate how linguistic elements interact and contribute to better writing performance. Reliable MEMs investigating all possible features of quality in L2 writing can be the steppingstones for more trustworthy automated essay scoring (AES). From a pedagogical perspective, the development of AES software may have an important effect on the teaching and learning of L2 writing skills in that it can facilitate learning and shape teaching and testing practices.

Conclusion

In this research study, we provided an overarching account of the relationship between certain linguistic features and writing quality in L2 English. In our analysis, we attempted to consider all the related aspects of writing performance (in parallel with standard writing evaluation criteria) and build a MEM scrutinizing lexical sophistication, syntactic complexity, textual cohesion and grammar accuracy. Using the model, we sought to explain specific associations between the lexical features under investigation and the mean writing scores in order to foreground the potential factors contributing to more advanced L2 texts. For a more comprehensive explanation of the phenomenon of performance in second language writing, more developmental studies should be conducted by considering proficiency levels and task types. Pedagogically, these findings suggest that instruction should prioritize the mastery of high-frequency academic lexical bundles and formulaic constructions rather than the broad constructional diversity, which may be associated with increased error rates. Furthermore, writing curricula should emphasize the development of local semantic overlap and the strategic use of abstract academic vocabulary, as these features are more immediate indicators of writing quality at the intermediate level than global cohesion markers. Finally, the integration of NLP tools into the classroom could provide students with automated, linguistically-grounded feedback, allowing them to monitor their own progress in their written production in real-time. Future studies should also consider working with more balanced data than ours, with essay scores evenly distributed across the dataset. Essay scores with a normal score distribution could reveal more reliable output, signifying other potential linguistic features and providing a better insight into the contributing factors to writing quality in L2. Another important point is that we need more readily available and user-friendly NLP tools with open-source codes for academic and educational purposes. They can be used for a thorough automatic linguistic analysis and thus pave the way for automated grading and feedback in L2 writing. Last but not least, while the inter-rater reliability in this study was within the acceptable range, the observed variance in human judgment may have slightly decreased the correlation coefficients between specific linguistic features and writing scores. Future studies might employ more extensive rater training or a consensus-based scoring

approach to further minimize measurement error and capture even more nuanced linguistic effects.

Financial Support

This publication was supported by project number 220K289 under the TÜBİTAK 3501 Career Development Program.

Ethics Statement

This study was found to comply with ethical guidelines in accordance with Decision No. 3/9, adopted at the meeting of the Ankara University Ethics Committee on March 29, 2021.

Conflict of Interest Statement

The authors of this article declare that there is no conflict of interest.

Statement on the Use of Generative AI

Generative AI tools were not used to create content during the preparation of this study; these tools were used only during the language editing and proofreading stages.

Author Contributions

The authors declare that they contributed equally to the planning, data collection, and writing of this article.

Data Availability Statement

Data supporting the findings of this article are available from the corresponding author upon reasonable request.

References

Abdi Tabari, M., & Johnson, M. D. (2023). Exploring new insights into the role of cohesive devices in written academic genres. *Assessing Writing*, *57*, 100749. https://doi.org/https://doi.org/10.1016/j.asw.2023.100749

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment*, *4*(3). Available from http://www.jtla.org

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Barker, F., Salamoura, A. & Saville, N. (2015). "Learner corpora and language testing". In S. Granger, G. Gilquin & F. Meunier, Eds., *The Cambridge Handbook of Learner Corpus Research*, Cambridge Handbooks in Language and Linguistics. Cambridge University Press, 511–534. https://doi.org/10.1017/CBO9781139649414.023

Berman, R. and Verhoeven, L. (2002). Cross-linguistic perspectives on the development of text-production abilities: Speech and writing. *Written Language and Literacy*, *5*(1), 1-43. https://doi.org/10.1075/wll.5.1

Bestgen, Y. (2017). Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System*, *69*, 65–78. https://doi.org/10.1016/j.system.2017.08.004

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). Springer.

Bybee, J. (2010). *Language, Usage and Cognition*. Cambridge University Press.

Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, *11*(3), 415–443. https://doi.org/10.17239/JOWR-2020.11.03.01

Crossley, S. A., Bradfield, F., & Bustamante, A. (2019a). Using human judgments to examine the validity of automated grammar, syntax, and mechanical errors in writing. *Journal of Writing Research, 11*(2), 251-270. https://doi.org/10.17239/jowr-2019.11.02.01

Crossley, S. A., Kyle, K., & Dascalu, M. (2019b). The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods, 51*(1), 14-27. https://doi.org/10.3758/s13428-018-1142-4

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing, 32*, 1-16. https://doi.org/10.1016/j.jslw.2016.01.003

Crossley, S. A., Kyle, K., Varner, L., Gou, L., & McNamara, D. S. (2014). Linguistic microfeatures to predict L2 writing proficiency: A case study in automated writing evaluation. *Journal of Writing Assessment*, *7*(1).

