applied sciences

MDPI

Article

Incorporating Fine-Grained Linguistic Features and Explainable AI into Multi-Dimensional Automated Writing Assessment

Xiaoyi Tang [1], Hongwei Chen [1,*], Daoyu Lin [2] and Kexin Li [1]

[1] School of Foreign Studies, University of Science and Technology Beijing, Beijing 100083, China; tangxiaoyi@ustb.edu.cn (X.T.); likexin@xs.ustb.edu.cn (K.L.)
[2] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; lindy@aircas.ac.cn
* Correspondence: chenhongwei@ustb.edu.cn

Abstract: With the flourishing development of corpus linguistics and technological revolutions in the AI-powered age, automated essay scoring (AES) models have been intensively developed. However, the intricate relationship between linguistic features and different constructs of writing quality has yet to be thoroughly investigated. The present study harnessed computational analytic tools and Principal Component Analysis (PCA) to distill and refine linguistic indicators for model construction. Findings revealed that both micro-features and their combination with aggregated features robustly described writing quality over aggregated features alone. Linear and non-linear models were thus developed to explore the associations between linguistic features and different constructs of writing quality. The non-linear AES model with Random Forest Regression demonstrated superior performance over other benchmark models. Furthermore, SHapley Additive exPlanations (SHAP) was employed to pinpoint the most powerful linguistic features for each rating trait, enhancing the model's transparency through explainable AI (XAI). These insights hold the potential to substantially facilitate the advancement of multi-dimensional approaches toward writing assessment and instruction.

Keywords: fine-grained linguistic features; principal component analysis (PCA); SHapley Additive exPlanations (SHAP); explainable AI (XAI); multi-dimensional automated writing assessment

check for updates

Citation: Tang, X.; Chen, H.; Lin, D.; Li, K. Incorporating Fine-Grained Linguistic Features and Explainable AI into Multi-Dimensional Automated Writing Assessment. *Appl. Sci.* **2024**, *14*, 4182. https://doi.org/10.3390/app14104182

Academic Editor: Antonio Fernández-Caballero

Received: 16 April 2024
Revised: 8 May 2024
Accepted: 12 May 2024
Published: 15 May 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

As technology-powered advances are being incorporated into large-scale writing assessments, automated essay scoring (AES) has received increasing attention, offering a viable alternative to the traditionally time-intensive and laborious manual grading processes [1–3]. Due to remarkable advances in corpus linguistics [4,5], natural language processing (NLP) [6,7], and deep learning [3,8,9], AES has the benefits of improved consistency, reduced subjectivity, and constructive feedback by exploiting extensive linguistic features or incorporating cutting-edging algorithms [10–14]. Given the importance of AES, it is unsurprising that the investigation into the power of linguistic features characterizing writing quality has become a critical focus within the domains of writing assessment and instruction in the past five decades. Traditionally, feature-based approaches utilize regression analysis to generate essay scores with manually extracted linguistic features [15–17]. The advantage of feature-based approaches involves the interpretability and explainability of linguistic features for writing quality [12,18].

However, there continue to be a plethora of issues associated with traditional approaches using handcrafted linguistic features. Firstly, even though the application of corpus-based methods in evaluating essay quality has proliferated in recent years, many of these applications have attached great importance to exploring linguistic features for argumentative writing quality as it is the most common genre type in large-scale writing assessment [19–21]. Relatively few studies have investigated linguistic features characterizing the writing quality of other genres. Additionally, feature-based approaches hang on

a single holistic score with different constructs of writing quality being ignored [5,22–24]. More attention should be paid to explaining the inner workings of linguistic features underlying each trait, ensuring that the AES models are both robust and reflective of diverse writing constructs. Moreover, scholars have resorted to micro-linguistic features [15,25] and aggregated linguistic features derived from averaged combinations of micro-features that are theoretically interconnected [4,26,27] in characterizing the holistic score of writing. The effectiveness of micro-features, aggregated features, or their combined impact in accurately capturing different constructs of writing quality remains an area ripe for further investigation.

Given the above arguments, there is a need to enhance the understanding of fine-grained linguistic features that define writing quality, encompassing both holistic and analytic aspects. Additionally, traditional linear approaches do not sufficiently capture the intricate relationships between linguistic features and different constructs of writing quality. Therefore, this study pivots to non-linear models and provides a refined lens for examining the associations between linguistic features and different constructs of writing quality. Utilizing Principal Component Analysis (PCA) [28] and Explainable AI (XAI) with SHapley Additive exPlanations (SHAP) [29], the present study explores the extent of influence fine-grained linguistic features have on different constructs of writing quality and the underlying reasons, thereby improving the explainability of the proposed AES model and providing diagnostic and multi-dimensional feedback for writing instruction about what kinds of linguistic features should be given more emphasis in writing assessment and instruction.

2. Related Work

2.1. Holistic and Analytic Rating Traits for Writing Assessment

Writing assessment presents multifaceted challenges and relies heavily on subjective judgment, making it both time-consuming and demanding on human resources [30–34]. It is critical that human raters have a thorough understanding of the rating traits for scoring essays. The existing literature strongly advocates for the idea that a clear set of rating traits is critical to the validity of writing assessments. Without a clear framework, even the most experienced raters may be overwhelmed and unable to effectively measure the various dimensions of writing quality [35,36]. Scholars have also highlighted how specific scoring traits can help raters and educators produce reliable scores and maintain consistency in assessments, thereby mitigating the risk of making premature judgments influenced by personal biases [37,38]. Therefore, it is crucial to develop precise rating traits and their associated descriptors, which underpin the validity and reliability of writing assessments.

Holistic and analytic rating traits have been adopted to identify writing proficiency for different purposes [39,40]. Holistic scoring assigns a singular, comprehensive score to represent the quality of a sample essay. In contrast, analytic scoring evaluates essays through multiple sub-scales (e.g., *content* , *language*, and *organization*) that may or may not be averaged together to a total score [41]. Scholars have noted the efficiency of holistic scoring in educational and large-scale assessment settings, highlighting its cost-effectiveness and time-saving attributes [42,43]. As a result, holistic scoring has gained widespread acceptance in the field of writing assessment over recent years. Its utility lies in that raters can assign a single score, sparing them the necessity of multiple readings to grasp the meaning of analytical traits.

Although holistic scoring is widely adopted, it has been criticized for its use in the field of writing assessment and instruction. As critics point out, a notable limitation is that a single score fails to offer detailed feedback to students, particularly those struggling with writing, on the specific areas they need to improve. Additionally, it does not allow evaluators and educators to recognize and value the distinct characteristics that qualitatively differentiate one essay from another [44]. For example, two essays could be awarded the same overall score for vastly different reasons: one might be commended for its coherent structure, while the other could be celebrated for the sophistication of its

sentence construction. Such instances illuminate a fundamental weakness: it may achieve substantial inter-rater reliability, but this comes at the potential cost of compromising the validity of writing assessment [33,45,46].

Depending on the purposes of writing assessment, analytic rating traits involve different aspects of writing quality: *vocabulary*, *content*, *grammar*, *register*, and *organization* [39,47]. Analytic rating traits are often used to provide diagnostic information on students' writing proficiency and areas needing improvement [48]. For instance, a student might exhibit a strong command of syntax while struggling with essay organization. Therefore, an ever-increasing number of scholars have attached great importance to analytic rating traits in writing assessment due to their informative details and diagnostic feedback [49]. Despite the resource-intensive nature of analytic scoring, its advantages are compelling: firstly, it tends to yield more reliable scores compared to holistic scoring methods; secondly, it offers particular benefits for learners and students, whose writing skills may develop at varying paces across different constructs; thirdly, it facilitates more targeted and effective rater training and writing instruction; and fourthly, it is less prone to validity concerns, making it a robust choice for assessing writing quality [33,50–52]. Thus, the current study incorporates the strengths of holistic and analytic scoring, thereby improving score reliability in multi-dimensional writing assessment and providing diagnostic feedback for different constructs of writing quality.

2.2. Linguistic Features for Writing Quality

Scholars have attached much importance to linguistic features and their interactions with writing quality [53–58]. Generally, linguistic features characterizing writing quality can be categorized into three aspects: vocabulary, syntax, and cohesion [59,60]. In terms of lexical level, scholars have demonstrated that first language (L1) writers tend to use sophisticated phrasal constructions, greater lexical variation, less common words and bi-grams, more imageable words, and academic vocabulary [25,60–67]. Similar results have been observed in second language (L2) writing, revealing that high-quality essays are characterized by longer texts, a broader and more sophisticated vocabulary, and the use of less common and familiar words [66,68–74].

