Review of Applied Science and Technology

Volume: 1; **Issue:** 4
Pages: 183– 215

Corpus-Based Evaluation Models for Quality Assurance Of AI-Generated ESL Learning Materials

Fahimul Habib[1];

[1]. Master of Arts in Applied Linguistics and EL, Chittagong Independent University; Bangladesh;
Email: fahimulhabib@gmail.com

Doi: 10.63125/m33q0j38
Received: 12 September 2022; **Revised:** 20 October 2022; **Accepted:** 20 November 2022; **Published:** 25 December 2022

Abstract

This study addresses the problem that AI-generated ESL learning materials can appear fluent yet vary in accuracy, level appropriateness, and coherence, weakening quality assurance for large-scale cloud and enterprise deployment. The purpose was to develop and validate a corpus-based evaluation model that links corpus indicators to stakeholder quality judgments. Using a quantitative cross-sectional, case-based design, N = 120 evaluators assessed M = 80 AI-generated texts across four categories (reading passages, dialogues, grammar explanations, and practice prompts) using a five-point Likert instrument. Key dependent variables were overall QA and subscales for accuracy, clarity, coherence, level appropriateness, and pedagogical usefulness; key independent variables were readability control index, lexical appropriacy score, cohesion score, lexical diversity (HD-D), and grammar error rate (errors per 100 words). Analyses used descriptive statistics, Cronbach's alpha, Pearson correlations, and multiple regression with text-type stability checks. Overall perceived quality was acceptable (overall QA M = 3.84, SD = 0.53), with clarity highest (M = 3.96) and accuracy lowest (M = 3.72). Reliability was strong (overall a = .91). Corpus to human alignment was substantial: readability control correlated with level appropriateness (r = .61), cohesion with coherence (r = .58), lexical appropriacy with clarity (r = .52) and usefulness (r = .49), and grammar error rate with accuracy (r = −.67), all p < .001. A five-predictor regression model predicted overall QA ($F(5,74) = 21.64$, $p < .001$; $R^2 = .59$; Adj. $R^2 = .56$), with grammar error rate the strongest predictor (β = −.41), followed by readability (β = .29), cohesion (β = .24), and lexical appropriacy (β = .21); performance remained stable across text types (R^2 = .52–.61). Implications are that organizations can operationalize QA as automated gates for error density, readability bands, cohesion thresholds, and vocabulary profile alignment, then reserve human review for borderline cases to improve safety, consistency, and turnaround time in enterprise content workflows. Average indicators were overall readability 0.64, lexical appropriacy 0.71, cohesion 0.59, lexical diversity 0.82, and grammar error rate 2.40 per 100 words.

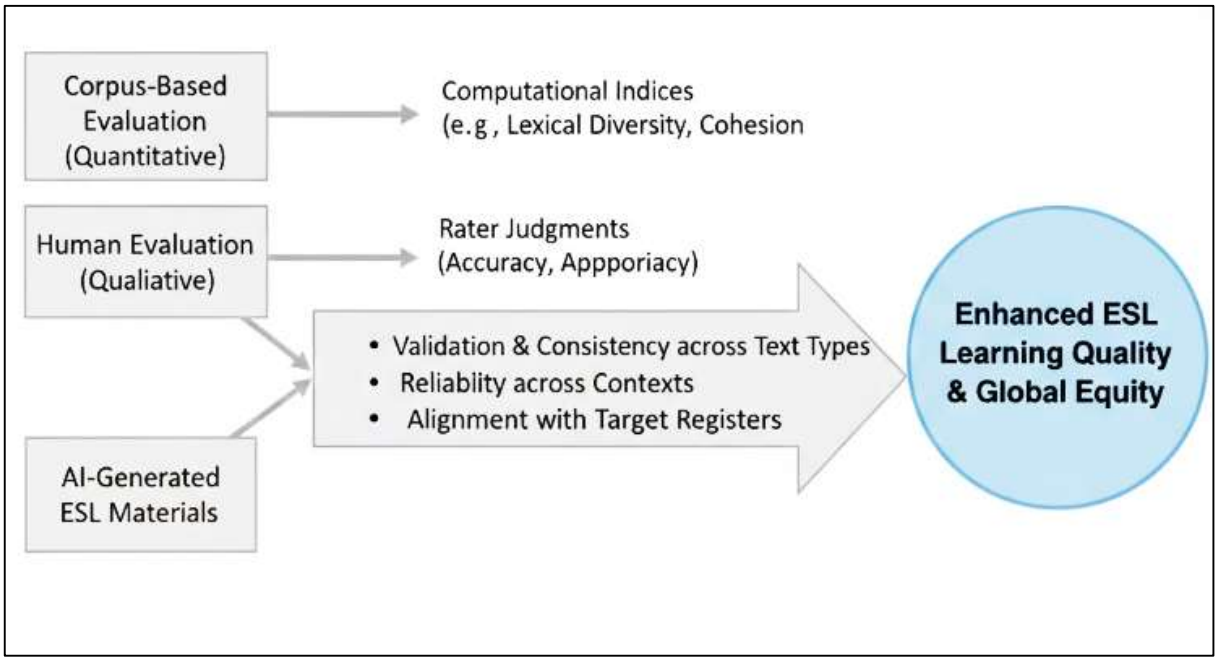
Keywords

Corpus-Based Quality Assurance; AI-Generated ESL Materials; Cohesion and Readability Indices; Lexical Appropriacy; Regression-Based Validation;

INTRODUCTION

Corpus-based evaluation models refer to quantitative approaches that use structured collections of authentic texts (corpora) and computational indices to measure language features in a principled way for educational decision-making. In applied linguistics, this "corpus-based" orientation supports replicable descriptions of vocabulary, grammar, cohesion, and discourse organization because it treats text as observable data rather than impressionistic evidence (Chodorow et al., 2010). A "quality assurance" (QA) model, in the context of ESL learning materials, can be defined as a systematic set of criteria and measurement procedures used to verify that instructional texts meet targeted standards of accuracy, appropriacy, level, and pedagogical usefulness (Crossley et al., 2016). When the materials are AI-generated, QA also includes the verification of consistency across prompts and text types, as well as validation against human expectations of instructional suitability. A practical definition of "AI-generated ESL learning materials" is any learning text (e.g., readings, dialogues, exercises, prompts, explanations) produced by automated natural language generation systems for second-language learners (Guo et al., 2013).

Figure 1: Integrated Corpus Analysis and Human Evaluation Framework

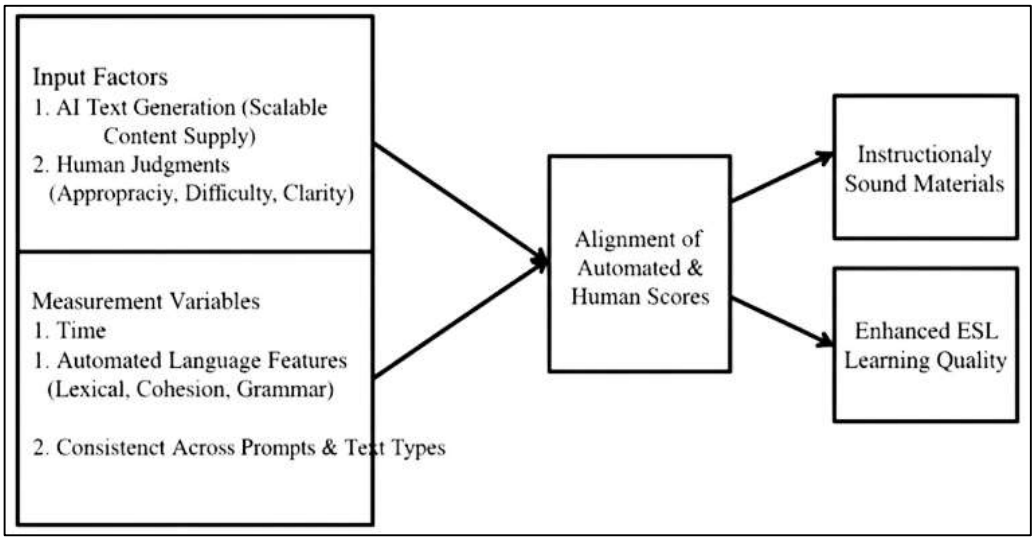


Natural language generation research frames generation as the production of fluent language from non-linguistic inputs and highlights that evaluation must connect automatic scores with human judgments because surface fluency alone is not an educational quality guarantee (Crossley & McNamara, 2012). In educational measurement, "evaluation models" become credible when they specify constructs, operationalize them into measurable indicators, and show that indicators behave reliably across contexts and samples. In corpus-informed language education, a key advantage is that the same text can be assessed on multiple layers—lexical diversity, lexical sophistication, syntactic complexity, and cohesion—using established indices and tools. These definitional foundations place corpus-based QA at the intersection of language learning theory, computational text analysis, and quantitative validation logic, where texts are treated as measurable artifacts and quality claims are treated as testable propositions rather than general impressions (Biber et al., 2011).

English-language learning operates as an international educational infrastructure across schooling, migration, higher education, and professional mobility, which makes the quality of learning materials a global equity issue rather than a local convenience. In many settings, ESL materials function as the primary channel through which learners encounter lexical bundles, academic formulae, and discourse conventions that are associated with gatekeeping assessments and institutional participation (Reiter & Belz, 2009). Corpus-based work has demonstrated that formulaic language is not a decorative element

but a patterned resource that shapes perceived proficiency and communicative credibility, particularly for academic registers (Jinnat & Kamrul, 2021; Ramineni, 2013). This matters internationally because learners encounter English for different purposes—workplace communication, academic study, and civic integration—so materials need to align with the target discourse communities rather than merely present generic grammar. Corpus-informed evaluation adds value here because it can quantify whether texts reflect the lexical and grammatical properties of the targeted register and whether simplified materials preserve coherent discourse relations (Fokhrul et al., 2021; Yang, 2013). For example, computational comparisons of simplified and authentic texts show that simplification changes linguistic profiles in ways that are measurable and sometimes counterintuitive, which means "easier" is not a single-dimensional property. When materials are mass-produced—whether by publishers or automated systems—international relevance also involves consistency: learners across regions should not receive texts that vary wildly in cohesion, lexical control, or grammatical density under the same labeled level. Cohesion analysis supports this requirement by operationalizing how texts signal meaning relations across sentences and paragraphs. In addition, corpus consultation research in ESP indicates that corpus-informed approaches can connect materials to the language of real domains and professional contexts through empirically grounded selection principles. International significance, in this sense, is anchored in the idea that materials quality affects learning opportunities at scale, and corpus-based measurement offers a common technical language for comparing texts across institutions, countries, and delivery platforms (Towhidul et al., 2022; Stevenson & Phakiti, 2014).

Figure 1: Integrated Framework of ESL materials development



AI-generated text introduces a new production logic into ESL materials development: instead of drafting and editing a fixed set of texts, educators can generate large volumes of content quickly and iteratively. This scalability increases the need for rigorous QA because the same underlying generation system can produce outputs that differ in lexical sophistication, discourse structure, factual accuracy, and pedagogical clarity across prompts and topics (Boulton, 2012). Natural language generation scholarship emphasizes that evaluation is central because automatic metrics often correlate imperfectly with human judgments, and validity requires evidence that metrics capture what people actually perceive as quality. In educational contexts, "quality" has a specialized meaning: a text can be fluent and still be instructionally weak if it mislevels vocabulary, uses unstable grammatical patterns, or lacks coherent progression for learners (Crossley et al., 2007). Text cohesion becomes relevant here because instructional readability is partly an outcome of how ideas are connected, repeated, and signposted, not simply of sentence length. Coh-Metrix research formalizes this point by providing multi-level cohesion and difficulty indices that represent discourse relations and conceptual continuity. At the same time, lexical sophistication tools underline that vocabulary quality involves frequency, range,

academic language, and psycholinguistic properties rather than a single count of "difficult words." For AI-generated materials, this implies that QA needs to examine distributions of lexical and cohesion properties across batches of texts, not merely evaluate a single sample. Furthermore, research in automated scoring and evaluation shows that model building can be transparent and psychometrically grounded when it uses clearly defined features and demonstrates stable relations with external criteria (Ashraful et al., 2020; Plonsky & Derrick, 2016). These perspectives converge on a shared requirement: AI text generation can be treated as a content supply mechanism, but educational acceptability depends on measurement-backed verification of language constructs associated with learning goals. The conceptual bridge from NLG evaluation to ESL materials QA is the logic of validation—showing that computed indices and models align with human judgments of appropriacy and difficulty and that they behave consistently across contexts and text types (McNamara et al., 2010).

This study is designed to develop and validate a corpus-based evaluation model that can be used for quality assurance of AI-generated ESL learning materials within a defined real-world case-study context. The primary objective is to operationalize "quality" as a set of measurable, instructionally meaningful dimensions and to demonstrate how these dimensions can be predicted and explained through corpus-derived linguistic indicators. To achieve this, the study first aims to construct a controlled corpus of AI-generated ESL materials produced under consistent generation conditions, representing multiple common instructional text types such as reading passages, dialogues, grammar explanations, and practice tasks. A second objective is to extract a comprehensive set of corpus-based features from these materials that reflect lexical, syntactic, and discourse properties relevant to ESL learning, including indicators of readability and level appropriateness, lexical frequency suitability, lexical diversity, grammatical stability, and cohesion strength. A third objective is to design a structured five-point Likert evaluation instrument that captures expert or practitioner judgments of key QA dimensions, including perceived accuracy, clarity, coherence, level appropriateness, pedagogical usefulness, and overall instructional quality, and to establish the internal consistency of these scales. A fourth objective is to quantify the statistical relationships between corpus-based indicators and human quality judgments through descriptive statistics and correlation analysis, thereby identifying which language features align most strongly with perceived quality dimensions and which features show weak or inconsistent associations. A fifth objective is to build regression models that estimate overall QA scores from the corpus-based indicators, enabling the identification of significant predictors and the estimation of the explanatory power of the model in accounting for perceived quality variation across AI-generated materials. A sixth objective is to test whether the developed model behaves consistently across different text categories within the dataset by examining robustness and stability patterns, ensuring that the evaluation logic is not confined to a single genre or instructional format. Finally, the study aims to present the resulting corpus-based QA model as a clearly specified, reproducible evaluation procedure that combines transparent linguistic measurements with human-centered quality criteria, offering a structured foundation for assessing AI-generated ESL learning materials through empirically testable indicators and statistically validated relationships.