Crossley, S. A., & McNamara, D. S. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 984-989). Austin, TX: Cognitive Science Society.

Crossley, S. A., & McNamara, D. S. (2011). Text coherence and judgments of essay quality: Models of quality and coherence. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society*. (pp. 1236-1241). Austin, TX: Cognitive Science Society.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35*(2), 115-135. https://doi.org/10.1111/j.1467-9817.2010.01449.x

Crossley, S. A., & McNamara, D. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology*, *17*(2), 171–192. https://doi.org/10.64152/10125/44329

Crossley, S. A., & McNamara, D. S. (2016). Say more and be more coherent: How text elaboration and cohesion can increase writing quality. *Journal of Writing Research*, *7*(3), 351–370. https://doi.org/10.17239/jowr-2016.07.03.02

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, *29*(2), 243–263. https://doi. org/10.1177/0265532211419331

Dalkey, N., & Helmer, O. (1963.) An Experimental Application of the Delphi Method to the Use of Experts. *Management Science*, *9*, 458-467. https://doi.org/10.1287/mnsc.9.3.458

Du, X., Afzaal, M., & Al Fadda, H. (2022). Collocation use in EFL learners' writing across multiple language proficiencies: A corpus-driven study. *Frontiers in Psychology*, *13*, 752134. https://doi.org/10.3389/fpsyg.2022.752134

Durrant, P., Moxley, J., & McCallum, L. (2019). Vocabulary sophistication in First-Year Composition assignments.

International Journal of Corpus Linguistics, *24*(1), 33–66. https://doi.org/10.1075/ijcl.17052.dur

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155–185. https://doi.org/10.1177/0265532207086780

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188. https://doi.org/10.1017/S0272263102002024

Ellis, N. C. (2022). *The usage-based approach to second language acquisition*. Routledge.

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, *42*(3), 375–396. https://doi.org/10.1002/j.1545-7249.2008.tb00137.x

Ellis, N. C., & Wulff, S. (2020). *Usage-based approaches to SLA*. For: Bill VanPatten, Gregory Keating and Stefanie Wulff (eds.), Theories in second language acquisition. 3rd edition. New York: Routledge, 63-82.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*(2), 139-155. https://doi.org/10.1016/1060-3743(95)90004-7

Field, A. (2016). *An adventure in statistics: The reality enigma*. Sage Publications.

Francis, W. N., & H. Kučera. (1964). A *Standard Corpus of Present-Day Edited American English, for use with Digital Computers*. Providence, Rhode Island: Department of Linguistics, Brown University. Revised 1971. Revised and amplified 1979.

Freedman, D. A. (1983). A note on screening regression equations. *Am. Stat.* , *37*(2), 152-155. https://doi.org/10.1080/00031305.1983.10482729

Gallucci, M. (2019). *GAMLj: General analyses for linear models*. [jamovi module]. Retrieved from https://gamlj.github.io/.

Gallucci, M. (2020). *Model goodness of fit in GAMLj*. link.

Garner, J. (2022). The cross-sectional development of verb–noun collocations as constructions in L2 writing. *International Review of Applied Linguistics in Language Teaching*, *60*(3), 909-935. https://doi.org/10.1515/iral-2019-0169

Garner, J., Crossley, S. A., & Kyle, K. (2018). Beginning and intermediate L2 writer collocational use: A phraseological approach. *ITL - International Journal of Applied Linguistics*, *169*(1), 22–44. https://doi.org/10.1075/itl.17010.gar

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2014). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). *In Proceedings of the 31st second language research forum (SLRF)*. Cascadilla Press

Gilquin, G. (2022). Written learner corpora to inform teaching. In R. R. Jablonkai, &

E. Csomay (Eds.), *Routledge handbook of corpora in English Language Teaching and Learning* (pp. 281–295). Routledge. https://doi.org/10.4324/9781003002901-23

Goldberg, A. (2006). *Constructions at Work: The Nature of Generalization in Language*. New York: Oxford University Press. *Cognitive Linguistics Bibliography (CogBib)*.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 13-32). John Benjamins. https://doi.org/10.1075/scl.33.04gra

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics*, *52*, 229–252. https://doi.org/10.1515/iral-2014-0011

Grant, L., & Ginther, A. (2000). Using Computer-Tagged Linguistic Features to Describe L2 Writing Differences. *Journal of Second Language Writing*, *9(*2), 123–145. https://doi.org/https://doi.org/10.1016/S1060-3743(00)00019-9

Gries, S., Hampe, B., & Schönefeld, D. (2005). Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics*, *16*(4), 635-676. https://doi.org/10.1515/cogl.2005.16.4.635

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, *18*(3), 218-238. https://doi.org/10.1016/j.asw.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.