From a syntactic-level perspective, scholars have indicated that syntactic sophistication is positively related to the writing quality of L1 writers [59,60,75,76]. Subsequent studies have revealed a positive correlation with larger noun phrases, more modifiers, and additional words before the main verb and a negative correlation with simple declarative sentences and essay quality [25,77]. However, not all studies have reported a significant relationship between syntactic-level features and writing quality [25,78]. For L2 writers, complex nominal per clause has been identified as an indicator of writing quality. Additionally, the use of clausal subordination [56,72,79] and the frequency of passive voice are also positively correlated with writing quality.

Regarding cohesion, high-quality essays written by L1 writers are characterized by fewer positive connectives, less content word overlap, and increased usage of hedges and contrast expressions [77,80–82]. Research on L2 writers, however, yields mixed outcomes: cohesive devices such as semantic and lexical overlap, connectives, and givenness are negatively correlated with writing quality [55,68].

While existing research has considerably advanced our understanding of how linguistic features correlate with overall writing quality, there has been limited investigation of the correlation between linguistic features and different constructs of writing quality. Exploring multi-dimensional rating traits offers insights into linguistic features that enhance writing instruction across different writing constructs and covers a coherent picture of qualified linguistic features for model construction per rating trait. Furthermore, the advancement of computational analytic tools now allows extracting a wide array of fine-grained linguistic features, further enriching multi-dimensional writing assessment and model development.

2.3. Automated Essay Scoring (AES) Systems

The field of AES has experienced remarkable transformations over the last fifty years, fueled by significant breakthroughs in interdisciplinary research spanning NLP, computational linguistics, artificial intelligence (AI), and education [9,83–85]. Initially, AES systems depended on primitive models that utilized handcrafted linguistic features. However, they have evolved into more advanced systems, leveraging the power of AI technology to enhance their capabilities and accuracy. The foundation of AES was laid by Page and his team with Project Essay Grader (PEG), designed to streamline essay scoring on a large scale [86]. It focuses on surface-level linguistic features like *paragraphs* and *word counts* and combines multiple regression and NLP to predict essay scores [86,87]. However, PEG has faced criticism for overlooking semantic aspects of essays [88]. Moreover, the inner workings of linguistic features for each rating trait and processes involved in generating the holistic score remain undisclosed to the public [89,90].

In contrast to PEG's reliance on surface linguistic features, the Intelligent Essay Assessor (IEA) places a strong emphasis on the essence and content-driven linguistic features, harnessing the capabilities of Latent Semantic Analysis (LSA) and information retrieval [83,91]. Through statistical analysis, LSA maps words and essays to their semantic relations, revealing topic-based connections with matrix decomposition [92–94]. In the model training process, IEA first extracts linguistic features measuring different constructs of writing quality and uses LSA alongside information retrieval to decipher how features blend to characterize diverse writing traits, shifting focus towards evaluating essays' semantic depth.

E-rater combines multiple regression, the vector space model, and NLP techniques to extract linguistic features from essays [90,95,96]. This multifaceted approach allows E-rater to evaluate both the quality of language and the content of essays, with a notable focus on organizational quality as well. In the model training process, micro-features are initially aggregated into feature scores, with the significance of each feature determined through multiple regression analysis. For content-based feature extraction, both E-rater and IEA utilize vector space models. However, E-rater distinguishes itself by employing a keyword-based approach that excels in managing synonyms and polysemy. In contrast, IEA adopts a singular value decomposition (SVD)-based model, which excels in uncovering semantic relationships between texts and words by decomposing matrices, thereby facilitating topic-based analysis [26,97]. This methodological distinction enhances the E-rater's ability to interpret nuanced linguistic variations more effectively.

Leveraging a blend of cognitive science, NLP, computational linguistics, and AI, IntelliMetric marks the advent of AI-driven AES tools and is adept at identifying linguistic features for high-quality essays, thereby crafting models that simulate human scoring [27,98]. Employing a multifaceted approach that includes linear regression, Bayesian ridge regression, and LSA for feature selection and model development, IntelliMetric distinguishes itself with an extensive array of over 400 linguistic features spanning semantic, syntactic, and discourse levels. It undergoes training with essays previously scored by human raters, learning to replicate human evaluative techniques and understanding the interplay between rating traits and linguistic features [27].

As writing quality needs to be judged from different constructs, AES systems and models thus need to be trained with different rating traits corresponding to each construct of writing quality. Therefore, the current study aims to investigate the most powerful linguistic features for different constructs of writing quality. With cutting-edge computational linguistic tools, PCA, and XAI with the SHAP model, a wide range of fine-grained linguistic features are extracted and transformed into weighted components to explore linguistic features for the writing quality at both holistic and analytic rating traits, thereby improving the explainability of the proposed AES model. Moreover, both linear and non-linear AES models are set up to capture the relationship between linguistic features and multifaceted constructs of writing quality.

The new trends and research gaps summarized above prompt the current study to examine the following research questions:

- Which types of features characterize writing quality more accurately: micro-features, aggregated features derived from averaged combinations of micro-features, or a combination of both?
- Which kinds of AES models, linear vs. non-linear models, perform better in encapsulating the association between linguistic features and different constructs of writing quality?
- Which linguistic features emerge as the most powerful for different constructs of writing quality, considering both holistic and analytic rating traits?

3. Method

3.1. Dataset

Within the realm of AES, the Automated Student Assessment Prize (ASAP) dataset is the largest open-access dataset for the advancement and benchmark of machine learning algorithms, encompassing over 16,000 student submissions across eight distinct topics [85]. The length of essays varies with the prompt, generally ranging from 150 to 650 words across the collections. Authored by students across various educational levels, these essays were rigorously evaluated and scored by expert human raters. Although numerous studies have investigated the relationship between linguistic features for the argumentative writing quality [17,20,21,55,72], the relationship between linguistic features and different constructs of writing quality of exposition remains under-explored. To address these needs, we concentrate on the seventh corpus of the ASAP dataset, chosen due to its noteworthy attributes that cover four analytic rating traits: *Ideas*, *Organization*, *Style*, and *Conventions*.

Dataset 7, which includes 1730 narrative essays written by seventh-grade students, calls for expository essays on the topic of *Patience*. Essays in this collection are concise yet thorough, with an average length of about 250 words. Essays have been evaluated and assigned scores on a 0–3 scale for each rating dimension by the human raters. The final holistic score was derived by summing the individual scores provided by these two human evaluators. We selected this subset due to its comprehensive scoring rubric that spans multiple evaluative dimensions rather than giving a holistic score. It offers significant advantages for our research objective to explore multi-dimensional writing assessment. The distribution of scores across the analytical rating traits by the two raters is illustrated in Figure 1. The results demonstrated that the vast majority of students fall within the intermediate range in terms of writing proficiency.

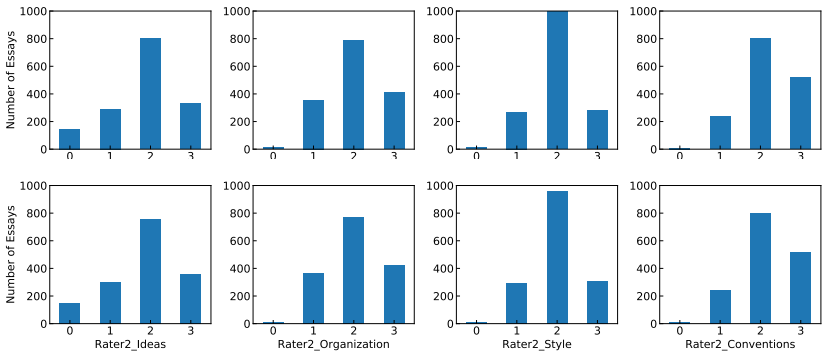


Figure 1. Distributionof essay scores across *Ideas*, *Organization*, *Style*, and *Conventions* rating traits by two independent raters in the ASAP dataset 7.

To assess the consistency and reliability of human grading, which serves as the gold standard for model construction, we computed the Pearson correlation coefficients between scores assigned by two independent evaluators across holistic and analytic rating traits.

The inter-rater reliability between human raters at both holistic and analytic rating traits was moderately acceptable, ranging from 0.545 to 0.722, as detailed in Table 1.

Table 1. Inter-raterreliability between human raters (*** $p < 0.001$).