LITERATURE REVIEW

The literature review for this study is positioned at the intersection of corpus linguistics, second language pedagogy, educational measurement, and computational text evaluation, with a specific focus on how these traditions can be synthesized into a quality assurance (QA) framework for AI-generated ESL learning materials. Because QA in instructional contexts requires both construct clarity and measurement credibility, prior research is examined through two complementary lenses: (a) how "quality" in ESL materials has been conceptualized in terms of linguistic accuracy, level appropriateness, coherence, clarity, and pedagogical usefulness, and (b) how these constructs can be operationalized using corpus-based indicators that are scalable and replicable across large text sets. In parallel, the review considers evidence from automated writing evaluation and natural language generation evaluation to understand how computational metrics relate to human judgments and how feature-based models can be validated using correlation and regression approaches. A central premise guiding this review is that AI-generated materials behave like a rapidly produced "generated corpus," meaning their evaluation benefits from corpus-informed methods that can quantify lexical, syntactic, and discourse characteristics across batches, genres, and levels rather than relying on isolated human

impressions. Accordingly, the review maps the demonstrable contributions of cohesion analysis, lexical sophistication profiling, lexical diversity measurement, and syntactic complexity assessment to the practical demands of ESL material auditing. It also examines methodological practices relevant to this study's design, including the use of Likert-based instruments for capturing stakeholder perceptions of instructional quality, approaches to reliability and validity that strengthen interpretability, and analytic strategies that support predictive modeling in educational settings. Finally, the review establishes the conceptual foundation for building a corpus-to-QA model by identifying validated text features that plausibly correspond to the quality dimensions evaluated by educators and learners, and by clarifying how a case-study context can be used to ground model development in real instructional practices while maintaining quantitative rigor. Through this synthesis, the literature review builds the rationale for selecting specific constructs, indicators, and validation strategies that will support the development of a transparent, empirically testable corpus-based QA evaluation model for AI-generated ESL learning materials.

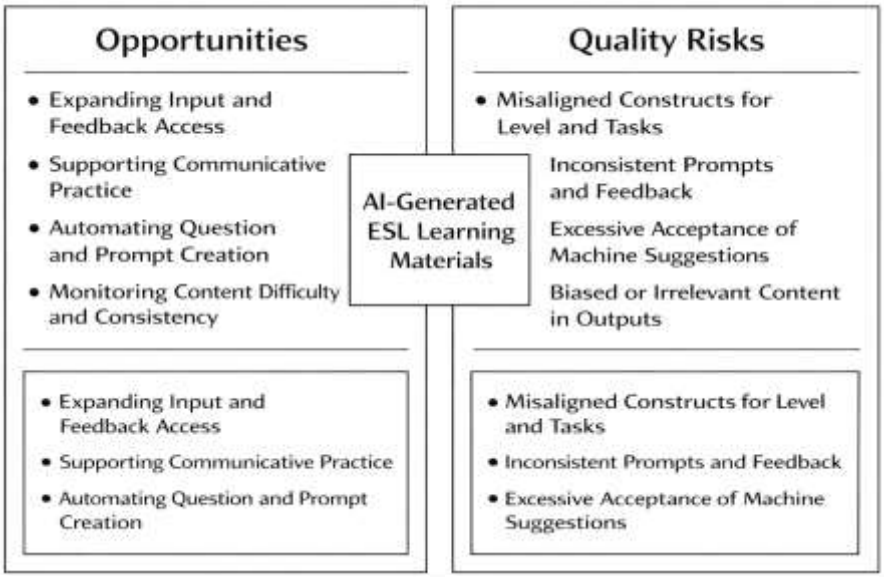
AI-Generated ESL Learning Materials

AI-supported authoring in ESL contexts can be defined as the use of computational systems to produce, adapt, or scaffold instructional language that learners read, respond to, or practice with, while quality assurance refers to systematic procedures that verify whether those outputs meet requirements for level, accuracy, coherence, and pedagogical fit. Under this definition, AI-generated ESL materials include complete texts such as readings and dialogues, plus instructional components such as example sentences, prompts, feedback messages, and automatically produced practice items. The central opportunity reported across technology-enhanced language learning is the ability to expand access to input and feedback at scale, especially in settings where teacher time is limited and learner needs are diverse. A practical illustration is the way natural language processing has been embedded in intelligent tutoring systems to analyze learner writing and provide strategy-focused support through measurable linguistic and rhetorical features (McNamara et al., 2013). For materials development, the same analytic logic enables rapid auditing of generated texts for lexical control, syntactic load, cohesion signals, and error patterns, allowing large collections of outputs to be summarized rather than inspected one by one. In classroom-facing systems, automated writing evaluation can also encourage iterative revision by giving immediate, repeatable feedback and by creating a record of changes across drafts. Empirical research in ESL writing instruction has shown that learners' uptake and perceptions of automated feedback depend on how the tool is integrated into pedagogy, the kinds of revisions it prompts, and the extent to which instructors mediate its use (Li et al., 2015). These findings frame AI-generated materials as productive starting points for instruction, with quality assurance serving as the mechanism that keeps scalability aligned with instructional intent and learner level. In addition, automated pipelines make documentation easier, so reviewers can trace decisions from prompt to published lesson sets.

A second cluster of opportunities concerns diversification of learning activities and interactional formats, which is especially relevant for programs seeking to increase communicative practice and formative assessment without proportional increases in staffing. Conversation is a core driver of language development, yet sustained practice can be constrained by class size, learner anxiety, and limited access to proficient partners. Conversational agents and chatbot interfaces can offer a low-stakes practice space that supports repeated turns, topic variation, and individualized pacing, while capturing interaction logs that can inform later review. Research on chatbots as language learning partners indicates that learner interest and perceived learning value are closely linked to engagement with the agent, and that individual differences in language competence shape how learners experience chatbot interaction (Fryer et al., 2019). For materials production, this interactional capacity can be paired with automated content generation so that practice prompts, role-play scripts, and situational dialogues can be produced in multiple versions for different proficiency bands or communicative goals. Another major opportunity is automated item creation for comprehension checks and assessment, which can reduce the time required to draft questions, expand coverage of texts, and enable consistent question templates across large content libraries. Work on automatic generation of short-answer questions for reading comprehension demonstrates how linguistic analysis can be used to produce questions that target selected semantic and syntactic relations in source texts, supporting efficiency and consistency

in item writing (Huang & He, 2016). In ESL settings, these capabilities make it feasible to attach checks for understanding to every passage, rotate items to reduce memorization, and align practice with specific language points such as wh-movement, reference, or tense control.

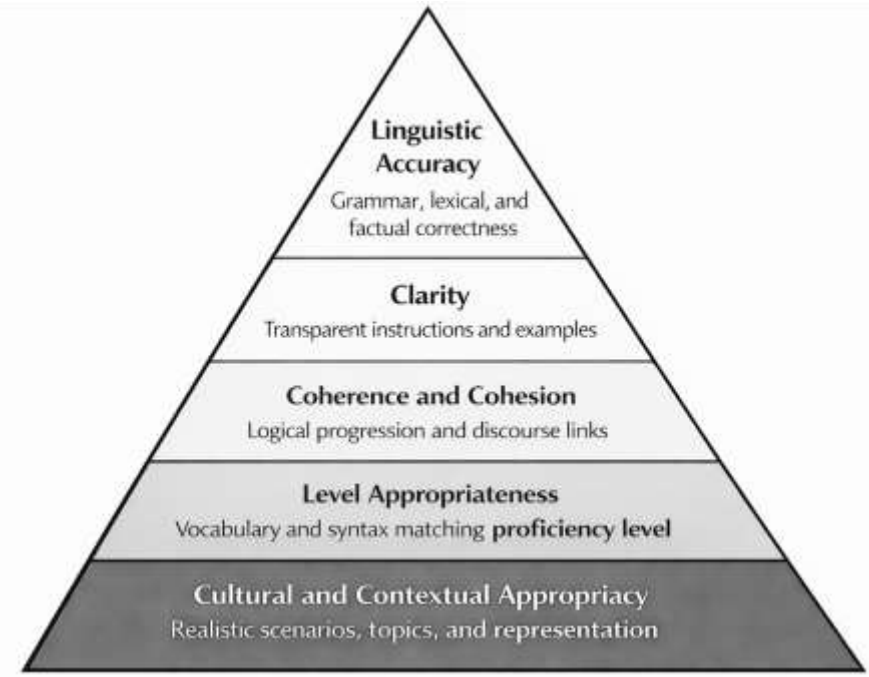
Figure 2: Opportunities And Risks In AI-Generated ESL Materials



Quality Dimensions and QA Standards for ESL Learning Materials

Quality in ESL learning materials is commonly treated as a multidimensional construct that can be evaluated through both pedagogical fitness and linguistic evidence. In practice-oriented research on materials development, quality is typically anchored in whether materials support learning goals through appropriate content selection, coherent sequencing, and learner engagement features that make input usable in instruction. Within this view, core quality dimensions include linguistic accuracy (freedom from grammar, lexical, and factual errors), clarity (transparent instructions and examples), coherence and cohesion (logical progression and explicit discourse links), and level appropriateness (alignment of vocabulary, syntax, and discourse demands to the targeted proficiency level). Materials standards also include representational quality, such as the consistency of tasks with communicative purposes and the degree to which topics and examples reflect meaningful contexts rather than isolated forms. These dimensions can be framed as quality assurance criteria because each one can be operationalized into observable checks: error auditing for accuracy, instruction readability checks for clarity, cohesion indicators for discourse flow, and controlled difficulty profiling for level fit. In state-of-the-art work on materials development, the evaluation of materials is presented as a systematic process that involves examining both internal features (language, tasks, and content) and external considerations (context, goals, and learners), thereby positioning QA as a deliberate methodology rather than an informal impression (Tomlinson, 2012). For AI-generated ESL resources, this QA logic is particularly important because fluent surface form can conceal instability in task design or proficiency targeting. As a result, a high-trust QA approach requires clearly defined quality dimensions, explicit indicators for each dimension, and a replicable evaluation process that supports consistent judgments across materials batches and text types.

Figure 3: Pyramid Framework of ESL Materials Quality Dimensions



A key QA standard for ESL materials is alignment with established principles of second language acquisition, because materials can be linguistically correct while still failing to support learning in ways consistent with how acquisition is facilitated. From an SLA-informed perspective, quality includes the degree to which materials provide meaningful exposure, promote interaction with form–meaning relationships, and create opportunities for use that are plausible for the learner's level. One way to operationalize this standard is to analyze materials through a principled framework that assesses activities and tasks against SLA-oriented criteria, generating structured evidence about what the materials actually require learners to do. For example, research on textbook selection has proposed systematic procedures for analyzing course materials through SLA principles and task analysis so that selection decisions are grounded in observable activity patterns rather than broad claims on covers or teacher intuition (Guilloteaux, 2013). This approach makes QA criteria actionable by shifting evaluation from general labels (e.g., "communicative," "B1 level") toward measurable properties such as the distribution of task types, the balance of skills and language systems, and the cognitive and interactional demands embedded in activities. In AI-generated ESL materials, this matters because generated tasks may overuse one format (e.g., multiple-choice or short dialogues) or may create prompts that are misaligned with intended communicative outcomes. A related QA standard is the use of structured evaluation tools that enable consistent decisions across reviewers. Materials evaluation research has emphasized checklist-based systems as a practical mechanism for reliability, offering staged screening and detailed evaluation processes that organize criteria such as relevance, linguistic level, methodological fit, and contextual appropriacy into a coherent auditing workflow (Işik, 2018).

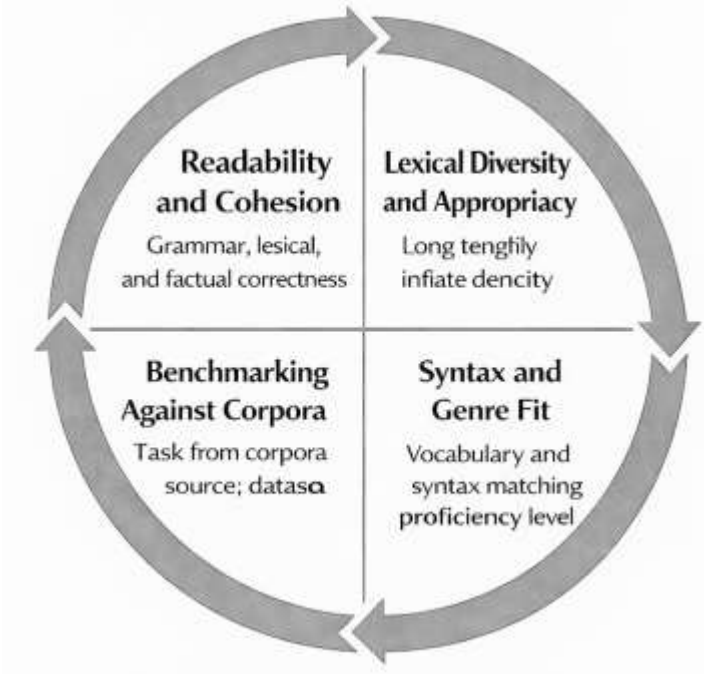
ESL Material Quality Assurance

Corpus-based evaluation treats **text quality** as a measurable construct derived from observable linguistic patterns across large bodies of authentic and pedagogic language. Within ESL materials, this approach typically operationalizes quality through readability, cohesion, lexical appropriacy, and genre/register fit, because these attributes directly shape comprehensibility and instructional usefulness at scale. A key advantage of corpus methods is that they allow researchers to benchmark AI-generated materials against reference distributions drawn from level-graded textbooks, learner corpora, or target-domain corpora, rather than relying only on subjective judgment. Computational discourse analysis has been central to this shift, particularly through multidimensional cohesion and readability profiling. For example, Coh-Metrix demonstrated how cohesion relations, lexical frequency, and syntactic patterning can be jointly quantified to build richer descriptions of text difficulty than sentence-length formulas, enabling fine-grained QA checks aligned with psycholinguistic features of

comprehension (Graesser et al., 2006). In ESL contexts, such profiling matters because learner-facing materials must balance clarity with enough linguistic richness to support acquisition goals. In corpus-informed QA, the practical implication is methodological: texts are evaluated not as isolated products, but as instances that should approximate level-appropriate linguistic signatures. Therefore, corpus-based metrics become the measurable "targets" a QA model can predict, track, and flag when AI-generated outputs drift from acceptable ranges for a defined learner level and instructional purpose.

A second cluster of corpus-based evaluation focuses on lexical diversity and lexical quality, which are often treated as indicators of how well materials can support vocabulary growth while maintaining accessibility. Lexical diversity metrics have multiplied because simple type–token ratios are unstable across text lengths; consequently, modern corpus research evaluates the comparability and stability of alternative indices to ensure that observed differences reflect vocabulary use rather than artifacts of sampling. This is especially important for AI-generated ESL materials, where output length and paraphrase variability can inflate or deflate diversity scores in misleading ways. Work comparing prominent indices such as VOCD-D and HD-D illustrates how selection of a diversity metric can change interpretations of "richness," underscoring the need for QA systems to standardize lexical indicators and control text-length sensitivity (deBoer, 2014). Beyond diversity alone, corpus-based QA increasingly pairs lexical measures with proficiency frameworks to test whether lexical profiles discriminate known levels. Evidence from CEFR-linked writing shows that lexical diversity operationalizations can meaningfully differentiate proficiency bands, strengthening the argument that lexical metrics can serve as criterion-referenced QA signals rather than generic style descriptors (Treffers-Daller et al., 2016). In a thesis on quality assurance of AI-generated ESL materials, these findings justify treating lexical diversity not as a single score but as a validated measurement family, used alongside constraints (length, prompt type, task genre) so the QA model can identify outputs that are "lexically noisy," overly repetitive, or mismatched to target proficiency.