Hur, W., & Ji, B. (2025). A Hybrid System for Automated Assessment of Korean L2 Writing: Integrating Linguistic Features with LLM. *Systems*, *13*(10), 851. https://doi.org/10.3390/systems13100851

Hwang, H., & Kim, H. (2023). Automatic analysis of constructional diversity as a predictor of EFL students' writing proficiency. *Applied Linguistics*, *44*(1), 127–147. https://doi.org/10.1093/applin/amac046

Janebi Enayat, M. (2025). Computationally derived linguistic features of L2 narrative essays and their relations to human-judged writing quality. *Language Testing in Asia*, *15*(1). https://doi.org/10.1186/s40468-025-00374-9

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19*(1), 57-84. https://doi.org/10.1191/0265532202lt220oa

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing*, *12*(4), 377-403. https://doi.org/10.1016/j.jslw.2003.09.001

Jiang, J., Yu, L., & Lin, C. (2021). Lexical sophistication and syntactic complexity in EFL writing quality: A multidimensional study. *International Review of Applied Linguistics in Language Teaching*, *59*(4), 1023–1051

Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal, 102*(1), 120-141. https://doi.org/10.1111/modl.12447

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155-163. https://doi.org/10.1016/j.jcm.2017.10.001

Kuiken, F., & Vedder, I. (2019). Syntactic complexity across proficiency and languages: L2 and L1 writing in Dutch, Italian and Spanish. *International Journal of Applied*

Linguistics, *29*(2), 192–210. https://doi.org/https://doi.org/10.1111/ijal.12256

Kyle, K. (2016). Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication [Doctoral dissertation, Georgia State University]. ScholarWorks @ Georgia State University. http://scholarworks.gsu.edu/alesl_diss/35

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly* *49*(4), 757-786. https://doi.org/10.1002/tesq.194

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing, 34*, 12-24. https://doi.org/10.1016/j.jslw.2016.10.003

Kyle, K., & Crossley, S. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, *34*(4), 513-535. https://doi.org/10.1177/0265532217712554

Kyle, K., & Crossley, S. A. (2018). Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices. *The Modern Language Journal*, *102*(2), 333–349. https://doi.org/https://doi.org/10.1111/modl.12468

Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods 50*(3), pp. 1030-1046. https://doi.org/10.3758/s13428-017-0924-4

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*, 325-343. https://doi.org/10.3758/s13428-011-0146-0

Li, P., Zhang, F., Yu, A., & Zhao, X. (2020). Language History Questionnaire (LHQ3): An enhanced tool for assessing multilingual experience. *Bilingualism: Language and Cognition, 23*(5), 938-944. https://doi.org/10.1017/S1366728918001153

Lohmann, J. F., Junge, F., Möller, J., Fleckenstein, J., Trüb, R., Keller, S., Jansen, T., & Horbach, A. (2024). Neural networks or linguistic features? Comparing different machine-learning approaches for automated assessment of text quality traits among L1- and L2-learners' argumentative essays. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-024-00426-w

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, *45*(1), 36-62. https://doi.org/ 10.5054/tq.2011.240859

Lu, X., & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, *29*, 16–27. https://doi.org/https://doi.org/10.1016/j.jslw.2015.06.003

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208. https://doi.org/10.3758/BF03204766

MacArthur, C. A., Jennings, A., & Philippakos, Z. A. (2019). Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction?. *Reading and Writing*, *32*(6), 1553-1574. https://doi.org/10.1007/s11145-018-9853-6

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural Language Processing in an Intelligent Writing Strategy Tutoring System. *Behavior Research Methods*, *45*(2), 499-515. https://doi.org/10.3758/s13428-012-0258-1

Menard, S. (2002). *Applied logistic regression analysis*. SAGE Publications, Inc., https://doi.org/10.4135/9781412983433

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *International Conference on Learning Representations*.