Rating Traits	Pearson Correlation Coefficient (*r*)
Ideas	0.696 ***
Organization	0.576 ***
Style	0.545 ***
Conventions	0.567 ***
Holistic	0.722 ***

3.2. Research Instruments

All the essays were processed with advanced computational analytic tools, including Constructed Response Analysis Tool (CRAT) [99], Tool for the Automatic Analysis of Lexical Diversity (TAALED) [100], Tool for the Automatic Analysis of Lexical Sophistication (TAALES) [66], Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC) [101], Tool for the Automatic Analysis of Cohesion (TAACO) [53,102], and GAMET [103]. Linguistic features pertinent to various dimensions of writing quality were extracted using specific tools selected for several key reasons: First, their public accessibility and cost-free nature made them highly accessible. Second, the ease and speed with which these tools can be downloaded and operated provided unmatched convenience for selecting features and training models. Third, the strong construct validity of the linguistic features is supported by their foundation in established linguistic theories. Research has consistently demonstrated that these tools are effective in extracting linguistic features critical to assessing different facets of writing quality [53,66,99–103].

3.2.1. CRAT

CRAT, a user-friendly text analysis tool, offers genre-specific frequency and diversity insights, enriching its utility for a broad spectrum of textual analyses [99]. This freely available, cross-platform tool is operationalized through a graphical user interface, enhancing user accessibility. Its analytical prowess extends to evaluating lexical and phrasal similarities through advanced methodologies such as keywords, synonym overlaps, and LSA [104]. Additionally, it delves into phrasal congruence by scrutinizing the overlap of critical bi-grams and tri-grams and employing part-of-speech–sensitive slot-grams.

3.2.2. TAALED and TAALES

TAALED extracts 38 lexical diversity indices ranging from surface-level features such as the number of tokens and number of unique tokens to deep-level features such as the moving average type-token ratio (TTR) and measure of textual lexical diversity (MTLD) [100]. Six categories of indices, including a series of different tokens, lexical density, TTR, mass index [105], moving average TTR (50-word window), mean segmental TTR (50-word window), the hypergeometric distribution's D (HD-D) index [106], and MTLD [67], are extracted for measuring lexical diversity.

TAALES extracts 424 indices, which can be categorized into 13 types: word range and frequency, academic words, n-gram range, frequency and strength of association, semantic network, psycholinguistic word information, and other indices [66]. Table 2 illustrates linguistic features for lexical diversity and sophistication. Eight categories of indices regarding lexical diversity and eleven categories concerning lexical sophistication are included to ensure the richness and comprehensiveness of linguistic features for lexical resources.

Table 2. Linguistic features measuring lexical resources.

Indices of Lexical Diversity (No.)	Tool	Indices of Lexical Sophistication (No.)	Tool
token (6)	TAALED	word frequency (68)	TAALES
lexical density (2)	TAALED	word range (48)	TAALES
simple TTR (9)	TAALED	academic words (15)	TAALES
mass index (3)	TAALED	n-gram (227)	TAALES
moving average TTR (3)	TAALED	semantic network (14)	TAALES
mean segmental TTR (3)	TAALED	psycholinguistic word information (14)	TAALES
HD-D (7)	TAALED	age of acquisition/exposure (7)	TAALES
MTLD (9)	TAALED	contextual distinctiveness (8)	TAALES
		word recognition norms (8)	TAALES
		word neighbors (14)	TAALES
		other indices (5)	TAALES

3.2.3. TAASSC

TAASSC [101] explores more than 350 indices aimed at gauging various aspects of syntactic complexity and sophistication. These metrics are systematically organized into several distinct categories, including L2 Syntactic Complexity Analysis (L2SCA) [107], fine-grained phrasal complexity, clausal complexity, and syntactic sophistication. Please refer to [101] for the complete list of linguistic features measuring syntactic sophistication and complexity.

3.2.4. TAACO

TAACO involves more than 150 linguistic indices concerning three levels of local, global, and text cohesion [53,102]. The linguistic indices can be categorized into five types: connectives, givenness, TTR, lexical, and semantic overlap. There are altogether 25 types of connectives, ranging from basic connectives (*and*, *but*) to positive and negative logical connectives (*consequently*, *notwithstanding that*) to measure local cohesion. Givenness is a crucial measure of text cohesion, reflecting how information can be inferred from the context provided by preceding sentences. TTR plays a significant role in this domain by quantifying lexical diversity within essays. It does so by calculating the proportion of unique words (types) to the total word count (tokens) [53,102]. As an essential measurement of text cohesion, a series of TTR-based indices regarding the content word TTR, lemma TTR, content lemma TTR, and TTR for bi-grams and tri-grams are included in TAACO.

3.2.5. GMAET

Six categories of errors concerning duplication, grammar, spelling, style, typography, and white space are calculated with Grammar and Mechanics Error Tool (GMAET) [103]. Grammar errors consist of noun errors with agreement, adjective errors related to comparative and superlative errors, adverb word order errors, connector errors, negation errors, fragment errors, and verb errors with verb usage, person, tense, and aspect [103]. Duplication errors refer to word duplication. Spelling errors involve lowercase, missing hyphens, and apostrophes. Style errors examine features including verbosity, redundancy, and inappropriate word selection, which can detract from the clarity and conciseness of writing. Typography errors, on the other hand, involve mistakes related to capitalization and associated punctuation, which are crucial for maintaining the formal structure and readability of the text. Additionally, white space errors, which pertain to improper spacing within the document, can significantly impact the visual presentation and organization of the content.

3.3. Research Procedures

Figure 2 demonstrates the flowchart of research procedures for feature-based approaches. Based on the computational analytic tools, linguistic features were extracted to characterize the writing quality at holistic and analytic rating traits. This pivotal stage enabled the distillation of each essay's core attributes into quantifiable metrics, laying the foundation for in-depth analysis. To avoid the problems of multicollinearity and over-fitting

caused by too many linguistic features, we first included all linguistic features extracted by computational analytic tools for standardization and normally distributed detection. Then, the multicollinearity tests were conducted to ensure the selected linguistic features were not overlapping. PCA was conducted to condense the information of linguistic indices and keep those features that contribute the most to the variance.

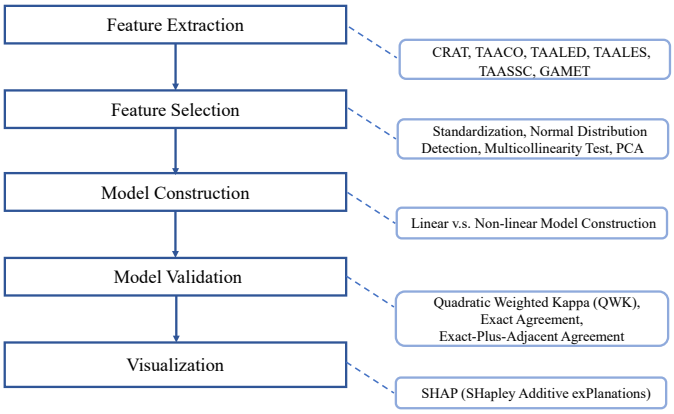


Figure 2. The flowchart of research procedures.

Qualified linguistic features were employed to develop AES models for holistic and analytic rating traits. Linear and non-linear AES models were trained to mirror the scoring patterns of human raters by systematically learning the characteristics of qualified linguistic features. To guarantee the robustness and reliability of the training and evaluation processes for these models, a five-fold cross-validation was implemented. This approach divided the dataset such that 60% was allocated for training purposes, 20% served as a validation set to fine-tune model parameters, and the remaining 20% was utilized for testing the models' performance. The training sets were used to select the linguistic indices for modeling, while the test sets were employed to verify whether the AES models were valid or not. Finally, SHAP values were applied to shed light on how linguistic features affect the writing quality across different constructs, enhancing our understanding of the inner workings of the proposed AES model.

3.3.1. Feature Extraction and Selection

Six computational analytic tools were used to extract linguistic features for model construction. The feature selection was a pivotal step in refining the array of linguistic attributes extracted from essays, guaranteeing that only the most relevant and robust features were incorporated into the model training process. Initially, an extensive set of 2044 features was extracted from each essay, utilizing advanced linguistic tools such as CRAT [99], TAALED [100], TAALES [66], TAASSC [101], TAACO [53], and GAMET [103].

The advantage of combining a great number of computational analytic tools was to ensure the richness and comprehensiveness of the linguistic features for measuring different constructs of writing quality. If these linguistic indices were directly included in the analysis, the model would be complicated and unstable, leading to over-fitting problems. Therefore, several methods were used to remove the redundant features and keep those that contribute the most to writing quality (see Table 3). The non-parametric Kolmogorov–Smirnov test was employed to assess the adherence of the data to a normal distribution. In addition, the variance threshold was used to remove the feature that did not vary much within itself (variance lower than 0.01) [8]. The variance threshold seeks to eliminate variables demonstrating negligible variation within the dataset on the premise that such features are improbable to yield significant revelations regarding paper quality.

Table 3. Progressive refinement of features during the feature selection phase, illustrating the reduction in the number of features at each step.