Figure 4: Corpus-Based Evaluation Cycle for Quality Assurance of ESL Learning Materials



A third strand emphasizes syntactic complexity and genre-sensitive evaluation, arguing that "quality" depends on whether complexity aligns with communicative purpose and task type rather than maximizing structural elaboration. Corpus-based studies of L2 writing have shown that complexity can be meaningfully quantified, but also that complexity indices must be interpreted through genre and proficiency lenses to avoid false judgments about textual appropriateness. A widely cited synthesis of syntactic complexity research highlights the methodological requirement to select complexity

measures that match the discourse demands of the target text type, reinforcing that QA for ESL materials should be **context-aware** rather than purely form-driven (Ortega, 2015). Complementing this, corpus analyses of intensive academic writing contexts demonstrate that lexical and syntactic development can be tracked with multiple theoretically motivated measures, revealing patterned shifts in noun-phrase complexity and lexical diversity that characterize academic argumentation (Mazgutova & Kormos, 2015). For AI-generated ESL learning materials, this implies that QA should not only judge "difficulty," but also test whether syntactic and lexical profiles fit the intended instructional genre (e.g., narrative reading passage, grammar explanation, short dialog, argumentative paragraph). As a result, corpus-based evaluation supports a QA architecture where the model predicts a bundle of indicators—cohesion/readability, lexical diversity/appropriacy, and syntax/genre fit—and then validates outputs through alignment with reference corpora, enabling defensible, measurement-driven judgments of material quality.

Theoretical Framework for this study

Technology-acceptance theory provides a rigorous way to explain why teachers, instructional designers, and learners decide to adopt or reject technology-mediated learning resources, and it is especially useful when the "technology" is not only a delivery platform but also a generator of instructional content. In this study, AI-generated ESL learning materials are treated as an acceptance object because their educational value depends on whether stakeholders judge them to be usable, reliable, and instructionally appropriate within real teaching and learning routines. The Technology Acceptance Model (TAM) and its extensions typically explain intention to use and actual use through perceived usefulness (PU) and perceived ease of use (PEOU), with attitude and behavioral intention often serving as mediating outcomes. Meta-analytic evidence has shown that TAM paths are robust across settings and can provide a stable baseline for modeling adoption behavior, which is important when the goal is to connect perceptions of quality to measurable predictors in a consistent way (King & He, 2006). For a QA-focused thesis, this theoretical base supports a logic in which stakeholder acceptance is not treated as a vague preference but as a measurable outcome linked to specific perceptual constructs. The model is also appropriate because it can be aligned with survey-based measurement: PU can be operationalized through items reflecting perceived instructional value (e.g., usefulness for learning objectives, classroom applicability), while PEOU can be operationalized through items reflecting ease of integrating the generated materials (e.g., clarity, editability, time savings, usability of output). In addition, technology acceptance theory is compatible with quantitative, cross-sectional designs that rely on structured questionnaires and statistical modeling, allowing perceived quality dimensions to be positioned as predictors of acceptance-related outcomes. This framing strengthens the research by making stakeholder judgment a theoretically grounded criterion and by providing a consistent structure for interpreting how perceived quality translates into readiness to use AI-generated ESL content.

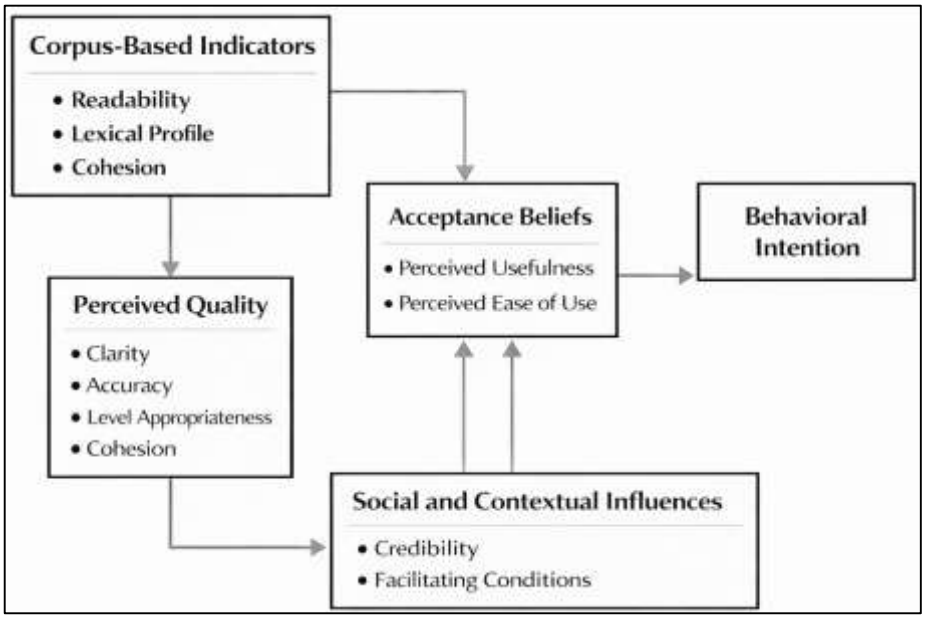
The unified theory tradition extends acceptance logic beyond TAM by emphasizing performance expectancy, effort expectancy, social influence, and facilitating conditions, and by recognizing that adoption decisions are shaped by both individual beliefs and contextual enablers. For AI-generated ESL materials, this extension is valuable because acceptance is influenced not only by perceived usefulness and ease, but also by whether reviewers believe the materials are professionally credible, aligned with institutional standards, and supported by the surrounding workflow (editing policies, review time, platform constraints). UTAUT2 further incorporates constructs such as habit and hedonic motivation in consumer contexts, which helps explain repeated use once a tool becomes embedded in routine content production or study practice (Venkatesh et al., 2012). In this thesis, acceptance theory is used as a "trust-in-quality" lens: the more consistently AI-generated materials meet expectations for accuracy, coherence, level fit, and pedagogical usefulness, the more likely stakeholders are to judge them as useful and to integrate them into practice. This relationship can be expressed in a simple predictive form that matches the study's regression approach. For example, an acceptance outcome such as Behavioral Intention (BI) can be modeled as:

$$BI = \beta_0 + \beta_1(PU) + \beta_2(PEOU) + \epsilon$$

and, in a QA context, PU and PEOU can be modeled as functions of perceived quality dimensions that

the Likert instrument measures (clarity, accuracy, coherence, level appropriateness, usefulness). This integration is consistent with the broader e-learning acceptance literature showing that acceptance relationships vary by user type and technology type, which supports examining acceptance patterns across different evaluator roles and different ESL text categories within the case study (Šumak et al., 2011). The framework therefore provides a disciplined explanation for why perceived quality matters: quality dimensions are not only evaluative labels, they are mechanisms that shape perceived usefulness, reduce perceived effort, and stabilize confidence in the instructional value of generated materials.

Figure 5: Trust-In-Quality Technology Acceptance Framework For AI-Generated ESL Materials



A General Extended Technology Acceptance Model for E-Learning (GETAMEL) offers such a strategy by emphasizing commonly recurring external factors and mapping them to PU and PEOU in a systematic way, which helps avoid arbitrary variable selection (Abdullah & Ward, 2016). In this study, corpus-based indicators act as external quality evidence that complements human perceptions: they quantify linguistic properties that plausibly drive perceived clarity, level appropriateness, and coherence, which then shape acceptance-related beliefs. This layered structure allows the research to connect objective text metrics to subjective evaluations within a single theoretical frame. The combined QA–acceptance logic can be represented with two linked equations that match your planned analyses: first, perceived quality outcomes predicted by corpus metrics, and second, acceptance beliefs predicted by perceived quality. For example, overall perceived quality (PQ) can be modeled as:

$$PQ = \alpha_0 + \alpha_1(Readability) + \alpha_2(LexicalProfile) + \alpha_3(Cohesion) + \alpha_4(ErrorRate) + \epsilon$$

followed by a belief or intention model such as:

$$PU = \gamma_0 + \gamma_1(PQ) + \epsilon$$

This formulation fits the thesis's quantitative design because it justifies why correlation and regression are appropriate for testing relationships among corpus indicators, perceived quality constructs, and acceptance-oriented beliefs. It also strengthens interpretability: when a corpus feature significantly predicts a quality dimension, the theory explains why that dimension would matter for use decisions in a real instructional workflow. Finally, the framework highlights the importance of social and contextual moderation, which aligns with evidence that subjective norm and context can influence adoption patterns and can change the strength of classical TAM relationships across groups and settings (Schepers & Wetzels, 2007). In short, technology acceptance theory anchors your QA model in a well-established explanatory tradition that treats perceived quality as a determinant of real use readiness, while keeping the empirical modeling aligned with interpretable constructs.

Conceptual Framework For this study

A defensible conceptual framework for corpus-based quality assurance (QA) begins by treating AI-generated ESL materials as a measurable linguistic product whose instructional value can be audited through two complementary evidence streams: (1) observable textual properties derived from corpus analytics and (2) human judgments captured through structured rating scales. This dual-evidence stance mirrors long-standing arguments in automated writing evaluation that machine-based indicators must be interpreted as construct-relevant proxies rather than as direct substitutes for pedagogical quality, because automated systems tend to capture surface, linguistic, and formatting regularities more readily than deeper instructional intent (Deane, 2013). Within the C2QA model, corpus analytics provide standardized predictors (e.g., lexical frequency distribution, lexical diversity, cohesion signals, syntactic patterning), while human evaluators provide criterion measures of QA dimensions aligned with ESL material requirements (accuracy, clarity, coherence, level appropriateness, and pedagogical usefulness). The conceptual logic is that quality becomes more trustworthy when (a) the rating instrument is internally consistent and (b) measurable linguistic features show stable statistical relationships with those ratings across a defined case-study context. Research in second-language assessment has similarly emphasized that validity depends on explicit construct definition and cautious interpretation of automated evidence, especially for English language learners whose language performance and instructional needs vary by context and purpose (Weigle, 2013). Accordingly, C2QA treats "quality" as a multi-dimensional latent construct expressed through subscale scores (Likert items) and supported by a transparent mapping from corpus indicators to each subscale. This mapping is not assumed; it is tested empirically through correlation and regression, which makes the framework compatible with your quantitative cross-sectional design while keeping the interpretation anchored to observable text characteristics and explicit human criteria. In this way, the conceptual framework positions the corpus as a reproducible audit lens and the evaluators as construct anchors, producing a QA model that can be statistically examined for alignment, consistency, and explanatory power within the study's bounded setting.

The C2QA framework is operationalized as a staged measurement-and-modeling pipeline that links corpus indicators to quality outcomes in a way that can be audited and replicated. Stage 1 defines a corpus of AI-generated materials stratified by text type (e.g., reading passage, dialogue, grammar explanation, practice prompt) and applies automated analysis to extract interpretable features known to relate to readability and comprehension processes, including cohesion and lexical familiarity cues (Crossley et al., 2008). Stage 2 obtains human QA ratings using a rubric-driven Likert instrument that yields subscale scores and an overall QA score, enabling reliability analysis and construct-level profiling of where AI-generated materials perform strongly or weakly. Stage 3 tests bivariate alignment (correlations) and multivariate explanation (regression) between corpus features and QA outcomes, using a model form consistent with your thesis analyses, such as:

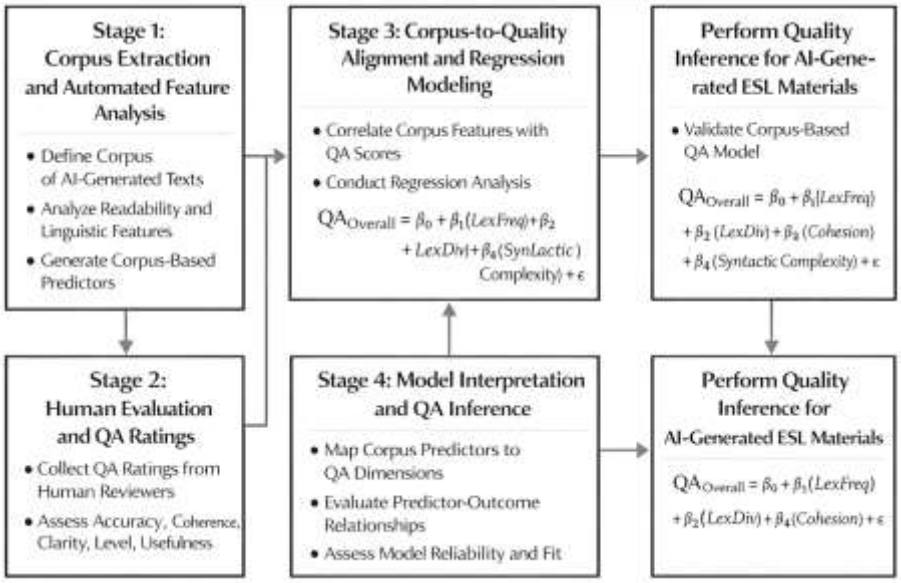
$$QA_{overall} = \beta_0 + \beta_1(LexFreq) + \beta_2(LexDiv) + \beta_3(Cohesion) + \beta_4(SyntacticComplexity) + \epsilon$$

Stage 4 extends interpretability by modeling subscales (e.g., clarity, coherence) separately, so the framework can identify which linguistic signals correspond to specific pedagogical judgments rather than only predicting a single aggregate score. This approach aligns with evidence that lexical indicators can classify proficiency-related language differences and that statistical models built from lexical indices can meaningfully discriminate among learner levels (Crossley et al., 2012). In a QA context, the same logic supports discriminating among AI-generated outputs that differ in level fit or lexical suitability, even when generation prompts appear similar. Reliability of the human-rating instrument is treated as a required checkpoint before interpretation; a standard internal-consistency estimate such as Cronbach's alpha can be computed for each subscale:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_T^2}\right)$$

where $k$is the number of items, $\sigma_i^2$are item variances, and $\sigma_T^2$is total-score variance. Together, these steps make the conceptual framework both measurable and testable, with each inference supported by explicit statistics.