Monteiro, K., Crossley, S., Botarleanu, R. M., & Dascălu, M. (2023). L2 and L1 semantic context indices as automated measures of lexical sophistication. *Language Testing*, *40*(3), 576-606. https://doi.org/10.1177/02655322221147924

Murakami, A., & Ellis, N. C. (2022). Effects of availability, contingency, and formulaicity on the accuracy of English grammatical morphemes in second language writing. *Language Learning*, *72*(4), 899–940. https://doi.org/10.1111/lang.12493

Myers, R.H. (1990) *Classical and modern regression with applications.* PWS-Kent Publishing, Boston.

Myles, F. (2021). Commentary: An SLA Perspective on Learner Corpus Research. In B. Le Bruyn & M. Paquot (eds.), *Learner Corpus Research Meets Second Language Acquisition* (pp. 258-273). Cambridge University Press. https://doi.org/10.1017/9781108674577.013

Norris, J. M., & Ortega, L. (2009). Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity. *Applied Linguistics*, *30*(4), 555–578. https://doi.org/10.1093/applin/amp044

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied linguistics*, *24*(4), 492-518. https://doi.org/10.1093/applin/24.4.492

Öksüz, D., Brezina, V., & Rebuschat, P. (2021). Collocational processing in L1 and L2: The effects of word Frequency, collocational frequency, and association. *Language Learning*, *71*(1), 55–98. https://doi.org/10.1111/lang.12427

Page, E. B. (1968). The use of the computer in analyzing student essays. *International Review of Education*, 14(2), 210–225. https://doi.org/10.1007/BF01257528

Paquot, M., & Plonsky, L. (2017). Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research*, *3(*1), 61–94. https://doi.org/10.1075/ijlcr.3.1.03paq

Perin, D., & Lauterbach, M. (2018). Assessing Text-Based Writing of Low-Skilled College Students. *International Journal of Artificial Intelligence in Education*, *28*(1), 56–78. https://doi.org/10.1007/s40593-016-0122-z

Qin, W., & Uccelli, P. (2016). Same language, different functions: A cross-genre analysis of Chinese EFL learners' writing performance. *Journal of Second Language Writing*, *33,* 3–17. https://doi.org/10.1016/j.jslw.2016.06.001

Reppen, R. (1994). *Variation in elementary student language: A multi-dimensional perspective*. Unpublished doctoral dissertation, Northern Arizona University, Flagstaff.

Robinson, P., & Gilabert, R. (2007). Task complexity, the cognition hypothesis and second language learning and performance. *IRAL - International Review of Applied Linguistics in Language Teaching*, *45*(3), 161–176. https://doi.org/10.1515/iral.2007.007

Römer, U., Skalicky, S., & Ellis, N. C. (2020). Verb-argument constructions in advanced L2 English learner production: Insights from corpora and verbal fluency tasks. *Corpus Linguistics and Linguistic Theory*, *16*(2), 303–331. https://doi.org/10.1515/cllt-2016-0055

Senel, B. (2025). Lexical bundle frequency as a construct-relevant candidate feature in automated scoring of L2 academic writing. arXiv preprint arXiv:2504.08537.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Sinclair, S., & Rockwell, G. (2016). *Voyant Tools*. Web. http://voyant-tools.org/.

Stefanowitsch, A., & Gries, S., Th. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, *8(*2), 209–243. https://doi.org/10.1075/ijcl.8.2.03ste

Struthers, L., Lapadat, J. C., & MacMillan, P. D. (2013). Assessing cohesion in children's writing: Development of a checklist. *Assessing Writing*, *18,* 187-201. https://doi.org/10.1016/ j.asw.2013.05.001

Taguchi, N., Crawford, W., & Wetzel, D. Z. (2013). What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. *TESOL Quarterly*, *47*, 420-430. https://doi.org/10.1002/tesq.91

The jamovi project (2023). *jamovi*. (Version 2.4) [Computer Software]. Retrieved from https://www.jamovi.org.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition.* Harvard University Press.

Tywoniw, R., & Crossley, S. (2019). The effect of cohesive features in integrated and independent L2 writing quality and text classification. *Language Education & Assessment*, *2*(3), 110–134. https://doi.org/10.29140/lea.v2n3.151

Uysal, I., & Doğan, N. (2021). How reliable is it to automatically score open-ended items? *International Journal of Assessment Tools in Education*, *8*(Special Issue), 1–12.

Uzun, K. (2023). *Predicting independent L2 writing score using grammatical accuracy indices*. Paper presented at the 12th International Istanbul Scientific Research Congress on Humanities and Social Sciences.

Wang, X. (2014). The relationship between lexical diversity and EFL writing proficiency. *University of Sydney Papers in TESOL*, *9*, 65-88.