Feature Selection Stage	Number of Features
Initial Features Extracted	2044
After Removing Features with Variance < 0.01	934
After Removing Non-Normally Distributed Features	916
After Removing Features with Multicollinearity ($r > 0.90$)	362

The multicollinearity tests were conducted to overcome the overlapping problems between features. These tests aimed to identify instances where two or more indices exhibited a high degree of correlation. In scenarios where indices were found to be highly correlated ($r > 0.90$), the index demonstrating the strongest correlation with the scores was retained for subsequent analysis, while redundant indices were excluded from the model [4]. After conducting filter methods, the number of linguistic features was reduced from 2044 to 362. The rigorous feature selection process honed the machine learning model's training on a refined feature set, ensuring precise and stable scoring of essays.

PCA was used to condense the information of linguistic indices and keep those that contribute the most to the variance. Scholars have indicated three standard methods for determining the number of principal components. The first method was an eigenvalue-based method, which followed the Kaiser–Harris criterion and retained components with eigenvalues greater than 1 [108]. The second method was Cattell's scree test [109]. By plotting the eigenvalues and principal components, the curve of the scree plot was displayed, and the principal components above the maximum change were retained. The last method was the parallel analysis method [110]. A principal component was considered significant and retained for further analysis if its eigenvalue surpassed the corresponding average eigenvalue derived from a collection of random matrices. The decision on the number of principal components to retain was informed by this eigenvalue criterion, in conjunction with the insights gained from the initial two methods employed in the analysis. To ensure the correlation of the indicators contained in each principal component, only the absolute value of the eigen loading greater than 0.35 was included for further analysis [4].

Figure 3 shows the proportion of the total data variance accounted for by each principal component. Seven principal components were retained by taking account of the first two methods [108,109]. Among them, the first principal component explained 10.1% (74 micro-features) of the variance of the total data variation. The last six principal components explained 5.9% (52 micro-features), 4.9% (26 micro-features), 4.0% (14 micro-features), 3.3% (19 micro-features), 2.5% (6 micro-features), and 2.2% (3 micro-features), respectively. Hence, the seven principal components (194 micro-features) explained 32.9% of the total variance of the original data set.

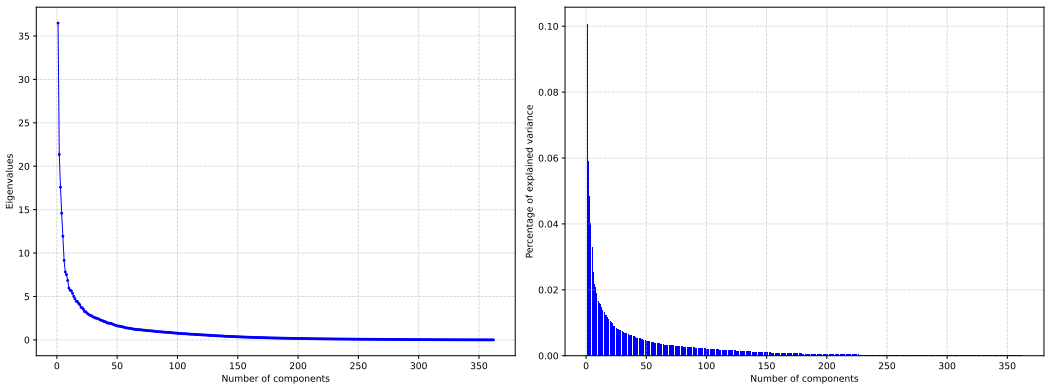


Figure 3. Scree plot illustrating the variance explained by the principal components.

3.3.2. Model Construction and Validation

Micro-features, aggregated features derived from averaged combinations of micro-features, and a combination of those were used to investigate the strengths of characterizing writing quality. With qualified linguistic features, both linear and non-linear models were developed using the following six machine learning algorithms, namely Bayesian Ridge Regression [111,112], Linear Regression [112,113], Stochastic Gradient Descent Regression [114], Random Forest Regression [115], Decision Tree Regression [116], and Support Vector Regression [117]. The first three algorithms were used for linear AES model construction, whereas the last three algorithms were used for non-linear AES model construction.

- **Bayesian Ridge Regression:** integrates ridge regression and Bayesian inference, providing a probabilistic perspective on linear regression, excelling in data-scarce environments or large feature domains by proficiently approximating a weight distribution.
- **Linear Regression**: establishes a linear correlation between the input linguistic features and the predicted scores. Its straightforwardness and interpretability render it a foundational benchmark for numerous regression endeavors.
- **Stochastic Gradient Descent Regression**: iteratively adjusts parameters to reduce the loss function, promoting efficiency and the capacity to scale.
- **Random Forest Regression**: An ensemble method that consolidates outputs from various decision trees, renowned for its precision, robustness to over-fitting, and adeptness in handling non-linear correlations.
- **Decision Tree Regression**: builds a tree-like model that bifurcates the dataset into branches predicated on attribute values, offering an intuitive method capable of deciphering intricate patterns.
- **Support Vector Regression**: applies support vector machine concepts to regression, focusing on maintaining errors within a defined threshold, excelling in high-dimensional contexts and situations where error tolerance is crucial.

Four evaluation metrics concerning R^2, Quadratic Weighted Kappa (QWK), Exact Agreement, and Exact-Plus-Adjacent Agreement were used to assess the discrepancy between the predicted scores generated by the model and the scores assigned by human raters.

R^2 serves to quantify the percentage of the variance in human raters' scores that can be accounted for by the linguistic features utilized in a model. If the R^2 of the proposed model was 0.60, 60% of the variability observed in the scores assigned by human raters could be explained by the inputs to the model. R^2 usually varied from 0% (the model did not explain any of the variations) to 100% (the model explained all the variations). Its formula is given by the following:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

where y_i represents the actual scores assigned by human raters, serving as a benchmark for the AES model's accuracy. The model aims to closely match its predictions $\hat{y}_i$ to these actual scores, effectively emulating human scoring behavior. $\bar{y}$ denotes the average of all actual scores, providing a reference point for assessing variations in both actual and predicted scores.

QWK was employed to evaluate the level of agreement between two raters. QWK values range from 0, indicating no match between raters, to 1, which signifies a perfect match. A great majority of studies have used QWK as a gold standard in evaluating the accuracy of AES models [3,6,83,118]. Therefore, the current study adopts QWK as the metric to gauge the accuracy between human raters' scores and the scores predicted by the model.

$$QWK = 1 - \frac{2 \times \sum_{i,j} w_{ij} o_{ij}}{\sum_{i,j} w_{ij} e_{ij}},$$

where w_{ij} is the weight assigned to the disagreement between raters for items in categories i and j, o_{ij} is the observed frequency of agreement, and e_{ij} is the expected frequency of agreement under chance.

Exact Agreement was designed to pinpoint the proportion of essays for which the scores assigned by human raters and those predicted by the model are precisely the same. This measure offers a straightforward assessment of the model's accuracy in replicating human scoring decisions without any margin for error.

Exact-Plus-Adjacent Agreement was used to measure the percentage of the number of essays with the same score or within 1 point of the score.

3.4. Visualization of Feature Importance: SHapley Additive exPlanations (SHAP)

XAI with the SHAP model [29], underpinned by game theory principles, was utilized to clarify the impact of linguistic features on different constructs of writing quality. SHAP provides a coherent and theoretically sound approach for attributing the contributions of input variables to a model's predictions. This method is distinguished by its capacity to illuminate the overall importance of features across the dataset and to offer detailed explanations for individual prediction instances. For each feature, SHAP values quantify its impact on the variance between the specific prediction and the dataset's mean prediction. Furthermore, an analysis of SHAP values reveals the key features for each evaluative criterion, thereby enhancing the transparency and explainability of the proposed AES model. The summary plot uses the horizontal axis to display SHAP values, showing how each linguistic feature affects the model's prediction: positive for enhancement and negative for detriment. The vertical axis ranks these features by their impact based on SHAP value aggregation. Color shifts from blue to red reflect the feature's value in observations (blue for lower values, red for higher), offering a quick visual insight into the distribution and impact of feature values on the model's output.

4. Results and Discussion

4.1. Micro-Features, Aggregated Features, and Their Combination in Characterizing Writing Quality

In the investigation of the power of different types of features for assessing writing quality, three distinct groups were utilized: micro-features, aggregated features synthesized from the averaged combinations of micro-features, and a comprehensive approach incorporating both micro-features and aggregated features. The selection criteria for linguistic features and components to be included in the further analysis were based on their demonstrated significant correlation with writing quality, accompanied by at least a small effect size ($r > 0.100$) [4]. To assess the influence of these different feature types on the evaluation of writing quality, we conducted experiments on the aforementioned six models using a validation dataset. We also delved into the influence of different feature types across holistic and analytic rating dimensions: *Ideas*, *Organization*, *Style*, *Conventions*, and *Holistic*. The analysis encompassed three scenarios: utilizing the qualified micro-features only (194 features), exploring these seven aggregated features generated by PCA only, and examining a combination of both.