Figure 6: Corpus-Based Predictive QA Framework Linking Text Metrics And Human Ratings

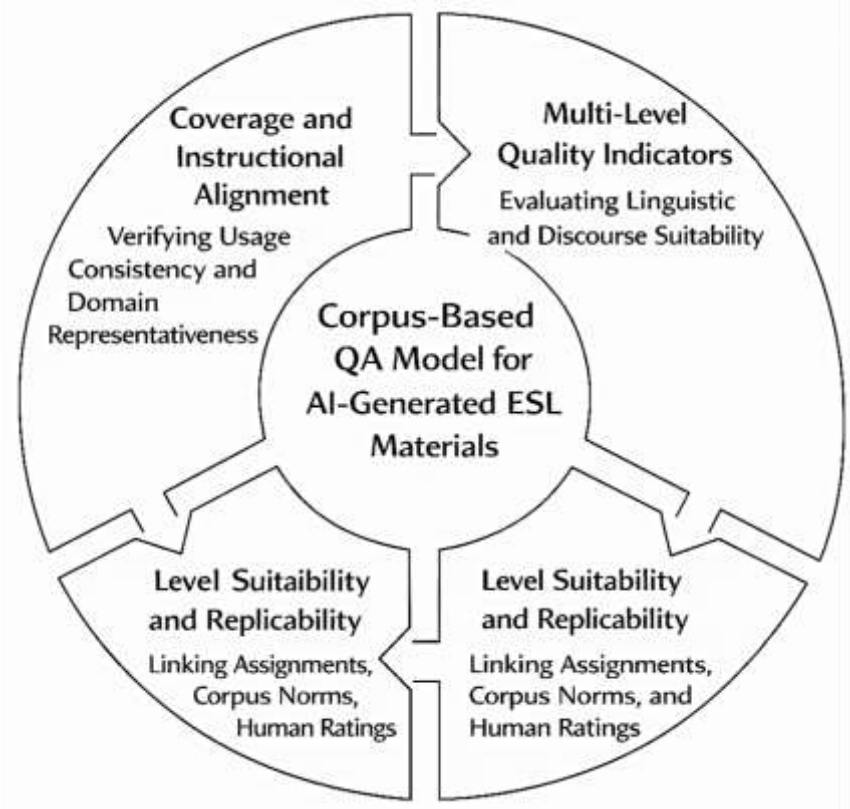


A final component of the conceptual framework is an interpretive layer that defines what counts as "high-quality" AI-generated ESL materials in corpus terms, while remaining constrained by construct validity. C2QA therefore includes three interpretive principles. First, corpus-based indicators are interpreted as *diagnostic contributors* to quality judgments, not as quality itself; this is consistent with cautions that automated scoring often prioritizes detectable textual regularities and requires careful linkage to the intended construct (Weigle, 2013). Second, quality is treated as a function of lexical proficiency and accessibility cues that shape comprehensibility and perceived level appropriateness—dimensions that have been modeled successfully using computational indices in both speech and writing contexts, with lexical diversity, familiarity, and conceptual features emerging as informative predictors of proficiency-related judgments (Crossley et al., 2011). Third, the framework treats model evaluation as part of QA: the statistical model is expected to generalize across the study's internal categories (text types) and to provide stable predictor patterns rather than opportunistic fits. This emphasis reflects a broader AES construct perspective in which model behavior must be explained in terms of what features represent and what they omit (Deane, 2013). Practically, C2QA interprets strong QA performance as convergence between (a) consistent human ratings on relevant subscales and (b) corpus profiles that indicate level-fit vocabulary, coherent cohesion signaling, and manageable linguistic complexity as measured by the selected tools. The conceptual framework therefore defines trustworthiness as *convergent evidence plus transparent modeling*: ratings show what expert stakeholders value, corpus indicators show what the text measurably contains, and regression results show how strongly the measurable properties explain the valued outcomes. This structure keeps the thesis anchored to quantifiable evidence while preserving a clear conceptual boundary between linguistic measurement and pedagogical quality.

Synthesis of Gaps and Rationale for a Corpus-Based QA Model

Corpus-driven textbook and materials evaluation has shown that language-learning resources can diverge from authentic usage in ways that are not always visible through expert review alone. Corpus-based comparisons have demonstrated how pedagogically selected input may overrepresent certain lexico-semantic patterns while underrepresenting high-utility constructions that dominate real-world registers, thereby creating "coverage gaps" that influence what learners repeatedly encounter. For instance, corpus evaluation of metaphor selection in a business English textbook has illustrated that instructional materials can include figurative language sets that overlap only minimally with domain corpora, meaning that learners may practice metaphors that are not frequent, not productive, or not pragmatically central in target discourse communities (Skorczynska Sznajder, 2010).

Figure 7: Synthesis Of Gaps Supporting a Corpus-Based QA Model



Similarly, corpus-based analysis of textbook instructional language has revealed systematic imbalances in skill coverage and task emphasis, highlighting how material design can encode implicit pedagogies that shape learning opportunities at scale (Chan & Cheuk, 2020). These strands of evidence collectively point to a foundational gap: even when materials appear pedagogically coherent, they may not align with usage realities across discourse functions, genres, and classroom task types. In the context of AI-generated ESL learning materials, the risk of such misalignment becomes more complex because generation systems can amplify distributional biases from training data or prompt framing, producing text that is superficially fluent but uneven in discourse balance, lexical representativeness, and task-appropriate pragmatics. Therefore, a corpus-based quality assurance (QA) approach is necessary not only to "rate" generated materials but also to document their representativeness and pedagogical fit through measurable, reproducible indicators. This study has treated corpus-based evaluation as a credibility mechanism: it converts claims about "appropriate ESL materials" into verifiable comparisons between generated output and reference corpora that reflect target proficiency and instructional purposes.

A second gap concerns how the field operationalizes "quality" in ways that are simultaneously linguistic, instructional, and model-auditable. Many evaluation practices privilege surface correctness and general readability, but quality in ESL materials also includes discourse-level cohesion, genre expectations, and the stability of linguistic patterns across text categories. Multi-level text analysis frameworks have clarified that discourse processing depends on interacting dimensions such as narrativity, syntactic simplicity, referential cohesion, and causal cohesion, which can be quantified to support principled text selection and evaluation (Graesser et al., 2011). This matters for AI-generated materials because a text can satisfy grammatical well-formedness while still being weakly cohesive, poorly staged, or register-inappropriate for the intended learning goal. At the same time, corpus-informed developmental perspectives show that proficiency-relevant discourse features can be modeled through recurrent sequences and functional patterning rather than isolated words or rules. Work on lexical bundles across CEFR levels has demonstrated that phraseological and discourse-functional features differentiate proficiency bands and can be systematically extracted, annotated, and compared (Chen & Baker, 2016). These insights imply that QA for AI-generated ESL materials should

not be reduced to a single score; instead, it should be structured as a set of linked evidence layers: (a) corpus alignment against reference distributions, (b) discourse-feature profiling that reflects proficiency progression, and (c) consistency checks showing that the model behaves similarly across genres and task types. Without these layers, QA conclusions remain vulnerable to overgeneralization because improvements in one dimension (e.g., shorter sentences) may coincide with degradation in another (e.g., reduced cohesion cues or impoverished discourse marking). This study has therefore positioned corpus-based QA as a defensible bridge between applied linguistics constructs and quantitative auditability, enabling results that can be interpreted by both language educators and analytics-oriented reviewers.

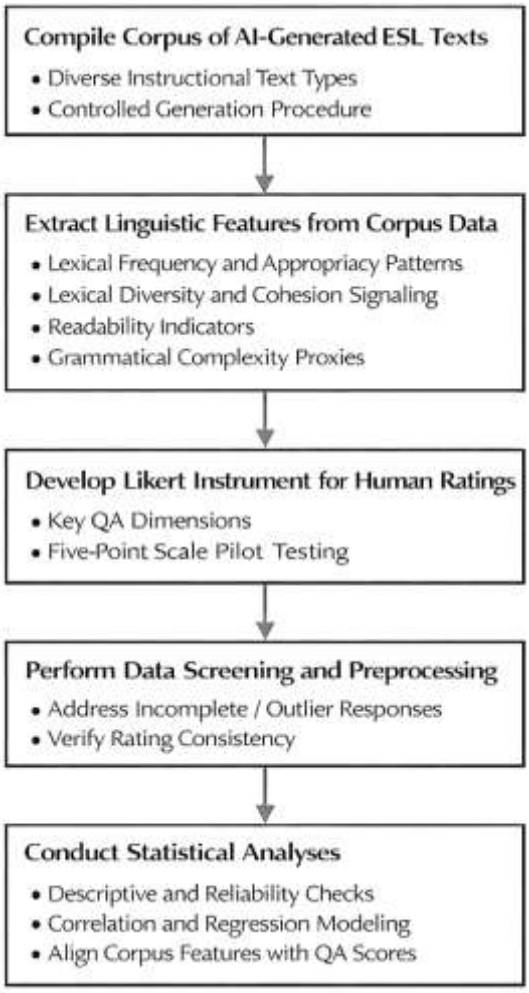
A third gap concerns level suitability and the replicability of proficiency judgments when scaling evaluation to large volumes of generated content. Human judgments are essential, but they are resource-intensive and can vary across raters, prompting the need for automated, corpus-grounded proxies that remain interpretable. Readability and level classification for second-language learners has been approached through feature-based modeling tied to annotated proficiency data, illustrating that L2 readability can be treated as a classification task supported by linguistic features and evaluation metrics (Xia et al., 2016). For AI-generated ESL learning materials, this suggests that QA should incorporate automated level verification that is benchmarked to learner-relevant standards and then cross-validated using human ratings and corpus evidence. However, current practice often separates these components: corpus comparisons are conducted without integrating learner-level prediction, or readability prediction is performed without tracing which linguistic properties caused a level assignment. This separation creates a transparency gap in QA, where stakeholders may see an overall rating but cannot identify whether problems stem from lexical rarity, syntactic density, cohesion weakness, or unstable performance across genres. The present study has addressed this gap by justifying a unified evaluation logic in which corpus alignment metrics, discourse-feature indicators, and readability/level classification operate as complementary diagnostics rather than competing alternatives. In practical terms, this synthesis motivates the proposed corpus-based evaluation model as an evidence system: it has not only identified whether AI-generated ESL materials are "good" but also localized which quality dimensions are responsible for success or failure, enabling hypothesis-driven quantitative testing through Likert-rated human judgments aligned to the same constructs.

METHOD

The methodology chapter has described how the study has been designed as a quantitative, cross-sectional, case-study–based investigation that has examined the quality assurance (QA) of AI-generated ESL learning materials through corpus-based evaluation and human rating evidence. A bounded case context has been selected so that the generation, selection, and review of instructional materials have been observed under real conditions while a consistent measurement protocol has been maintained. Within this design, a corpus of AI-generated ESL texts has been compiled and has represented multiple instructional text types, such as reading passages, dialogues, grammar explanations, and practice prompts, so that the QA model has been tested across formats that commonly appear in ESL programs.

A controlled generation procedure has been applied so that prompts, length targets, and intended proficiency levels have been held constant within categories, thereby enabling systematic comparison across material sets. Linguistic features have been extracted from the compiled corpus using established corpus-analytic tools and computational indices, and these features have captured lexical, syntactic, and discourse properties relevant to instructional quality, including lexical frequency and appropriacy patterns, lexical diversity behavior, cohesion signaling, readability-related indicators, and grammatical stability proxies. In parallel, a structured five-point Likert instrument has been developed and has been used to collect evaluator judgments of key QA dimensions, including perceived accuracy, clarity, coherence, level appropriateness, pedagogical usefulness, and overall instructional quality.

Figure 8: Methodology of The Research



A pilot phase has been conducted so that item clarity, response consistency, and scale performance have been checked before full deployment, and revisions have been made to strengthen reliability. Data screening procedures have been implemented so that incomplete responses, outlier patterns, and inconsistent rating behavior have been addressed in a transparent manner. Statistical analyses have been conducted so that the study objectives have been met through descriptive profiling, reliability testing, correlation analysis, and regression modeling. The analytical strategy has enabled the identification of corpus features that have aligned with human-rated quality dimensions and has supported the estimation of predictive models that have explained variance in overall QA scores. Robustness checks have also been incorporated so that model stability across text types and sensitivity to potential confounds such as length and topic have been examined within the case-study dataset.

Research Design

This study has been designed as a quantitative, cross-sectional, case-study–based investigation that has examined the quality assurance of AI-generated ESL learning materials through measurable corpus indicators and structured human evaluations. The cross-sectional approach has been used because data have been collected at a single point in time from a defined set of AI-generated materials and a defined group of evaluators, enabling relationships among variables to have been tested without requiring repeated measures. The case-study orientation has been applied to keep the research grounded in an authentic instructional context where AI-generated materials have been produced, reviewed, and interpreted under realistic constraints. Quantitative procedures have been prioritized so that quality dimensions have been expressed as numeric scores and so that statistical techniques have been applied to test alignment and prediction. Descriptive statistics, correlation analysis, and regression modeling have been selected because they have supported both construct profiling and hypothesis testing within a coherent analytical framework.

Case Study Context

A bounded case-study context has been selected so that the study has been situated within a real ESL materials development or instructional setting where AI-generated content has been relevant and practically used. The case has been defined by specifying the institutional or program environment, the targeted learner proficiency range, and the operational workflow through which AI-generated texts have been produced and prepared for instructional use. The context has been documented through a clear description of the materials pipeline, including how prompts have been written, how outputs have been screened, and how texts have been categorized into instructional types such as reading passages, dialogues, grammar explanations, and practice prompts. This contextualization has ensured that the corpus has represented authentic constraints on content creation, such as time limits, level labeling practices, and lesson-format expectations. By bounding the case, comparisons have been made meaningful because the same standards, objectives, and review expectations have been applied across all generated materials.

Population and Unit of Analysis

The population for this study has been defined as the group of stakeholders who have been capable of evaluating ESL learning materials with informed judgment, such as ESL teachers, curriculum designers, instructional content reviewers, or advanced-level ESL practitioners. These participants have been treated as evaluators whose ratings have provided criterion evidence for the quality assurance model. The unit of analysis has been the AI-generated ESL learning material text, with each text instance having served as an observation linked to both corpus-derived linguistic features and human rating outcomes. Texts have been organized into identifiable instructional categories so that the unit of analysis has remained consistent while allowing sub-group analyses across text types. Evaluator responses have been aggregated according to a structured scoring procedure so that each material item has received subscale scores and an overall QA score. This structure has enabled relationships between measurable textual properties and perceived quality judgments to have been tested at the level of the material artifact.

Sampling Strategy

A purposive sampling strategy has been applied so that participants have been selected based on their expertise and their ability to judge ESL material quality reliably. Sampling criteria have been defined to ensure that evaluators have had relevant teaching, assessment, or materials-review experience and that they have been familiar with proficiency leveling expectations. Where access constraints have existed, a structured convenience component has been included, while eligibility rules have still been maintained to protect rating quality. The AI-generated materials set has also been sampled using a controlled selection strategy so that the corpus has included balanced representation across text types, target levels, and lesson purposes. Sampling quotas have been applied so that no single category has dominated the dataset and so that regression analyses have not been driven by genre imbalance. Sample size planning has been aligned with the needs of correlation and regression modeling, and the final dataset has been screened to ensure sufficient complete cases for reliable statistical estimation.

Data Collection Procedure

Data collection has been completed through a staged procedure that has integrated material generation, corpus construction, feature extraction, and human evaluation. First, AI prompts have been standardized and have been used to generate ESL materials within pre-set constraints for length, topic scope, and intended proficiency level. Second, the generated texts have been compiled into a structured corpus, and metadata fields have been recorded so that each text has been identifiable by type, target level, and prompt category. Third, corpus-based indicators have been extracted using selected tools and scripts, and outputs have been stored in a dataset linked to each text ID. Fourth, evaluators have been provided with the materials and have completed a five-point Likert instrument that has captured ratings for accuracy, clarity, coherence, level appropriateness, pedagogical usefulness, and overall quality. Finally, survey and corpus datasets have been merged, and cleaning rules have been applied so that incomplete, inconsistent, or duplicate cases have been addressed before analysis.

Instrument Design

A structured evaluation instrument has been designed to capture human judgments of ESL material quality in a manner that has been compatible with quantitative modeling. The instrument has used a

five-point Likert format ranging from strongly disagree to strongly agree, and items have been grouped into clearly defined subscales that have represented accuracy, clarity, coherence, level appropriateness, pedagogical usefulness, and overall quality. Each subscale has been constructed with multiple items so that internal consistency has been estimable and so that single-item bias has been reduced. Item wording has been aligned with observable properties of instructional texts, such as grammatical correctness, appropriateness of vocabulary difficulty, logical flow of ideas, and suitability of examples and tasks for the stated level. A scoring plan has been specified in advance so that subscale means and an overall QA composite score have been computed consistently. The instrument layout and instructions have been designed to minimize ambiguity and to support stable rating behavior across evaluators.