Weigle, S. C. (2013). *Assessing writing*. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal0052

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge University Press.

Yavuz, F. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*, *56*(2), 612–634.

Zhang, X. (2022). The relationship between lexical use and L2 writing quality: A case of two genres. *International Journal of Applied Linguistics*, *32*(3), 540–558. https://doi.org/10.1111/ijal.12420

Yayıncı Notu

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Publisher's Note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Appendix

Writing Evaluation Rubric

	Weak (0.5)	Needs Improvement (1.0)	Good (1.5)	Strong (2.0)
VOCABULARY	Predominantly high-frequency, contextually inappropriate, non-natural vocabulary and collocations. High frequency of repetitions without using synonyms. Frequent lexical and/or derivation errors.	Frequent high-frequency, contextually inappropriate, non-natural vocabulary and collocations with some deviations from the topic. Certain parts with repetitions. More than a few lexical or derivation errors.	Mainly low-frequency, contextually appropriate, natural vocabulary and collocations related to the topic with a few deviations. Generally avoids repetition with synonyms. A few lexical or derivation errors.	Predominantly low-frequency, contextually appropriate, natural vocabulary and collocations related to the topic. Avoids repetition with good use of synonyms. Little or no lexical or derivation errors.
GRAMMAR	A lack of appropriate, semantically-accurate, complex structures and affixation used in academic writing. Erroneous, distractingly repetitive, too short or too long sentences making the text difficult to understand and causing communication breakdown. Frequent tense and aspect errors.	Frequent inappropriate, semantically-inaccurate, simple structures and affixation that deviate from academic writing conventions. Frequent sentences that are too long or too short making the text difficult to read at times. Some communication breakdowns. More than a few tense and aspect errors.	Generally appropriate, semantically-accurate, complex structures and affixation used in academic writing with a few deviations. Generally appropriate sentence length. A few tense and aspect errors.	A wide range of appropriate, semantically-accurate, complex structures and affixation used in academic writing. Appropriate sentence length. Little or no tense and aspect error.
TASK COMPLETION	Unclear introduction, main body and conclusion with no thesis statement. Does not present, support and conclude arguments in the main body or develops them weakly in a run-on manner in the same paragraph. Does not consolidate the thesis in the conclusion. Irrelevant to the topic, unable to reach the word limits or neatness requires attention.	Visible clarity problems with introduction, main body and conclusion. Weak thesis statement. Clarity problems with presentation, support and conclusion for arguments that may be run-on in the same paragraph. Consolidation is present but not well-connected to the rest of the essay. Content irrelevant to the topic at times, a bit far off of the word limits and is not neat.	Generally clear introduction, main body and conclusion with a thesis statement to develop arguments from. Main body arguments are mostly related to the thesis and a paragraph is dedicated per argument with slight deviations. The thesis is consolidated in the conclusion. Mostly relevant to the topic, around the word limits, sufficiently neat and assembled with a professional look.	A clear introduction, main body and conclusion. Includes a thesis statement with a few points to develop arguments from. Presents, supports and concludes arguments in the main body directly related to the thesis. A paragraph is dedicated per argument. The thesis is consolidated in the conclusion. Relevant to the topic, within the word limits, neat and correctly assembled with a professional look.
COHERENCE – COHESION	Conceptually unrelated paragraphs that do not form a whole. Ideas do not flow in a logical order, marked by serious disorganisation in their connection. Thought pattern exists but difficult to follow. A lack of appropriate connectives, pronouns and articles.	Paragraphs only slightly related and loosely form a whole. Many ideas do not flow in a logical order, causing disorganisation in many parts. Thought pattern loosely expressed. Frequent problems in the use of connectives, pronouns and articles.	Generally related paragraphs that form a whole. Ideas mostly clear, flow in a logical order and are connected coherently with only a few deviations. Connectives, pronouns and articles are mostly used appropriately with only a few problems.	Conceptually related paragraphs that form a whole. Ideas clearly presented, flow in a logical order and are connected coherently. Appropriate use of connectives, pronouns and articles where necessary.
SPELLING – PUNCTUATION	Frequent spelling, punctuation or capitalization errors which distract the reader and make the text difficult to read.	More than a few spelling, punctuation or capitalization errors. More than a few problems with the use of conventions that may partially reduce readability.	A few spelling, punctuation or capitalization errors. A few problems with the use of conventions that may slightly reduce readability.	Little or no spelling, punctuation or capitalization error. Conventions effectively used to enhance readability.