To better examine the difference in characterizing writing quality, we computed the mean of the evaluation metrics of the six methods for comparative analysis to facilitate the comprehension of how micro-features, aggregated features, and their combination contribute to the prediction of writing quality. Table 4 indicates the effectiveness of different feature sets: micro-features only, aggregated features from the averaged combinations of micro-features, and their combination in characterizing writing quality.

Table 4. Comparison between micro-features, aggregated features, and a combination of both.

Evaluation Metrics	Aggregated Features Only	Micro-Features Only	Combined
R^2	0.517	0.685	**0.690**
QWK	0.705	0.820	**0.822**
Exact Agreement	0.225	**0.288**	0.285
Exact-Plus-Adjacent Agreement	0.470	**0.560**	0.555

The most optimal results are highlighted in bold for clarity and emphasis.

The coefficient of determination R^2 showed that the combined feature approach yielded the highest value (R^2 = 0.690). This indicated a marginally superior fit to the data when compared to using micro-features (R^2 = 0.685) alone and a notably better performance over aggregated features (R^2 = 0.517). Paired *t*-tests were conducted to explore whether significant improvement exists. The results showed that the micro-feature ($p < 0.050$) and combined feature approaches ($p < 0.050$) were significantly improved when compared with the aggregated feature approach. However, no significant differences were found between the micro-feature and combined feature approaches ($p > 0.050$).

In terms of QWK, the results also supported the superiority of the combined features model with a score of 0.822, marginally outperforming the micro-features model (0.820) and significantly surpassing the aggregated features model (0.705). This increment, although slight, indicated that there were no significant differences between the micro-feature approach and the combined feature approach ($p > 0.050$). As for Exact Agreement and Exact-Plus-Adjacent Agreement, a slight advantage for micro-features in characterizing writing quality has been reported when compared with the combined feature approach ($p > 0.050$).

Overall, similar results have also indicated that the micro-feature and combined feature approaches provided a robust basis for characterizing writing quality when compared with the aggregated feature approach [4]. Additionally, the results demonstrated no differences between AES models with micro-features and AES models combining micro-features and aggregated components ($p > 0.50$). Therefore, further analysis will include only micro-features under the aggregated components.

4.2. Comparison between Linear and Non-Linear AES Models in Capturing Writing Quality

Both linear and non-linear methods were used to explore the relationship between linguistic features and essay scores across different rating traits. Table 5 demonstrates the performance of linear and non-linear AES models in delineating the relationship between linguistic features and the multifaceted constructs of writing quality.

Following these tests, Random Forest Regression emerged as the most effective model, demonstrating superior performance across all rating traits and evaluation metrics. It achieved the highest (R^2) values, indicating a strong explanatory power for variations in writing quality, with notable scores in *Ideas* (0.777), *Organization* (0.769), *Style* (0.740), *Conventions* (0.725), and *Holistic* (0.771). Additionally, this model outperformed others regarding QWK, with scores indicating a robust agreement between predicted and actual ratings, peaking at 0.863 in the holistic rating trait. Decision Tree Regression showed remarkable strength in Exact Agreement, particularly in the *Ideas* and *Organization*, suggesting its capability to mirror the scoring patterns of human raters precisely. However, the Random Forest model excelled in the Exact-Plus-Adjacent Agreement metric across all the rating traits.

The Random Forest Regression model stood out for its robust capability across multiple dimensions of writing quality, supported by high agreement rates with human raters. The results suggested that non-linear AES models better captured the relationship between linguistic features and writing quality across different rating traits, which could enhance the overall accuracy and reliability of AES tasks, ensuring a more systematic assessment of writing quality that mirrors the complex judgment process of human raters.

The confusion matrices for each rating trait, illustrated in Figures 4 and 5, provided further insights into the accuracy of AES models, revealing the distribution of predictions across the scoring traits. These matrices are crucial for understanding where the model excels and where it may confuse certain score categories. With qualified micro-features, the proposed AES model not only achieves high accuracy but also aligns closely with the nuanced judgment of human raters across multiple dimensions of writing quality.

Table 5. AES models' performance based on micro-features across different rating traits (E: Exact Agreement; E + A: Exact-Plus-Adjacent Agreement).

Model	Ideas	Organization	Style	Conventions	Holistic
Bayesian Linear Regression					
R^2	0.629	0.556	0.592	0.547	0.680
QWK	0.767	0.679	0.731	0.670	0.806
E	0.481	0.506	0.576	0.481	0.178
E + A	0.892	0.917	0.975	0.936	0.503
Linear Regression					
R^2	0.637	0.542	0.582	0.496	0.672
QWK	0.783	0.685	0.722	0.670	0.811
E	0.433	0.497	0.551	0.503	0.159
E + A	0.911	0.914	0.965	0.917	0.487
Stochastic Gradient Descent Regression					
R^2	0.614	0.554	0.598	0.515	0.670
QWK	0.775	0.699	0.753	0.668	0.801
E	0.439	0.519	0.576	0.478	0.185
E + A	0.911	0.914	0.981	0.927	0.487
Random Forest Regression					
R^2	**0.777**	**0.679**	**0.740**	**0.725**	**0.771**
QWK	**0.863**	**0.778**	**0.832**	**0.819**	**0.863**
E	0.646	0.640	**0.720**	**0.659**	0.280
E + A	**0.952**	**0.949**	**0.987**	**0.965**	**0.643**
Decision Tree Regression					
R^2	0.666	0.469	0.518	0.524	0.527
QWK	0.831	0.742	0.762	0.754	0.763
E	**0.678**	**0.682**	0.720	0.659	**0.557**
E + A	0.904	0.898	0.924	0.908	0.618
Support Vector Regression					
R^2	0.704	0.616	0.685	0.631	0.732
QWK	0.816	0.746	0.770	0.741	0.839
E	0.573	0.621	0.650	0.608	0.376
E + A	0.924	0.930	0.978	0.933	0.586

The most optimal results are highlighted in bold for clarity and emphasis.

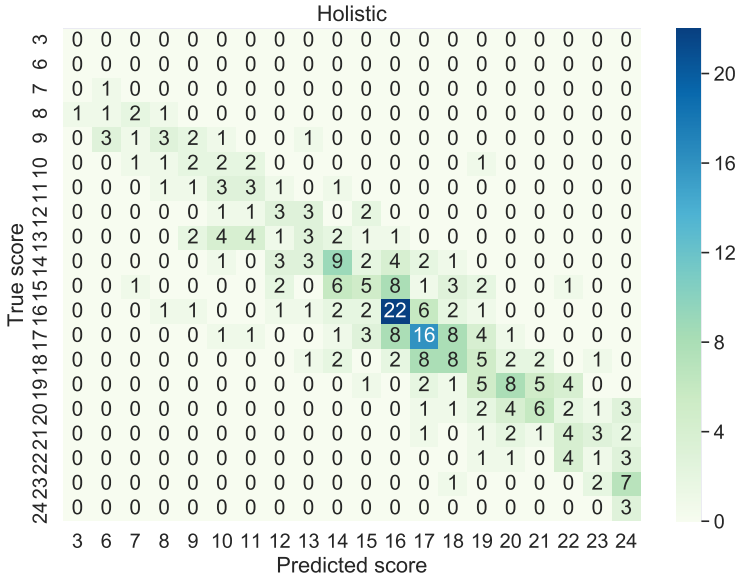
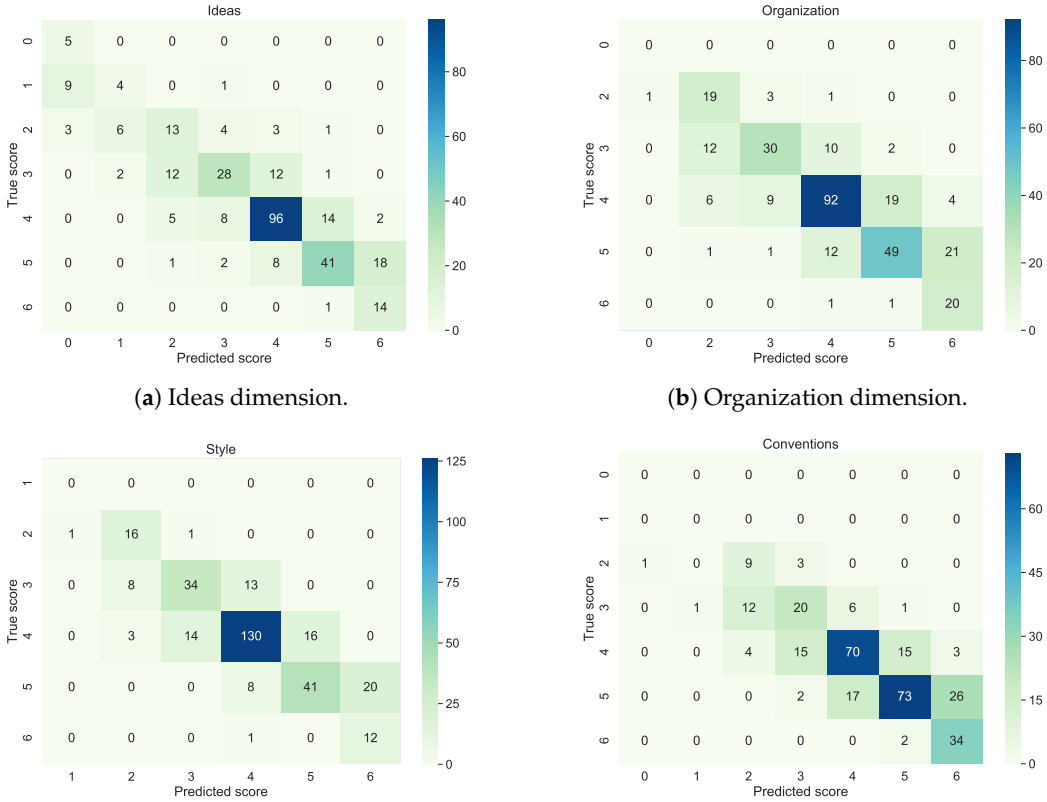