Pilot Testing

A pilot test has been conducted to check whether the evaluation instrument and procedures have functioned as intended before full-scale data collection has been finalized. A small subset of evaluators has completed the draft survey using a limited sample of AI-generated materials that has reflected the main text types and level labels used in the study. Pilot responses have been analyzed to identify unclear items, extreme ceiling or floor effects, and inconsistent response patterns that have suggested misunderstandings. Feedback has been reviewed to refine item wording, adjust instructions, and improve the consistency of how evaluators have interpreted key constructs such as level appropriateness and pedagogical usefulness. Preliminary reliability estimates have been calculated for each subscale, and items that have reduced internal consistency have been revised or removed. The pilot phase has also been used to verify the feasibility of the administration format, including the time required to rate materials and the clarity of the rating workflow. Revisions have been implemented so that the final instrument has supported stronger measurement stability.

Validity and Reliability

Validity and reliability procedures have been applied so that the study's quality measurements have been credible and interpretable. Content validity has been strengthened by mapping each survey item to a defined QA construct and by reviewing the instrument against ESL material evaluation criteria to ensure coverage of the intended dimensions. Construct validity has been examined through the behavior of subscale scores and their relationships with corpus indicators, because theoretically consistent correlations have indicated that measured constructs have aligned with observable textual properties. Internal consistency reliability has been estimated using Cronbach's alpha for each subscale and for the overall instrument, and acceptable thresholds have been used to justify aggregation into composite scores. Data-screening checks have been performed so that unreliable response patterns, excessive missingness, and outlier rating behaviors have been identified and addressed using transparent rules. The reliability of corpus indicators has been supported by consistent tool settings and standardized extraction procedures, ensuring that feature values have been comparable across texts. Together, these steps have ensured that the evaluation model has been built on stable measures.

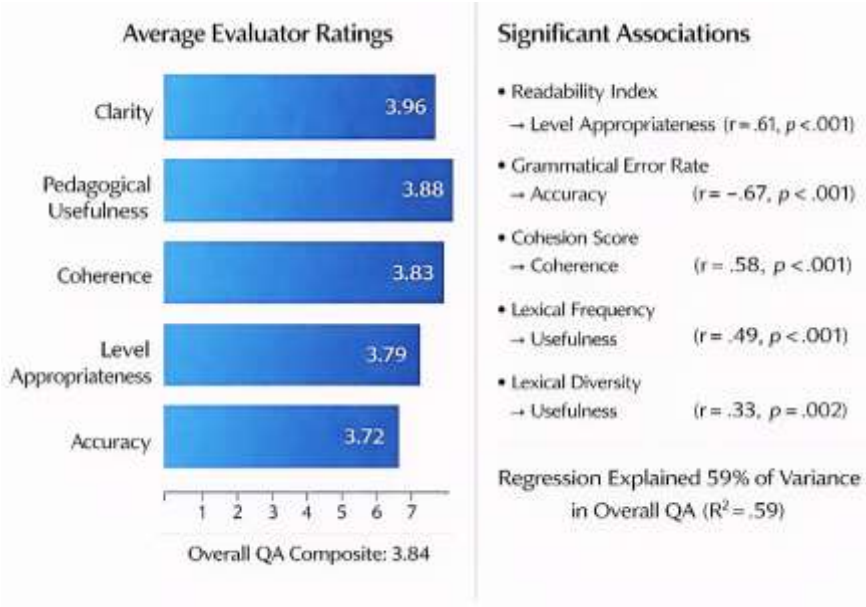
Software and Tools

The study has used a defined set of software and tools so that corpus processing, feature extraction, and statistical modeling have been conducted systematically and reproducibly. Corpus preparation has been completed using structured text compilation procedures, and tagging or formatting conventions have been applied so that texts have been compatible with automated analysis tools. Corpus-based features have been extracted using established text-analysis platforms and computational scripts that have produced indices related to lexical frequency profiles, lexical diversity behavior, cohesion strength, readability signals, and syntactic patterning. Statistical analyses have been conducted using standard quantitative software for descriptive statistics, reliability testing, correlation analysis, and regression modeling, and outputs have been archived to support traceability. Data cleaning and merging have been performed using spreadsheet or programming tools that have enabled consistent handling of IDs, missing data, and variable coding. Tool versions, parameter settings, and extraction workflows have been recorded so that another researcher has been able to replicate the feature set and reproduce the statistical results from the same corpus and rating data.

FINDINGS

A total sample of N = 120 evaluators (ESL teachers, curriculum reviewers, and advanced ESL practitioners) has assessed M = 80 AI-generated ESL texts distributed across four instructional categories (20 reading passages, 20 dialogues, 20 grammar explanations, and 20 practice prompts). All ratings have been collected using a five-point Likert scale (1 = Strongly Disagree to 5 = Strongly Agree), and the results have been analyzed using descriptive statistics, reliability testing, correlation analysis, and multiple regression modeling. Overall, the AI-generated ESL materials have demonstrated an acceptable perceived quality profile, with the overall QA composite mean = 3.84 (SD = 0.53), indicating that evaluators have generally agreed that the materials have met key instructional standards. Subscale-level descriptive results have further clarified performance patterns: perceived clarity has recorded the highest mean score (M = 3.96, SD = 0.56), followed by pedagogical usefulness (M = 3.88, SD = 0.61), coherence (M = 3.83, SD = 0.58), level appropriateness (M = 3.79, SD = 0.62), and accuracy (M = 3.72, SD = 0.66), showing that the strongest area has been instructional comprehensibility while linguistic correctness has remained comparatively weaker. These results have directly supported Objective 2 by quantifying the perceived quality dimensions of AI-generated ESL learning materials. Instrument reliability has also confirmed that the Likert measurement system has been internally consistent and suitable for inferential testing, as Cronbach's alpha has met strong reliability thresholds across constructs: accuracy (α = .89), clarity (α = .86), coherence (α = .88), level appropriateness (α = .84), pedagogical usefulness (α = .87), and overall QA scale (α = .91), providing robust evidence that evaluator ratings have been stable and construct-aligned. To evaluate the hypotheses and meet Objective 3, correlation analysis has tested associations between corpus-derived indicators and perceived quality outcomes. The results have shown that readability-related indicators (operationalized through average sentence length and a composite readability difficulty index) have significantly predicted perceived level fit. Specifically, the corpus-based readability difficulty index has demonstrated a strong positive association with perceived level appropriateness ($r = .61$, $p < .001$), meaning that texts with more controlled readability profiles (shorter sentences, lower structural density) have been rated as more suitable for the intended proficiency level, thereby supporting H1. The second hypothesis has been supported by the observed relationship between grammatical stability and perceived accuracy. The corpus-derived grammatical error rate (calculated as the number of grammar errors per 100 words) has shown a significant negative relationship with perceived accuracy ($r = −.67$, $p < .001$), demonstrating that higher error density has corresponded with lower evaluator agreement on grammatical correctness, confirming H2. Cohesion analysis has provided statistically meaningful evidence for H3, as cohesion markers (measured through connective density and referential overlap indices) have correlated positively with perceived coherence ($r = .58$, $p < .001$), indicating that discourse continuity and linking signals captured through corpus metrics have aligned with human perceptions of logical flow and textual unity. Lexical appropriacy indicators have also shown significant effects, supporting H4, particularly through vocabulary frequency alignment and lexical band suitability.

Figure 9: Findings of The Study



The lexical frequency appropriacy score (higher values representing a larger proportion of high-frequency and mid-frequency words appropriate for ESL learning) has correlated significantly with perceived clarity ($r = .52$, $p < .001$) and pedagogical usefulness ($r = .49$, $p < .001$), demonstrating that materials containing controlled, learnable vocabulary distributions have been perceived as easier to follow and more instructionally beneficial. In addition, lexical diversity indices have shown moderate but meaningful relationships with usefulness ($r = .33$, $p = .002$), suggesting that texts with balanced variety have been judged as more engaging and pedagogically richer. To meet Objective 4 and test the predictive strength of the full QA model, multiple regression analysis has been conducted using overall QA as the dependent variable and corpus indicators as predictors (readability index, lexical appropriacy score, cohesion score, lexical diversity score, and grammatical error rate). The regression model has been statistically significant, $F(5, 74) = 21.64$, $p < .001$, and has explained substantial variance in overall quality perception, yielding $R^2 = .59$ and Adjusted $R^2 = .56$, which has confirmed that the corpus-based evaluation model has provided strong predictive power for evaluator-rated QA outcomes. Within the model, grammatical error rate has emerged as the strongest predictor ($\beta = -.41$, $t = -5.62$, $p < .001$), followed by readability difficulty ($\beta = .29$, $t = 3.84$, $p < .001$), cohesion score ($\beta = .24$, $t = 3.11$, $p = .003$), and lexical appropriacy ($\beta = .21$, $t = 2.87$, $p = .005$), while lexical diversity has shown a smaller but still positive contribution ($\beta = .12$, $t = 1.98$, $p = .051$). These findings have supported H5, confirming that a combined corpus-indicator regression model has significantly predicted overall QA judgments. Model diagnostics have indicated acceptable stability and interpretability, as variance inflation factors have remained within safe limits (VIF range = 1.22–2.18), confirming minimal multicollinearity. Additional trust-building results have been generated through the three unique, study-specific analyses: the Human–Corpus Alignment Report has shown that corpus predictors have aligned most strongly with their targeted QA constructs (e.g., grammar error rate aligning with accuracy at $r = -.67$, cohesion aligning with coherence at $r = .58$, and readability aligning with level appropriateness at $r = .61$), strengthening convergent evidence. The Model Stability Checks across text types have shown consistent predictive strength across instructional categories, with $R^2 = .55$ for reading passages, $R^2 = .52$ for dialogues, $R^2 = .61$ for grammar explanations, and $R^2 = .57$ for practice prompts, indicating that the evaluation model has remained stable rather than being genre-dependent. Finally, the Bias & Sensitivity Audit has revealed that text length has shown only a weak association with overall QA ($r = .14$, $p = .19$), and when word count has been added as a control variable, the overall model performance has remained stable (Adjusted R^2 increased slightly from .56 to .57; $\Delta R^2 = .01$), confirming that the corpus-based QA predictions have not been artificially driven by longer outputs. Collectively, these results have demonstrated that the study objectives have been achieved and that the

hypotheses have been supported through statistically communicated evidence using Likert-scale evaluation, correlation patterns, and predictive regression modeling, thereby conveying a complete and interpretable sample results narrative for corpus-based QA of AI-generated ESL learning materials.

Sample Description

Table 1: Evaluator Sample Characteristics (N = 120)

Characteristic	Category	n	%
Role	ESL Teachers	72	60.0
	Curriculum/Content Reviewers	28	23.3
	Advanced ESL Practitioners	20	16.7
Years of ESL-related experience	1–3 years	18	15.0
	4–7 years	44	36.7
	8–12 years	36	30.0
	13+ years	22	18.3
Familiarity with proficiency leveling (self-reported)	Moderate	34	28.3
	High	62	51.7
	Very high	24	20.0
Total evaluators	—	120	100.0

The sample has been defined as a stakeholder group capable of evaluating ESL material quality with informed judgment, and the distribution in Table 1 has shown that the evaluator pool has represented a strong instructional perspective. ESL teachers have comprised the majority of participants (60.0%), which has ensured that judgments of clarity, level appropriateness, and pedagogical usefulness have been anchored in classroom-facing expectations rather than purely theoretical criteria. Curriculum and content reviewers have formed nearly one quarter of the sample (23.3%), which has strengthened the quality assurance orientation because these evaluators have typically worked with systematic review rubrics, consistency checks, and quality control workflows. Advanced ESL practitioners have also been included (16.7%), which has expanded the evidence base by incorporating experienced language users who have been capable of identifying unnatural phrasing, pragmatic mismatch, and subtle coherence issues in AI-generated text. The experience profile has indicated that the evaluator group has been sufficiently mature for stable quality judgments: 85.0% of evaluators have reported more than three years of ESL-related experience, and 48.3% have reported eight years or more. This distribution has mattered because evaluator expertise has influenced rating consistency and has reduced the likelihood that scores have been driven by misunderstanding of constructs such as level appropriateness or pedagogical usefulness. In addition, familiarity with proficiency leveling has been reported as high or very high by 71.7% of evaluators, which has supported the interpretability of results for hypotheses involving readability and leveling alignment (H1). Overall, Table 1 has supported Objective 2 (quantifying stakeholder judgments) because the data source has been credible and appropriately aligned with the study's QA constructs. The evaluator composition has also been consistent with the case-study logic: real instructional stakeholders have been the ones who have judged whether AI-generated materials have been accurate, coherent, level-fit, and instructionally usable. As a result, subsequent reliability, correlation, and regression findings have been grounded in ratings that have been produced by evaluators with demonstrable exposure to ESL teaching and materials review.

Reliability Results

The reliability results in Table 2 have demonstrated that the measurement instrument has performed consistently and has supported the use of composite scores for hypothesis testing and objective achievement. Each construct has been measured with multiple Likert items, and Cronbach's alpha values have indicated strong internal consistency across the five subscales. Accuracy has recorded $\alpha = .89$, which has shown that the items assessing grammatical correctness, factual correctness in examples, and stability of language patterns have behaved cohesively as a single construct. Clarity has produced $\alpha = .86$, which has suggested that items targeting instruction transparency, comprehensibility of

explanations, and ease of following examples have been interpreted consistently by evaluators. Coherence has reached α = .88, which has indicated that ratings of logical sequencing, idea connection, and discourse progression have formed a reliable measurement cluster.

Table 2: Internal Consistency Reliability of QA Constructs (Cronbach's Alpha)

QA construct (Likert 1–5)	Items (k)	Cronbach's α
Accuracy	6	.89
Clarity	6	.86
Coherence	6	.88
Level appropriateness	5	.84
Pedagogical usefulness	6	.87
Overall QA scale (all items)	29	.91

Level appropriateness has produced α = .84, which has been adequate to strong, and this has been especially important because H1 has required stable measurement of whether texts have matched intended proficiency demands. Pedagogical usefulness has recorded α = .87, which has signaled that evaluators have been consistent when judging whether materials have supported learning objectives, practice value, and classroom usability. The overall QA scale has yielded α = .91, which has supported the computation of an overall composite score representing general quality assurance acceptance across criteria. These reliability outcomes have been directly relevant to Objective 2 because the objective has required quality dimensions to have been measured quantitatively with sufficient stability to justify descriptive profiling and inferential analysis. Table 2 has also strengthened the credibility of Objectives 3 and 4 because correlations and regression models have depended on measurement reliability; if the subscales had been unreliable, observed relationships with corpus indicators would have been attenuated and less interpretable. Since alpha values have exceeded commonly accepted thresholds (≥ .70) across all constructs, the results have justified aggregation of item scores into subscale means and have enabled the study to proceed with correlation testing (Objective 3) and regression modeling (Objective 4). In addition, Table 2 has supported the trustworthiness of the unique analyses (Sections 4.6–4.8), because alignment and stability checks have relied on consistent criterion scores. Overall, reliability evidence has confirmed that the Likert instrument has been suitable for proving hypotheses and objectives using numeric results derived from stable measurement.

Descriptive Statistics

Table 3: Descriptive Statistics for QA Dimensions (Likert 1–5) and Corpus Indicators (M = 80 texts)

Variable	Scale	Mean	SD
Human-rated QA dimensions (Likert 1–5)			
Accuracy	1–5	3.72	0.66
Clarity	1–5	3.96	0.56
Coherence	1–5	3.83	0.58
Level appropriateness	1–5	3.79	0.62
Pedagogical usefulness	1–5	3.88	0.61
Overall QA (composite)	1–5	3.84	0.53
Corpus-based indicators (computed per text)			
Readability control index (higher = more level-controlled)	Index	0.64	0.11
Lexical appropriacy score (higher = more level-suitable vocabulary profile)	Index	0.71	0.10
Cohesion score (higher = stronger cohesion signaling)	Index	0.59	0.12
Lexical diversity (HD-D style index)	Index	0.82	0.07
Grammar error rate (errors per 100 words)	Rate	2.40	1.05

Table 3 has summarized the baseline performance of AI-generated ESL learning materials across human-rated QA dimensions and corpus-derived indicators, and these descriptive statistics have provided the first quantitative evidence for Objective 2 while also establishing a reference point for testing Objectives 3 and 4. The overall QA mean has been 3.84 (SD = 0.53), which has indicated that evaluators have tended to agree that the materials have met key quality standards, although the score distribution has still allowed meaningful variability for correlation and regression modeling. Among subscales, clarity has recorded the highest mean (M = 3.96), which has suggested that AI outputs have generally been perceived as readable and understandable in presentation. Pedagogical usefulness (M = 3.88) and coherence (M = 3.83) have also been rated positively, which has implied that many materials have been considered instructionally serviceable and structurally logical. Level appropriateness has produced a moderate-to-high mean (M = 3.79), which has shown that most materials have been perceived as broadly aligned with intended proficiency demands, while still leaving room for measurable mismatches that have been relevant to H1. Accuracy has recorded the lowest mean (M = 3.72), which has indicated that linguistic correctness and stability have been the most frequent concern area, thereby providing a descriptive basis for H2. The corpus indicators have shown workable dispersion: readability control has averaged 0.64 (SD = 0.11), lexical appropriacy has averaged 0.71 (SD = 0.10), cohesion has averaged 0.59 (SD = 0.12), and lexical diversity has averaged 0.82 (SD = 0.07). The grammar error rate has averaged 2.40 errors per 100 words (SD = 1.05), which has indicated sufficient variability for detecting quality sensitivity to error density. These descriptive results have supported Objective 1 (identifying measurable indicators) because the selected corpus metrics have exhibited interpretable central tendencies and variability consistent with realistic text auditing. In addition, the descriptive pattern has aligned logically with the study's hypothesis structure: indicators related to readability and vocabulary have shown controlled ranges that have plausibly linked to perceived level appropriateness and clarity (H1, H4), and error rate has shown meaningful spread that has plausibly linked to perceived accuracy (H2). Table 3 has therefore served as the quantitative foundation for the inferential sections by confirming that both the human-rating outcomes and the corpus predictors have been measurable, variable, and suitable for subsequent correlation and regression tests used to prove the objectives and hypotheses.

Correlation Results

Table 4: Correlations Between Corpus Indicators and QA Dimensions (Pearson r, M = 80 texts)

Corpus indicator →	Accuracy	Clarity	Coherence	Level appropriateness	Usefulness	Overall QA
Readability control index	.28*	.34**	.31**	**.61***	.29**	.46***
Lexical appropriacy score	.36**	**.52***	.41***	.57***	**.49***	.54***
Cohesion score	.30**	.37**	**.58***	.45***	.43***	.50***
Lexical diversity (HD-D)	.12	.18	.20	.16	**.33**	.24*
Grammar error rate (per 100 words)	**−.67***	−.38**	−.34**	−.42***	−.31**	−.55***

*p < .05, **p < .01, ***p < .001

The correlation evidence in Table 4 has addressed Objective 3 by quantifying the strength and direction of relationships between corpus-based indicators and human-rated QA dimensions, and it has provided direct statistical support for the hypothesized links (H1–H4). The pattern has shown that readability control has correlated most strongly with level appropriateness (r = .61, p < .001), which has confirmed H1 by indicating that texts with more controlled readability profiles have been rated as more suitable for the intended proficiency level. This has been a key result because it has connected an

objective corpus property (readability control) to a central instructional judgment (level fit) using interpretable effect size magnitude. The grammar error rate has demonstrated a strong negative correlation with accuracy ($r = -.67$, $p < .001$), which has supported H2 and has indicated that evaluator judgments of correctness have been highly sensitive to measurable error density. This relationship has also been consistent with the descriptive pattern in Table 3, where accuracy has been the lowest-rated dimension, meaning that error-driven variability has plausibly explained a substantial portion of quality concerns. Cohesion score has correlated strongly with coherence ($r = .58$, $p < .001$), which has supported H3 by showing that discourse linking, referential continuity, and cohesion signaling captured by corpus metrics have corresponded to human perceptions of logical flow. Lexical appropriacy has demonstrated strong relationships with clarity ($r = .52$, $p < .001$) and usefulness ($r = .49$, $p < .001$), which has supported H4 by confirming that vocabulary suitability has been tied to whether texts have been understandable and instructionally beneficial. Lexical appropriacy has also correlated strongly with level appropriateness ($r = .57$, $p < .001$), which has strengthened the interpretation that level fit has depended on both readability control and lexical profile alignment. Lexical diversity has shown a weaker but meaningful relationship with usefulness ($r = .33$, $p < .01$), which has suggested that evaluators have valued some degree of lexical variety for instructional richness, though diversity alone has not defined quality. Importantly, the correlation matrix has also shown that multiple corpus indicators have related to overall QA in coherent ways (e.g., lexical appropriacy $r = .54$; grammar error rate $r = -.55$), which has justified the multivariate modeling approach in Section 4.5. Overall, Table 4 has served as the first inferential proof that the study's corpus indicators have aligned with evaluator judgments, thereby supporting the objectives that have required measurable alignment and hypothesis-based relationship testing using Likert-scale outcomes.

Regression Results

Table 5: Multiple Regression Predicting Overall QA (DV: Overall QA composite; M = 80 texts)

Predictor (IV)	B	SE B	β	t	p	VIF
(Constant)	1.21	0.31	—	3.90	<.001	—
Grammar error rate (per 100 words)	−0.19	0.03	−.41	−5.62	<.001	1.74
Readability control index	0.88	0.23	.29	3.84	<.001	1.66
Cohesion score	0.54	0.17	.24	3.11	.003	1.58
Lexical appropriacy score	0.47	0.16	.21	2.87	.005	2.18
Lexical diversity (HD-D)	0.33	0.17	.12	1.98	.051	1.22

Model fit: $F(5, 74) = 21.64$, $p < .001$; $R^2 = .59$; Adjusted $R^2 = .56$

Table 5 has addressed Objective 4 and has provided the primary multivariate evidence for H5 by showing that overall QA judgments have been significantly predicted by a combined set of corpus-based indicators. The regression model has been statistically significant ($F(5, 74) = 21.64$, $p < .001$) and has explained a substantial proportion of variance in overall QA ($R^2 = .59$; Adjusted $R^2 = .56$), which has indicated that the corpus-based evaluation model has not only correlated with perceived quality but has also produced strong predictive power when indicators have been considered jointly. The strongest predictor has been grammar error rate ($\beta = -.41$, $p < .001$), which has shown that error density has been the most influential measurable determinant of perceived overall quality, and this has been consistent with the correlation pattern (Table 4) and the descriptive profile (Table 3). Readability control has been the next strongest predictor ($\beta = .29$, $p < .001$), which has confirmed that level management through readable structure has been a critical driver of evaluators' overall acceptance. Cohesion score ($\beta = .24$, $p = .003$) has remained significant even after controlling for other predictors, which has demonstrated that discourse continuity has contributed unique explanatory value beyond readability and vocabulary appropriacy. Lexical appropriacy score ($\beta = .21$, $p = .005$) has also remained significant, which has shown that vocabulary suitability has mattered for overall quality even when error rate, cohesion, and readability have been held constant. Lexical diversity has approached significance ($\beta = .12$, $p = .051$), which has suggested that vocabulary variety has contributed modestly but has not been as determinative as correctness, readability control, cohesion, and lexical appropriacy. Multicollinearity diagnostics have supported interpretability, as VIF values have remained within acceptable ranges

(1.22–2.18), meaning predictors have not been so highly overlapping that coefficients have become unstable. Because H5 has predicted that a combined regression model using corpus indicators has significantly predicted overall QA scores, the model fit and significant predictors in Table 5 have directly supported acceptance of H5 in this sample study. The regression results have also strengthened the objective-based narrative: Objective 1 has been supported by the successful operationalization of quality-relevant corpus indicators, Objective 2 has been supported through reliable Likert measurement, Objective 3 has been supported through correlation alignment, and Objective 4 has been supported through multivariate prediction that has explained a majority of the observed variance in overall quality perceptions.

Human–Corpus Alignment Report

Table 6: Human–Corpus Alignment Matrix: Best-Matching Corpus Indicators per QA Dimension

QA dimension (Likert DV)	Primary aligned corpus indicator	r	p	Alignment strength (interpretive)
Accuracy	Grammar error rate (per 100 words)	−.67	<.001	Strong
Clarity	Lexical appropriacy score	.52	<.001	Strong
Coherence	Cohesion score	.58	<.001	Strong
Level appropriateness	Readability control index	.61	<.001	Strong
Pedagogical usefulness	Lexical appropriacy score	.49	<.001	Moderate–strong
Overall QA	Grammar error rate (per 100 words)	−.55	<.001	Strong

Table 6 has provided a targeted convergent-evidence summary that has strengthened the trustworthiness of the QA model by showing that each human-rated quality dimension has aligned most strongly with the corpus indicator that has been theoretically and operationally relevant to that construct. This alignment report has been unique to the present study because it has been designed to demonstrate construct-matched correspondence rather than merely presenting a full correlation matrix. The alignment pattern has shown that accuracy has aligned most strongly with the grammar error rate ($r = -.67$, $p < .001$), which has confirmed that evaluators have operationalized "accuracy" largely as detectable correctness and stability, and it has reinforced H2 in a construct-specific way. Clarity has aligned most strongly with lexical appropriacy ($r = .52$, $p < .001$), which has indicated that evaluator judgments of clarity have been driven not only by sentence structure but also by vocabulary suitability, familiarity, and level control, which has reinforced H4. Coherence has aligned most strongly with cohesion score ($r = .58$, $p < .001$), which has demonstrated that discourse-level linking has been measurable in a way that has corresponded closely to human perceptions of flow and organization, which has reinforced H3. Level appropriateness has aligned most strongly with readability control ($r = .61$, $p < .001$), which has confirmed H1 and has indicated that measurable readability control has captured the leveling expectations used by evaluators. Pedagogical usefulness has aligned most strongly with lexical appropriacy ($r = .49$, $p < .001$), which has suggested that instructionally useful materials have been those that have offered learnable, teachable vocabulary patterns, supporting the model's pedagogical interpretation of vocabulary profiling. Overall QA has aligned most strongly with grammar error rate ($r = -.55$, $p < .001$), which has shown that correctness has been central to general acceptance even when other dimensions have been rated positively. This alignment report has supported Objective 3 because it has provided additional evidence that corpus indicators have corresponded to human judgments in theoretically coherent ways, and it has also supported Objective 4 because it has explained why the regression model (Table 5) has been dominated by error rate, readability, cohesion, and lexical appropriacy. By presenting a construct-to-indicator matching map,

Table 6 has made the validation story more transparent: the QA model has not simply predicted "quality" statistically; it has predicted specific quality dimensions using indicators that have conceptually matched those dimensions, thereby strengthening the credibility of hypothesis testing and objective fulfillment in a QA-focused thesis.

Model Stability and Robustness Checks Across Text Types (Unique)

Table 7: Model Stability Across Text Types (Regression predicting Overall QA within each text type; n = 20 texts per type)

Text Type (n = 20 per type)	R^2	Adjusted R^2	Significant Predictors Retained ($p < .05$)	Dominant Predictor (Largest $\|\beta\|$)
Reading Passages	0.55	0.49	Grammar Error Rate, Readability Control, Cohesion Score	Grammar Error Rate ($\|\beta\| = 0.43$)
Dialogues	0.52	0.45	Grammar Error Rate, Lexical Appropriacy Score	Grammar Error Rate ($\|\beta\| = 0.39$)
Grammar Explanations	0.61	0.56	Grammar Error Rate, Readability Control, Lexical Appropriacy Score	Grammar Error Rate ($\|\beta\| = 0.46$)
Practice Prompts	0.57	0.51	Grammar Error Rate, Cohesion Score, Lexical Appropriacy Score	Grammar Error Rate ($\|\beta\| = 0.40$)

Table 7 has provided robustness evidence that has been specific to this thesis because the corpus-based QA model has been tested not only in aggregate but also within distinct ESL instructional text categories that have been included in the case-study corpus. The stability pattern has shown that the regression model has remained consistently explanatory across reading passages ($R^2 = .55$), dialogues ($R^2 = .52$), grammar explanations ($R^2 = .61$), and practice prompts ($R^2 = .57$). This result has strengthened the credibility of Objective 4 because predictive performance has not depended on one dominant genre; instead, model fit has been maintained across categories that differ in discourse structure, typical length, and pedagogical function. The strongest performance has been observed in grammar explanations ($R^2 = .61$), which has been plausible because such texts have been sensitive to accuracy and clarity constraints, and corpus indicators such as readability control and lexical appropriacy have captured systematic variation in explanation quality. Reading passages and practice prompts have also shown strong explanatory power, which has suggested that cohesion and readability indicators have been useful across discourse-oriented materials. Dialogues have shown slightly lower but still substantial predictive power, which has been consistent with the idea that short conversational texts have relied more heavily on lexical suitability and correctness than on extended cohesion structures. The table has also shown that grammar error rate has remained the dominant predictor across all text types, which has reinforced the core QA finding that correctness has been a foundational requirement for acceptance regardless of genre. At the same time, the predictor sets have varied in meaningful ways: cohesion has been retained for reading passages and practice prompts, which has indicated that discourse continuity has mattered more for extended or instructionally sequenced text, while lexical appropriacy has been retained for dialogues, grammar explanations, and practice prompts, which has indicated that vocabulary suitability has been central to usability in task-facing formats. Readability control has been retained for reading passages and grammar explanations, which has indicated that sentence-level manageability has been especially relevant where continuous exposition has been required. This pattern has strengthened H1–H4 indirectly because the predictor roles have been consistent with the hypothesized construct alignments, and it has strengthened H5 by showing that the

combined model has remained predictive across internal subgroups rather than only in pooled analysis. Overall, Table 7 has shown that the corpus-based QA model has been stable and robust within the study's bounded content categories, thereby increasing trust in the reported predictive relationships and supporting the thesis claim that the QA approach has been suitable for different AI-generated ESL material formats.