Figure 4. Confusion matrix for the Holistic dimension.



(**a**) Ideas dimension.

(**b**) Organization dimension.

(**c**) Style dimension.

(**d**) Conventions dimension.

Figure 5. Confusion matrices for the Random Forest Regression model's performance across various essay scoring dimensions: Ideas, Organization, Style, and Conventions.

We also compared the modeling results with other baseline models. Numerous scholars have conducted deep learning techniques to construct AES models for writing assessments. Table 6 shows the QWK scores at the holistic rating trait on the ASAP dataset 7. MN [119] utilizes a novel approach by evaluating the relationship between unlabeled essays and a repository of selected essays for score prediction. HISK + BOSWE and v-SVR [120] explored a method that encompassed both surface-level features and deep-level semantic features. TSLF-ALL [121] proposed a dual-phase learning architecture that integrated the advantages of feature-engineered approaches. Qe-C-LSTM [122] involved a method employing an advanced deep convolutional recurrent neural network for enhanced essay scoring accuracy.

Table 6. Comparison of QWK values for the holistic rating trait with Random Forest Regression and the baselines.

Approaches	QWK Scores
MN [119]	0.790
HISK + BOSWE and v-SVR [120]	0.804
TSLF-ALL [121]	0.801
Qe-C-LSTM [122]	0.815
Random Forest Regression	**0.863**

The most optimal results are highlighted in bold for clarity and emphasis.

The results demonstrated that our proposed approach with Random Forest Regression outperformed the QWK score by 4.8% over the benchmark Qe-C-LSTM. The paired *t*-test results indicated that the Random Forest Regression approach significantly improved performance ($p < 0.050$) when compared with those baselines. The observed improvement in performance can be attributed to a comprehensive and coherent representation of linguistic

features that characterize various constructs of writing quality. By encompassing a broad spectrum of linguistic features characterizing different constructs of writing quality, the proposed AES model is better equipped to capture the multifaceted characteristics that constitute effective writing.

4.3. Most Powerful Linguistic Features per Rating Trait: Visualization via SHAP Model

To identify the most powerful linguistic features per rating trait, we calculated SHAP values to offer a rigorous quantification of each feature's impact on the model's prediction, thereby elucidating the relative importance of various linguistic features in determining different constructs of writing quality. Figures 6 and 7 encapsulate the top 10 influential features for the holistic and analytic rating traits. Detailed information of linguistic features can be found in Appendix A. The variability in the impact of each feature, as indicated by the spread of SHAP values, underscores the complex interplay between different linguistic attributes when assessing writing quality. The insights derived from this analysis can inform educational interventions, automated writing assessment tools, and further research into the features that contribute to effective writing.

Figure 6 illustrates the SHAP values associated with the top 10 linguistic features that characterize writing quality at the holistic rating trait, thereby offering insights into the relative importance of these linguistic features. It was observable that certain features such as (`hdd42_cw`) and (`nsentences`) predominantly had a positive SHAP value distribution, indicating that higher values of these features typically correspond to better writing quality. In contrast, (`Freq_N_CW`) exhibited a predominantly negative distribution of SHAP values, suggesting that higher values of these features may be associated with lower writing quality. The results offered support to previous findings that less proficient writers tended to utilize more high-frequency words compared to their more proficient counterparts [15,55].

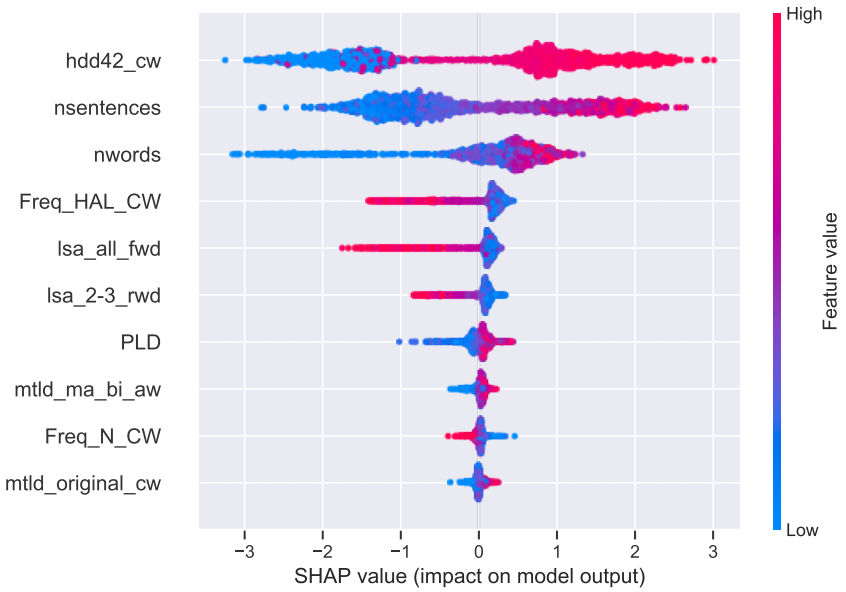
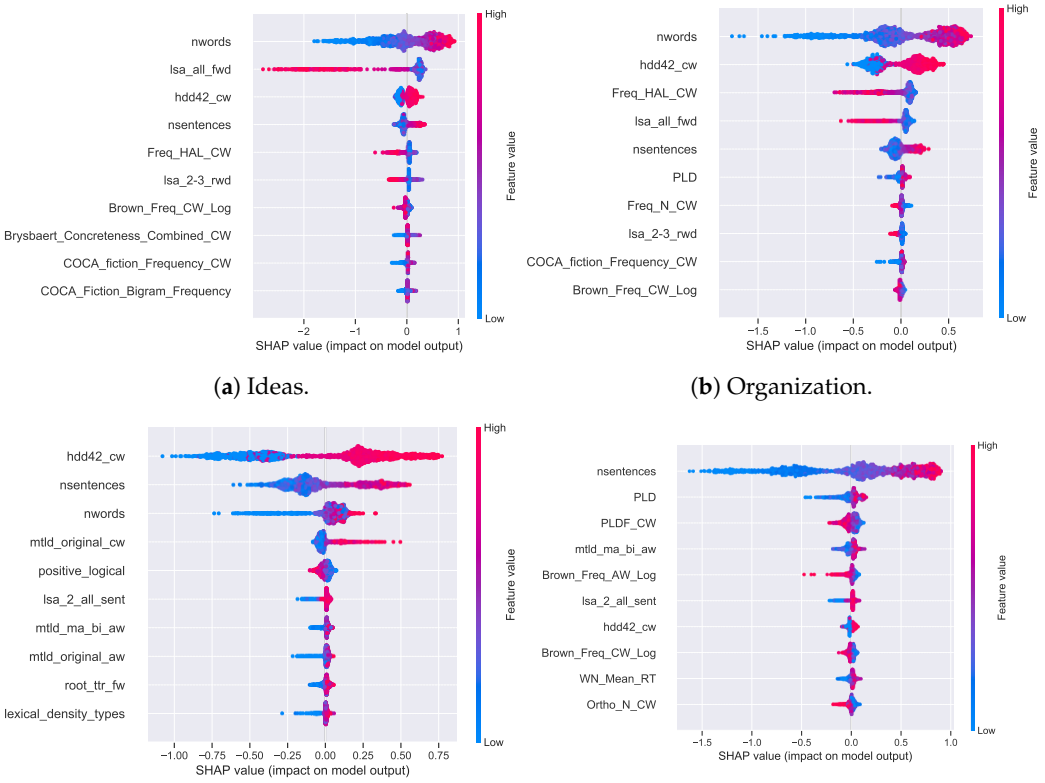


Figure 6. SHAP values: visualization of the top 10 features impacting the holistic rating trait.