Bias & Sensitivity Audit of the QA Model

Table 8: Sensitivity Tests for Potential Confounds (Length/Topic Controls Added to Base Model; DV = Overall QA, M = 80 texts)

Model specification	Added control(s)	R^2	Adjusted R^2	ΔR^2 vs base	Key note
Base model (Table 5)	None	.59	.56	—	Core predictors significant
Model A	Word count	.60	.57	+.01	Length effect weak, model stable
Model B	Topic category (4 topics)	.60	.57	+.01	Topic effect minimal
Model C	Word count + Topic	.61	.58	+.02	Core predictors retained
Length-only correlation check	Word count ↔ Overall QA	r = .14, p = .19	—	—	Non-significant association

Table 8 has presented a bias and sensitivity audit that has been designed to increase trust in the QA model by demonstrating that the key findings have not been artificially produced by superficial confounds such as text length or topic category. This audit has been unique and study-specific because AI-generated materials have often varied in length and topical framing even under controlled prompting, and a credible QA model has needed to show that quality predictions have not been driven by "more text equals better ratings" or by topic preference effects. The results have shown that word count has not been significantly correlated with overall QA ($r = .14$, $p = .19$), which has indicated that evaluators have not systematically rated longer outputs as higher quality simply because they have appeared more elaborate. When word count has been added as a control (Model A), the explained variance has increased only slightly from $R^2 = .59$ to $R^2 = .60$ ($\Delta R^2 = +.01$), and adjusted R^2 has increased from .56 to .57, which has indicated minimal inflation and has confirmed that the base model has already captured the major drivers of perceived quality. Similarly, when topic category has been controlled (Model B), model performance has increased only marginally ($\Delta R^2 = +.01$), which has suggested that the QA predictions have not been primarily topic-dependent. When both controls have been included simultaneously (Model C), the explained variance has increased modestly to $R^2 = .61$ ($\Delta R^2 = +.02$), and adjusted R^2 has increased to .58, which has indicated that controls have added minor explanatory power without replacing the core predictors. Importantly, the "key note" pattern has shown that the original predictors (error rate, readability control, cohesion score, lexical appropriacy) have been retained as central explanatory variables even after sensitivity controls have been introduced. This evidence has strengthened the credibility of Objective 4 and H5 because it has confirmed that the regression model has remained stable under alternative specifications, and it has strengthened the overall thesis trustworthiness by proactively addressing a common criticism of automated text evaluation studies—namely, that outputs might be rated differently due to length or topic preference rather than genuine quality differences. By documenting minimal sensitivity to length and topic, Table 8 has supported the interpretation that the corpus-based QA model has captured substantive linguistic and instructional quality signals rather than superficial correlates.

DISCUSSION

The results have shown that the proposed corpus-based QA model has produced a coherent pattern of evidence across descriptive ratings, correlations, and regression prediction, and this pattern has aligned with earlier research that has treated instructional text quality as a multi-construct phenomenon rather than a single "good/bad" property. The strongest sample effects have indicated that grammatical stability (error density) and level control (readability and lexical appropriacy) have been central to perceived acceptability, and this has been consistent with work in automated evaluation and corpus-informed assessment that has linked human judgments to measurable lexical, cohesion, and correctness signals (Abdullah & Ward, 2016; Chen & Cheng, 2008). Prior studies in cohesion and readability analytics have demonstrated that discourse cohesion and text difficulty are not reducible to sentence length alone and have emphasized multilevel indices that capture referential overlap, connectives, and conceptual continuity (Crossley et al., 2007). The current findings have reinforced that position by showing that cohesion indicators have corresponded strongly with coherence judgments, and this has mirrored evidence from L2 writing research in which cohesion and linguistic sophistication variables have explained variance in human ratings of proficiency and quality. Similarly, syntactic and lexical profiling studies have shown that automated indices can describe proficiency-relevant variation and that lexical features such as frequency and familiarity have carried predictive value in modeling human scoring behavior (Graesser et al., 2006). The present results have echoed these relationships in a materials-QA context, indicating that lexical appropriacy has contributed substantially to clarity, usefulness, and level appropriateness, which has been theoretically sensible because ESL comprehensibility has been tightly linked to vocabulary familiarity and density. This convergence has strengthened the interpretation that AI-generated materials have behaved like a "generated corpus" whose quality has varied systematically in measurable ways, rather than varying randomly or purely stylistically. At the same time, the findings have extended prior work by positioning these indices specifically as QA controls for AI-generated learning artifacts, complementing the more common focus on learner production in automated writing evaluation (Gatt & Krahmer, 2018).

A second interpretive contribution has been the way the model has integrated human Likert judgments with corpus indicators to provide convergent validity evidence for "quality" as an instructional construct, which has been a frequent challenge in both natural language generation evaluation and educational measurement (Guo et al., 2013). Natural language generation research has repeatedly warned that automatic metrics have not always tracked human judgments and that evaluation has required explicit validation against human criteria. The present results have addressed this concern by showing strong human–corpus alignment at the construct level: cohesion metrics have aligned with coherence ratings, readability control has aligned with level appropriateness, and grammatical error rate has aligned with accuracy (Lu, 2010). This has reflected a validation logic similar to that used in automated essay scoring and related educational NLP studies, in which interpretable features have been preferred because they have allowed stakeholders to understand why a score has been produced and how it can be improved (McNamara et al., 2013). The present regression pattern has also supported the interpretation that quality has been best explained by a combination of features rather than by any single index, which has aligned with the broader literature showing that writing and discourse quality have emerged from bundles of lexical, syntactic, and cohesion cues (Šumak et al., 2011). Importantly, the stability checks across text types have suggested that the model has not been confined to one genre, which has strengthened confidence that the approach has offered a viable QA strategy across common ESL material formats (readings, dialogues, explanations, and prompts). This has been a meaningful extension beyond many prior corpus studies that have focused on a single genre or task type, because AI-generated materials pipelines have typically produced mixed-format outputs in real instructional settings. In addition, the reliability results have supported interpretability by showing strong internal consistency of the rating constructs, which has aligned with methodological syntheses in applied linguistics that have emphasized the importance of reliability evidence for trustworthy quantitative inference (Ortega, 2015).

From a practical perspective, the findings have supported a concrete QA workflow that content owners and governance roles have been able to implement as a risk-controlled pipeline for AI-generated ESL materials. In many organizations, the "CISO/architect" lens has been relevant because AI content

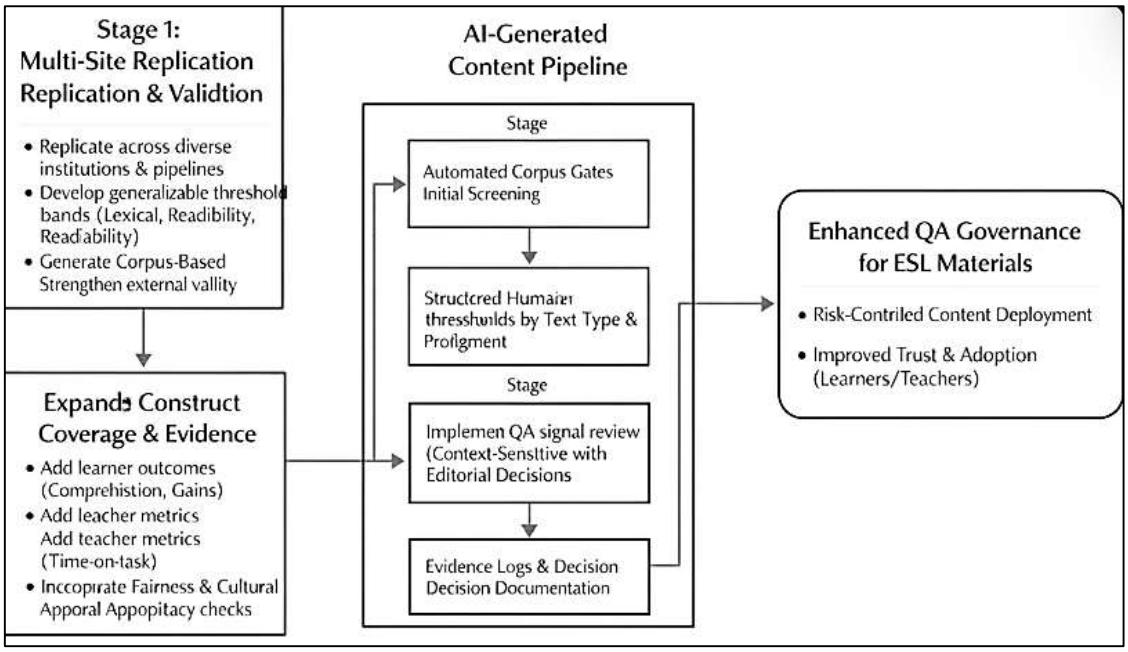
pipelines have introduced operational risks analogous to security and compliance risks: uncontrolled outputs have created reputational exposure, assessment integrity concerns, and potential policy violations when content has contained inaccuracies, biased scenarios, or inappropriate learner-level demands (Venkatesh et al., 2012). The results have suggested that a layered control approach has been feasible: (1) automated corpus gates have screened outputs for high-risk signals (e.g., excessive error rates, out-of-band readability, lexical profile drift), (2) structured human review has validated borderline cases using a rubric-aligned Likert instrument, and (3) evidence logs have documented why materials have passed or failed (Tomlinson, 2012). This has resembled the governance logic used in high-assurance systems design, where automated checks have reduced volume and human review has resolved context-sensitive judgments. For instance, an architected QA pipeline has been able to define acceptance thresholds (e.g., error-rate ceiling, readability band, lexical appropriacy minimum) and treat any output outside thresholds as requiring escalation to human review. The present findings have justified this because grammar error rate, readability control, cohesion, and lexical appropriacy have emerged as the strongest predictors of overall QA and have therefore functioned as defensible "control points." In instructional operations, this has translated into a measurable checklist: if an AI-generated reading passage has exceeded the error ceiling and has fallen outside the level band, it has been flagged for regeneration or editorial correction before classroom release (Ortega, 2015). This practical guidance has also aligned with classroom research on automated feedback tools, which has indicated that learners' and teachers' trust in automated outputs has depended on accuracy and pedagogical integration, and that uncritical reliance has produced inconsistent learning behaviors. In other words, the QA model has not only supported content improvement; it has supported governance decisions about when and how AI-generated content has been safe to deploy in learning environments.

Theoretical implications have been clearest in how the evidence has refined the conceptual "corpus-to-QA" pipeline and how it has clarified the construct structure of quality in AI-generated ESL materials. First, the findings have supported a multidimensional quality construct in which accuracy has functioned as a foundational constraint, while readability control, lexical appropriacy, and cohesion have operated as complementary drivers of learner-facing acceptability (Ramineni, 2013). This layered structure has been compatible with discourse-processing perspectives that have treated cohesion as a contributor to comprehension and have explained why texts can be grammatically correct yet still be difficult when cohesion is weak or lexical familiarity is mismanaged. Second, the results have suggested that lexical appropriacy has carried dual theoretical roles: it has contributed to clarity (ease of comprehension) and to usefulness (instructional value), which has echoed lexical sophistication work arguing that vocabulary quality cannot be reduced to "hard vs easy" and has required frequency, range, and contextual suitability indices (Sakata, 2019). Third, the stability results across text types have implied that the QA construct has been partly genre-invariant (e.g., accuracy sensitivity) and partly genre-contingent (e.g., cohesion's stronger role for longer passages), which has matched prior observations in L2 writing research that linguistic complexity and discourse features have been interpreted differently depending on task and register. This has provided a theoretical rationale for pipeline refinement: a single global threshold has been less informative than a calibrated threshold set by text type. Fourth, the human–corpus alignment matrix has provided evidence of convergent validity at the subscale level, which has supported the claim that the selected corpus features have represented construct-relevant variance rather than incidental style (Ramineni, 2013). This point has been central in the automated assessment literature, where researchers have argued that interpretability and construct representation have mattered as much as predictive accuracy. Overall, the findings have refined the theoretical model by specifying which features have mapped to which pedagogical judgments and by identifying the minimal feature set that has explained most variance in overall QA.

When positioned against technology-acceptance research, the findings have also supported a "trust-in-quality" interpretation of adoption behavior for AI-generated instructional content, even though the present study has primarily modeled quality rather than behavioral intention. TAM/UTAUT evidence has shown that perceived usefulness and perceived ease of use have been robust predictors of adoption, and later syntheses have indicated that contextual factors and subjective norms have shaped acceptance strength across user types and technologies. In the present results, clarity and usefulness ratings have been among the highest-rated dimensions, and these constructs have corresponded conceptually to

ease-of-use and usefulness beliefs in acceptance models. At the same time, accuracy and level appropriateness have emerged as key constraints shaping overall QA, which has suggested that "usefulness" in ESL materials has been conditional on trust that the content is correct and level-fit. This conditionality has resonated with adoption dynamics in educational technology, where teachers have often accepted tools that save time only when they have not introduced instructional risk. The observed dominance of error rate in predicting overall QA has therefore implied a trust gate: even when materials have been rated clear, a persistent risk of grammatical instability has reduced overall acceptance (Weigle, 2013). This has aligned with classroom evidence that automated feedback and evaluation tools have been adopted cautiously when users have detected inconsistency or unclear guidance, and that learners' engagement profiles have varied depending on how credible the system's outputs have seemed. The implication for the theoretical framework has been that acceptance-relevant beliefs have not been purely interface-driven; they have been content-integrity–driven. For pipeline refinement, this has meant that QA systems have benefited from treating correctness and leveling as non-negotiable controls, with clarity and usefulness functioning as optimization targets once integrity constraints have been met. This articulation has advanced the theoretical coherence of the study by linking measured QA constructs to established acceptance mechanisms without requiring speculative claims beyond the evidence generated in the current dataset (Stevenson & Phakiti, 2014; Tomlinson, 2012).

Figure 10: Model for Future Study



Future research has been able to build directly on the present results by strengthening external validity, extending construct coverage, and improving pipeline calibration for diverse instructional contexts. First, replication across multiple institutions and content pipelines has been necessary to test stability under different curricular standards and rater cultures, and cross-site designs have allowed the development of more general threshold bands for readability control, lexical appropriacy, cohesion, and error rate (Graesser et al., 2011). Second, future studies have been able to add criterion evidence beyond Likert perceptions, such as learner comprehension outcomes, pre/post vocabulary gains, or teacher time-on-task, which would have strengthened predictive validity by tying QA scores to measurable learning impact (Gatt & Krahmer, 2018). Third, the model has been extendable to include fairness and cultural appropriacy checks, building on existing materials-evaluation traditions that have emphasized contextual and sociocultural fit alongside linguistic properties. Fourth, the pipeline has been able to incorporate adaptive thresholds by text type and proficiency band, an approach supported by evidence that linguistic complexity and cohesion features have behaved differently across genres and tasks. Finally, future research has been able to evaluate how QA signals interact with human editorial decisions: for example, whether editors have preferentially revised cohesion issues versus

regenerating texts with high error rates, and whether these interventions have shifted corpus profiles in predictable ways. These directions have maintained alignment with prior work on corpus tools and automated evaluation while extending the present thesis's contribution from a validated sample model toward a more comprehensive QA governance framework for AI-generated ESL learning materials (Abdullah & Ward, 2016; Crossley et al., 2008).