The features were arranged in descending order of their mean absolute SHAP values, signifying their average impact across the dataset. *HD-D* (`hdd42_cw`), measuring the lexical diversity, had the highest mean absolute SHAP value, indicating it was the most influential in the model's assessment of writing quality. Additionally, lexical diversity indices, including the measure of textual lexical diversity (`mtld_ma_bi_aw` & `mtld_original_cw`) were identified as significant linguistic features for expository writing quality. The findings corroborated earlier research demonstrating that features based on the type-token ratio were key indicators of writing quality, as substantiated by [65,67,105].

Other features including the number of sentences (`nsentences`), number of words (`nwords`), word frequencies (`Freq_HAL_CW`), LSA-based indices (`lsa_all_fwd & lsa_2-3_rwd`), mean phonological Levenshtein distances (`PLD`), and neighborhood frequency of content word (`Freq_N_CW`) were identified as predictive linguistic features for the overall writing quality of exposition. The results offered support to previous studies showing that LSA-based features [81,82] and word frequencies [15,55] are significantly correlated with writing quality. With the XAI technique (SHAP), thorough and transparent linguistic features characterizing the overall writing quality of exposition are provided, which enhances the explainability of the proposed AES model and offers diagnostic and individualized feedback for multi-dimensional writing instruction and assessment.



(**a**) Ideas. (**b**) Organization.

(**c**) Style. (**d**) Conventions.

Figure 7. SHAP values: visualization of the top 10 features for different constructs of writing quality.

In terms of *Ideas*, SHAP values were plotted on the horizontal axis, quantifying the degree to which each feature affected the predictive model's output (see Figure 7a). The vertical listing of features indicated their relative importance based on the magnitude of their mean absolute SHAP values, with number of words (`nwords`) appearing as the most predictive feature. Upon examination of the plot, it was apparent that number of words (`nwords`) was associated with higher SHAP values, suggesting that it was positively related to writing quality for the *Ideas* dimension. Conversely, the semantic overlap of function words (`lsa_all_fwd`) was associated with negative SHAP values, indicating a potential association with lower quality in the *Ideas* dimension. Other features regarding *HD-D* (`hdd42_cw`), number of sentences (`nsentences`), word frequencies (`Freq_HAL_CW` & `COCA_fiction_Frequency_CW` & `COCA_Fiction_Bigram_Frequency`), LSA-based indices (`lsa_all_fwd & lsa_2-3_rwd`), number of content words in text with frequency score (`Brown_Freq_CW_Log`), and word concreteness (`Brysbaert_Concreteness_Combined_CW`) were identified as predictive linguistic features for the writing quality of *Ideas*.

As for *Organization*, analysis of the summary plot (see Figure 7b) revealed that features such as number of words (`nwords`) and *HD-D* (`hdd42_cw`) displayed a predominantly positive distribution of SHAP values, implying that these features were associated with a

more favorable assessment of the 'Organization' dimension. On the other hand, features such as word frequencies of content words (`Freq_HAL_CW`) exhibited a distribution towards negative SHAP values, suggesting a negative association with the writing quality in terms of *Organization*. Scholars have indicated words that were more frequently used would be acquired easily and earlier by students [59,64,66]. Thus, more frequent uses tended to get lower scores in writing tasks.

With Regard to *Style*, the results revealed that *HD-D* (`hdd42_cw`) had the highest mean absolute SHAP value, indicating it was the most influential in the model's assessment of writing quality (see Figure 7c). Additionally, linguistic features including number of sentences (`nsentences`), number of words (`nwords`), lexical diversity indices (`mtld_original_cw` & `mtld_ma_bi_aw` & `mtld_original_aw` & `root_ttr_fw`), the similarity between two adjacent sentences (`lsa_2_all_sent`), number of positive logical connectives (`positive_logical`), and lexical density (`lexical_density_types`) were identified as predictive linguistic features for the overall writing quality of exposition. The results aligned with previous research, showing that the richness of vocabulary within a text was an important indicator of expository writing quality [66,71,74].

For *Conventions* (see Figure 7d), the analysis revealed that the number of sentences (`nsentences`) and measure of textual lexical diversity (`mtld_ma_bi_aw`) were the features with a predominantly positive distribution of SHAP values. This suggested that these features were likely to correlate positively with writing quality. In contrast, the feature measuring the number of words in the text with a frequency score (`Brown_Freq_AW_Log`) displayed a distribution towards negative SHAP values, potentially indicating a negative relationship with writing quality. The findings corroborate earlier research, indicating that less proficient writers tend to employ a higher number of high-frequency words compared to their more proficient counterparts. This pattern suggests a reliance on more common vocabulary among writers with lower proficiency levels [15,55].

The above analysis through SHAP vividly illustrates the multifaceted nature of linguistic features characterizing writing quality. Existing AES models, leveraging machine learning algorithms, are often criticized for their opaque, black-box nature, which hinges critically on the caliber of the training datasets. Integrating XAI methodologies, particularly SHAP, into these algorithms marks a pivotal advancement in demystifying the underlying mechanisms and providing a transparent analysis of linguistic features that characterize different constructs of writing quality.

5. Conclusions

With advances in computational linguistics, NLP, and AI, scholars have provided thought-provoking insights on incorporating feature-based approaches and machine learning algorithms into AES tasks. However, little has been done to explore the relationship between deep-level linguistic features and different constructs of expository writing quality. Additionally, the inner workings of linguistic features in the black box are still unknown. The present study takes full advantage of the interdisciplinary fields combining corpus linguistics, NLP, and XAI. It also provides insightful perspectives on developing linear and non-linear models to examine the relationship between linguistic features and essay scores from holistic and analytic rating scales.

With cutting-edge computational analytic tools and PCA, deep-level linguistic features were extracted to condense and reduce the noise, thereby eliminating redundancy and enhancing data quality. The results demonstrated that the micro-feature and combined feature approach provided a robust basis for characterizing writing quality compared to the aggregated feature approach. Furthermore, linear and non-linear AES models were constructed to explore the complex relationship between linguistic features and writing quality. The results indicated that the proposed non-linear AES model with Random Forest Regression outperformed the QWK score by 4.8% over the benchmark Qe-C-LSTM, which offered support to the idea that the relationship between linguistic features and writing quality was complicated as multi-dimensional constructs were involved. Lastly, SHAP

values were plotted to identify the most powerful linguistic features per rating trait, thereby improving the explainability of the proposed model. Linguistic features for different constructs of writing quality can provide diagnostic feedback for writing instruction and revision. Teachers can also carry out diversified and individualized teaching according to the feedback of linguistic features, characterizing the writing quality through holistic and analytic rating traits.

While our findings contribute to the development of more sophisticated, fair, and transparent AES systems, they also highlight the need for a broader representation of writing genres and a more diverse range of educational backgrounds. Broadening the scope of the datasets would enhance the robustness and generalizability of the proposed AES models. This expansion is crucial for achieving a comprehensive understanding of multi-dimensional writing assessment across large-scale educational contexts. Furthermore, future studies could benefit from integrating large language models [7,123], which promise further to improve the accuracy and generalization capabilities of AES systems. Such advancements are essential for keeping pace with the evolving demands of multi-dimensional assessment.