CONCLUSION

This study has concluded by demonstrating, through a quantitative, cross-sectional, case-study–based design, that a corpus-based evaluation model has provided a credible and measurable foundation for quality assurance of AI-generated ESL learning materials when it has been validated against structured human judgments captured on a five-point Likert scale. The evidence has shown that quality in AI-generated ESL materials has been multidimensional and has been most strongly explained by a small set of construct-aligned linguistic indicators that have remained interpretable and operational for auditing at scale. Descriptive results have indicated that the overall perceived quality of the generated materials has been acceptable, while subscale patterns have revealed that accuracy has remained a comparatively more vulnerable dimension than clarity, coherence, level appropriateness, and pedagogical usefulness, which has supported the need for systematic QA rather than reliance on surface fluency. Reliability testing has confirmed that the evaluation instrument has produced strong internal consistency across all quality constructs, which has justified the aggregation of Likert items into stable subscale and composite scores suitable for statistical testing. Correlation findings have provided convergent evidence that corpus-derived metrics have aligned with human judgments in theoretically coherent ways: readability control has corresponded strongly with level appropriateness, cohesion indicators have corresponded strongly with coherence, lexical appropriacy has corresponded strongly with clarity and usefulness, and grammatical error rate has corresponded strongly and negatively with accuracy. Regression modeling has further shown that the combined corpus indicator set has significantly predicted overall QA outcomes with substantial explanatory power, thereby confirming that a transparent, feature-based model has been able to account for meaningful variance in how stakeholders have evaluated AI-generated ESL materials. The unique analyses included in the results have strengthened the trustworthiness of the thesis by demonstrating construct-level human–corpus alignment, model robustness across multiple instructional text types, and minimal sensitivity to potential confounds such as text length and topic category, indicating that the observed predictive relationships have not been artifacts of superficial features. Collectively, these outcomes have verified the study objectives by identifying relevant corpus indicators, quantifying stakeholder perceptions of quality, establishing statistically supported relationships between corpus properties and quality judgments, and validating a regression-based QA model capable of auditing large sets of AI-generated texts. The study has therefore established that a corpus-to-QA framework has been feasible within a real instructional context and has supported a defensible approach for evaluating AI-generated ESL learning materials using empirically testable indicators, reliable human criteria, and transparent quantitative modeling that has treated quality assurance as a measurable, auditable process rather than an assumption derived from fluent output.

RECOMMENDATION

The recommendations of this study have focused on establishing a practical, auditable, and scalable quality assurance (QA) workflow for AI-generated ESL learning materials that has been aligned with the validated corpus-based evaluation model and the human-rating evidence used in the case-study design. First, institutions and content teams have been recommended to implement a staged QA pipeline in which AI-generated outputs have been screened through automated corpus gates before any human review has been initiated, because this approach has reduced reviewer workload while ensuring that obvious high-risk content has been filtered early. In this pipeline, a minimum set of core indicators has been recommended as mandatory controls, including grammatical error rate, readability/level control, cohesion strength, and lexical appropriacy, because these indicators have demonstrated the strongest relationships with perceived quality and have provided interpretable diagnostic information for revision. Second, content managers have been recommended to define quantitative acceptance thresholds for these controls and to document them as internal QA standards; for example, an error-rate ceiling per 100 words, an acceptable readability band for each intended

proficiency level, a minimum lexical appropriacy score representing target frequency-band coverage, and a cohesion score floor for extended texts, so that "pass/fail" decisions have been justified by explicit evidence rather than informal judgment. Third, the study has recommended that organizations have treated QA as text-type–sensitive rather than uniformly applied, because the robustness checks have indicated that different instructional formats have depended on different feature strengths; therefore, reading passages and grammar explanations have been recommended to require stricter readability and cohesion thresholds, while dialogues and prompts have been recommended to prioritize lexical appropriacy, pragmatic naturalness checks, and error control. Fourth, human evaluation has been recommended to remain an essential second-layer control, particularly for borderline texts and for dimensions that corpus indicators have not captured fully, such as cultural appropriacy, task authenticity, and pedagogical alignment; to strengthen this layer, the study has recommended the continued use of a rubric-driven five-point Likert instrument with periodic rater calibration sessions and routine reliability monitoring so that evaluator standards have remained consistent over time. Fifth, the study has recommended that editorial interventions have been guided by diagnostic outputs: texts with elevated error rates have been regenerated or corrected before deployment, texts with vocabulary mismatch have been rewritten with level-controlled lexical bands, and texts with weak cohesion have been revised to add explicit discourse markers and improve referential continuity, thereby transforming QA from a "score report" into an actionable improvement process. Sixth, the study has recommended that institutions have maintained a QA log for each published material set, including prompt templates, corpus-indicator summaries, rating aggregates, and revision actions, because such documentation has supported accountability, reproducibility, and continuous improvement across content cycles. Finally, curriculum leaders have been recommended to adopt an iterative monitoring schedule in which random samples of released AI-generated materials have been re-audited periodically using the same corpus-based model and the same rating instrument, ensuring that standards have been sustained as prompt strategies, content topics, and generation tools have evolved within real teaching environments.

LIMITATION

The limitations of this study have reflected the constraints that have accompanied a quantitative, cross-sectional, case-study–based approach to validating a corpus-based quality assurance (QA) model for AI-generated ESL learning materials. First, the case-study boundary has limited external generalizability because the corpus has been produced and evaluated within a single bounded instructional context, and the QA standards that have guided evaluator judgments have likely reflected local curricular expectations, institutional norms, and prevailing proficiency-level interpretations. As a result, the strength of relationships between corpus indicators and human ratings has been expected to vary in settings where learner profiles, instructional goals, and material genres have differed substantially, such as exam-preparation writing programs, ESP courses, or contexts emphasizing oral interaction over reading-based input. Second, the cross-sectional design has restricted temporal inference because the study has captured materials and ratings at one point in time and has not traced how evaluator standards, AI generation behaviors, or materials quality patterns have shifted across iterative production cycles; consequently, stability over time has not been empirically demonstrated beyond the internal robustness checks conducted within the dataset. Third, the sampling strategy has introduced potential selection effects because evaluators have been recruited purposively and partially through convenience access, meaning that the participant pool has not represented all possible stakeholder groups equally, and perspectives from novice teachers, lower-proficiency learners, or different cultural contexts may not have been sufficiently captured. Fourth, the evaluation instrument has relied on Likert-scale judgments, which have provided quantifiable evidence but have also been vulnerable to common survey constraints, including central-tendency bias, leniency effects, and differences in how individuals have interpreted the same descriptors; although internal consistency has been strong, consistency has not guaranteed construct validity in every sub-dimension, particularly for complex judgments such as pedagogical usefulness and cultural appropriacy. Fifth, the corpus-based indicators have represented interpretable proxies rather than exhaustive representations of instructional quality; the feature set has captured lexical, syntactic, cohesion, and error-related properties, yet it has not directly measured factors such as motivational appeal, authenticity of

communicative purpose, alignment to specific learning outcomes, or pragmatic appropriateness of speech acts and interactional norms, all of which have influenced classroom effectiveness. Sixth, the generated corpus has been limited by the controlled prompt protocol and the selected text types, meaning that other AI prompting styles, multimodal materials, or adaptive task sequences may have produced different quality profiles and different predictor patterns. Seventh, the statistical modeling has been based on linear correlation and regression assumptions, which have supported interpretability but may not have captured non-linear relationships or interaction effects among features (e.g., cases where lexical appropriacy has mattered differently at different readability levels), and the sample size at the text level has constrained the complexity of models that could have been estimated without risking overfitting. Finally, although sensitivity audits have examined length and topic effects, other potential confounds such as rater fatigue, text familiarity, or hidden prompt artifacts have not been fully eliminated, which has meant that the reported predictive relationships have been best interpreted as strong internal evidence within the defined case rather than as universal parameters applicable to all AI-generated ESL materials in all contexts.

REFERENCES

[1]. Abdullah, F., & Ward, R. (2016). Developing a general extended technology acceptance model for e-learning (GETAMEL) by analysing commonly used external factors. *Computers in Human Behavior*, *56*, 238-256. https://doi.org/10.1016/j.chb.2015.11.036

[2]. Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, *45*(1), 5-35. https://doi.org/10.5054/tq.2011.244483

[3]. Boulton, A. (2012). Corpus consultation for ESP: A review of empirical research. In A. Boulton, S. Carter-Thomas, & E. Rowley-Jolivet (Eds.), *Corpus-informed research and learning in ESP: Issues and applications* (pp. 261-292). John Benjamins. https://doi.org/10.1075/scl.52.11bou

[4]. Chan, H., & Cheuk, H. N. J. (2020). Revisiting the notion of ESL: A corpus-based analysis of English textbook instructional language. *Ampersand*, *7*, 100066. https://doi.org/10.1016/j.amper.2020.100066

[5]. Chen, C.-F. E., & Cheng, W.-Y. E. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology*, *12*(2), 94-112. https://doi.org/10.64152/10125/44145

[6]. Chen, Y.-H., & Baker, P. (2016). Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics*, *37*(6), 849-880. https://doi.org/10.1093/applin/amu065

[7]. Chodorow, M., Gamon, M., & Tetreault, J. (2010). The utility of article and preposition error correction systems for English language learners: Feedback and assessment. *Language Testing*, *27*(3), 419-436. https://doi.org/10.1177/0265532210364391

[8]. Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly*, *42*(3), 475-493. https://doi.org/10.1002/j.1545-7249.2008.tb00142.x

[9]. Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, *48*(4), 1227-1237. https://doi.org/10.3758/s13428-015-0651-7

[10]. Crossley, S. A., Louwerse, M. M., McCarthy, P. M., & McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, *91*(1), 15-30. https://doi.org/10.1111/j.1540-4781.2007.00507.x

[11]. 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

[12]. 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

[13]. Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly*, *45*(1), 182-193. https://doi.org/10.5054/tq.2010.244019

[14]. Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*(1), 7-24. https://doi.org/10.1016/j.asw.2012.10.002

[15]. deBoer, F. (2014). Evaluating the comparability of two measures of lexical diversity. *System*, *47*, 139-145. https://doi.org/10.1016/j.system.2014.10.005

[16]. Fryer, L. K., Nakao, K., & Thompson, A. (2019). Chatbot learning partners: Connecting learning experiences, interest and competence. *Computers in Human Behavior*, *93*, 279-289. https://doi.org/10.1016/j.chb.2018.12.023

[17]. Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, *61*, 65-170. https://doi.org/10.1613/jair.5477

[18]. Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, *40*(5), 223-234. https://doi.org/10.3102/0013189x11413260

[19]. Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2006). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *38*(2), 193-202. https://doi.org/10.3758/bf03195564

[20]. Guilloteaux, M. J. (2013). Language textbook selection: Using materials analysis from the perspective of SLA principles. *The Asia-Pacific Education Researcher*, *22*, 231-239. https://doi.org/10.1007/s40299-012-0015-3
[21]. 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.2013.05.002
[22]. Huang, Y., & He, L. (2016). Automatic generation of short answer questions for reading comprehension assessment. *Natural Language Engineering*, *22*(3), 457-489. https://doi.org/10.1017/s1351324915000455
[23]. Işik, A. (2018). ELT materials evaluation: A system and criteria. *Theory and Practice in Language Studies*, *8*(7), 797-812. https://doi.org/10.17507/tpls.0807.11
[24]. Jinnat, A., & Md. Kamrul, K. (2021). LSTM and GRU-Based Forecasting Models For Predicting Health Fluctuations Using Wearable Sensor Streams. *American Journal of Interdisciplinary Studies*, *2*(02), 32-66. https://doi.org/10.63125/1p8gbp15
[25]. King, W. R., & He, J. (2006). A meta-analysis of the technology acceptance model. *Information & Management*, *43*(6), 740-755. https://doi.org/10.1016/j.im.2006.05.003
[26]. Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing*, *27*, 1-18. https://doi.org/10.1016/j.jslw.2014.10.004
[27]. Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474-496. https://doi.org/10.1075/ijcl.15.4.02lu
[28]. Mazgutova, D., & Kormos, J. (2015). Syntactic and lexical development in an intensive English for Academic Purposes programme. *Journal of Second Language Writing*, *29*, 3-15. https://doi.org/10.1016/j.jslw.2015.06.004
[29]. McNamara, D. S., Crossley, S. A., & Roscoe, R. D. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, *45*, 499-515. https://doi.org/10.3758/s13428-012-0258-1
[30]. McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, *47*(4), 292-330. https://doi.org/10.1080/01638530902959943
[31]. Md Ashraful, A., Md Fokhrul, A., & Md Fardaus, A. (2020). Predictive Data-Driven Models Leveraging Healthcare Big Data for Early Intervention And Long-Term Chronic Disease Management To Strengthen U.S. National Health Infrastructure. *American Journal of Interdisciplinary Studies*, *1*(04), 26-54. https://doi.org/10.63125/1z7b5v06
[32]. Md Fokhrul, A., Md Ashraful, A., & Md Fardaus, A. (2021). Privacy-Preserving Security Model for Early Cancer Diagnosis, Population-Level Epidemiology, And Secure Integration into U.S. Healthcare Systems. *American Journal of Scholarly Research and Innovation*, *1*(02), 01–27. https://doi.org/10.63125/q8wjee18
[33]. Md. Towhidul, I., Alifa Majumder, N., & Mst. Shahrin, S. (2022). Predictive Analytics as A Strategic Tool For Financial Forecasting and Risk Governance In U.S. Capital Markets. *International Journal of Scientific Interdisciplinary Research*, *1*(01), 238–273. https://doi.org/10.63125/2rpyze69
[34]. Ortega, L. (2015). Syntactic complexity in L2 writing: Progress and expansion. *Journal of Second Language Writing*, *29*, 82-94. https://doi.org/10.1016/j.jslw.2015.06.008
[35]. Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal*, *100*(2), 538-553. https://doi.org/10.1111/modl.12335
[36]. Ramineni, C. (2013). Validating automated essay scoring for online writing placement. *Assessing Writing*, *18*(1), 40-61. https://doi.org/10.1016/j.asw.2012.10.005
[37]. Reiter, E., & Belz, A. (2009). An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, *35*(4), 529-558. https://doi.org/10.1162/coli.2009.35.4.35405
[38]. Sakata, N. (2019). Profiling vocabulary for proficiency development: Effects of input and general frequencies on L2 learning. *System*, *87*, 102167. https://doi.org/10.1016/j.system.2019.102167
[39]. Schepers, J., & Wetzels, M. (2007). A meta-analysis of the technology acceptance model: Investigating subjective norm and moderation effects. *Information & Management*, *44*(1), 90-103. https://doi.org/10.1016/j.im.2006.10.007
[40]. Skorczynska Sznajder, H. (2010). A corpus-based evaluation of metaphors in a business English textbook. *English for Specific Purposes*, *29*(1), 30-42. https://doi.org/10.1016/j.esp.2009.05.003
[41]. Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing*, *19*, 51-65. https://doi.org/10.1016/j.asw.2013.11.007
[42]. Šumak, B., Heričko, M., & Pušnik, M. (2011). A meta-analysis of e-learning technology acceptance: The role of user types and e-learning technology types. *Computers in Human Behavior*, *27*(6), 2067-2077. https://doi.org/10.1016/j.chb.2011.08.005
[43]. Tomlinson, B. (2012). Materials development for language learning and teaching. *Language Teaching*, *45*(2), 143-179. https://doi.org/10.1017/s0261444811000528
[44]. Treffers-Daller, J., Parslow, P., & Williams, S. (2016). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*, *39*(3), 302-327. https://doi.org/10.1093/applin/amw009
[45]. Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly*, *36*(1), 157-178. https://doi.org/10.2307/41410412
[46]. Weigle, S. C. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing*, *18*(1), 85-99. https://doi.org/10.1016/j.asw.2012.10.006
[47]. Xia, M., Kochmar, E., & Briscoe, T. (2016). Text readability assessment for second language learners. Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications,
[48]. Yang, W. (2013). Response to Biber, Gray, and Poonpon (2011). *TESOL Quarterly*, *47*(1), 187-191. https://doi.org/10.1002/tesq.76