Author Contributions: Conceptualization, X.T., H.C. and D.L.; methodology, X.T. and D.L.; software, X.T. and D.L.; validation, X.T., H.C., D.L. and K.L.; writing—original draft preparation, X.T.; writing—review and editing, H.C., D.L. and K.L.; supervision, H.C., D.L. and K.L.; funding acquisition, X.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by China Postdoctoral Science Foundation under Grant Number 2023M740222.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Definitions of Linguistic Features for Expository Writing Quality

Linguistic features	Description
Brown_Freq_AW_Log	Number of words in text with frequency score, all words, log transformed
Brown_Freq_CW_Log	Brown Frequency CW Logarithm; number of content words in text with frequency score
Brysbaert_Concreteness_Combined_CW	Brysbaert Concreteness Combined (content words only)
COCA_fiction_Frequency_CW	COCA Fiction Frequency (content words only)
COCA_Fiction_Bigram_Frequency	Mean bigram frequency score
Freq_HAL_CW	Word frequencies from Hyperspace Analogue to Lang (HAL) corpus norms
Freq_N_CW	Ave freq (Freq_HAL) of orthographic neighborhood
hdd42_cw	HD-D uses the hypergeometric distribution to calculate the probability of encountering one of its tokens in a random sample of 42 tokens (content words only)
lexical_density_types	Number of content words types divided by number of total number of types.
lsa_all_fwd	LSA Cosine Score (entire text to source text) (frequent words dominate)
lsa_all_rwd	LSA Cosine Score (entire text to source text) (rare words dominate)
lsa_2_all_sent	Average latent semantic analysis cosine similarity between all adjacent sentences (with a two-sentence span).
lsa_2-3_rwd	Latent Semantic Analysis Cosine Score (second third of text to source text) (rare words dominate)
mtld_ma_bi_aw	The final value is the mean score when MTLD is run forwards and backwards (but partial factors are not calculated)
mtld_original_aw	MTLD is based on the average number of tokens it takes to reach a given TTR value (.720).
mtld_original_cw	MTLD is based on the average number of content word tokens it takes to reach a given TTR value (.720).
nsentences	Number of sentences
nwords	Number of words
Ortho_N_CW	Number of orthographic neighbors (content words)
PLD	Mean phonological Levenshtein distances from a word to its 20 closest neighbors
PLDF_CW	Mean log HAL frequency of a word's 20 closest neighbors; neighbors determined using phonological Levenshtein distance
positive_logical	Positive logical connectives; number of positive logical connectives
root_ttr_fw	Number of function word types divided by the square root of the number of function word tokens.
WN_Mean_RT	Word Naming Response Time; Mean naming reaction time in milliseconds across all participants for this word.

References

1. Han, C.; Lu, X. Can automated machine translation evaluation metrics be used to assess students' interpretation in the language learning classroom? *Comput. Assist. Lang. Learn.* **2023**, *36*, 1064–1087. [CrossRef]
2. Lee, A.V.Y.; Luco, A.C.; Tan, S.C. A human-centric automated essay scoring and feedback system for the development of ethical reasoning. *Educ. Technol. Soc.* **2023**, *26*, 147–159.
3. Shin, J.; Gierl, M.J. More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms. *Lang. Test.* **2021**, *38*, 247–272. [CrossRef]
4. Crossley, S.A.; Kyle, K.; McNamara, D.S. To aggregate or not? Linguistic features in automatic essay scoring and feedback systems. *J. Writ. Assess.* **2015**, *8* . Available online: www.journalofwritingassessment.org/article.php?article=80 (accessed on 11 May 2024).
5. Zhang, X.; Lu, X.; Li, W. Beyond differences: Assessing effects of shared linguistic features on L2 writing quality of two genres. *Appl. Linguist.* **2022**, *43*, 168–195. [CrossRef]
6. Beseiso, M.; Alzubi, O.A.; Rashaideh, H. A novel automated essay scoring approach for reliable higher educational assessments. *J. Comput. High. Educ.* **2021**, *33*, 727–746. [CrossRef]

7. Mizumoto, A.; Eguchi, M. Exploring the potential of using an AI language model for automated essay scoring. *Res. Methods Appl. Linguist.* **2023**, *2*, 100050. [CrossRef]
8. Kumar, V.S.; Boulanger, D. Automated essay scoring and the deep learning black box: How are rubric scores determined? *Int. J. Artif. Intell. Educ.* **2021**, *31*, 538–584. [CrossRef]
9. Ramesh, D.; Sanampudi, S.K. An automated essay scoring systems: A systematic literature review. *Artif. Intell. Rev.* **2022**, *55*, 2495–2527. [CrossRef]
10. Latifi, S.; Gierl, M. Automated scoring of junior and senior high essays using Coh-Metrix features: Implications for large-scale language testing. *Lang. Test.* **2021**, *38*, 62–85. [CrossRef]
11. Ramnarain-Seetohul, V.; Bassoo, V.; Rosunally, Y. Similarity measures in automated essay scoring systems: A ten-year review. *Educ. Inf. Technol.* **2022**, *27*, 5573–5604. [CrossRef]
12. Uto, M. A review of deep-neural automated essay scoring models. *Behaviormetrika* **2021**, *48*, 459–484. [CrossRef]
13. Uto, M.; Okano, M. Learning automated essay scoring models using item-response-theory-based scores to decrease effects of rater biases. *IEEE Trans. Learn. Technol.* **2021**, *14*, 763–776. [CrossRef]
14. Zhang, Z.V. Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assess. Writ.* **2020**, *43*, 100439. [CrossRef]
15. Crossley, S.A.; Allen, L.K.; Snow, E.L.; McNamara, D.S. Incorporating learning characteristics into automatic essay scoring models: What individual differences and linguistic features tell us about writing quality. *J. Educ. Data Min.* **2016**, *8*, 1–19.
16. Lee, C.; Ge, H.; Chung, E. What linguistic features distinguish and predict L2 writing quality? A study of examination scripts written by adolescent Chinese learners of English in Hong Kong. *System* **2021**, *97*, 102461. [CrossRef]
17. MacArthur, C.A.; Jennings, A.; Philippakos, Z.A. Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction? *Read. Writ.* **2019**, *32*, 1553–1574. [CrossRef]
18. Kumar, V.; Boulanger, D. Explainable automated essay scoring: Deep learning really has pedagogical value. In *Frontiers in Education*; Frontiers Media SA: Lausanne, Switzerland, 2020; Volume 5, p. 572367.
19. Crossley, S.; McNamara, D. Text coherence and judgments of essay quality: Models of quality and coherence. In Proceedings of the Annual Meeting of the Cognitive Science Society, Boston, MA, USA, 20–23 July 2011; Volume 33.
20. Taylor, K.S.; Lawrence, J.F.; Connor, C.M.; Snow, C.E. Cognitive and linguistic features of adolescent argumentative writing: Do connectives signal more complex reasoning? *Read. Writ.* **2019**, *32*, 983–1007. [CrossRef]
21. Vögelin, C.; Jansen, T.; Keller, S.D.; Machts, N.; Möller, J. The influence of lexical features on teacher judgements of ESL argumentative essays. *Assess. Writ.* **2019**, *39*, 50–63. [CrossRef]
22. Barkaoui, K. Explaining ESL essay holistic scores: A multilevel modeling approach. *Lang. Test.* **2010**, *27*, 515–535. [CrossRef]
23. Becker, A. Distinguishing linguistic and discourse features in ESL students' written performance. *Mod. J. Appl. Linguist.* **2010**, *2*, 406–424.
24. Li, Z.; Link, S.; Ma, H.; Yang, H.; Hegelheimer, V. The role of automated writing evaluation holistic scores in the ESL classroom. *System* **2014**, *44*, 66–78. [CrossRef]
25. McNamara, D.S.; Crossley, S.A.; Roscoe, R. Natural language processing in an intelligent writing strategy tutoring system. *Behav. Res. Methods* **2013**, *45*, 499–515. [CrossRef] [PubMed]
26. Burstein, J.; Tetreault, J.; Madnani, N. The e-rater automated essay scoring system. In *Handbook of Automated Essay Evaluation: Current Applications and New Directions*; Routledge: Oxford, UK, 2013; pp. 55–67.
27. Schultz, M.T. The IntelliMetric™ automated essay scoring engine–A review and an application to Chinese essay scoring. In *Handbook of Automated Essay Evaluation*; Routledge: Oxford, UK, 2013; pp. 89–98.
28. Abdi, H.; Williams, L.J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433–459. [CrossRef]
29. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the Advances in Neural Information Processing Systems;* Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Glasgow, UK, 2017; Volume 30.
30. Camp, H. The psychology of writing development—And its implications for assessment. *Assess. Writ.* **2012**, *17*, 92–105. [CrossRef]
31. Condon, W. Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assess. Writ.* **2013**, *18*, 100–108. [CrossRef]
32. Deane, P. On the relation between automated essay scoring and modern views of the writing construct. *Assess. Writ.* **2013**, *18*, 7–24. [CrossRef]
33. Weigle, S.C. *Assessing Writing*; Ernst Klett Sprachen: Stuttgart, Germany, 2002.
34. Zheng, Y.; Yu, S. What has been assessed in writing and how? Empirical evidence from Assessing Writing (2000–2018). *Assess. Writ.* **2019**, *42*, 100421. [CrossRef]
35. Bennett, C. Assessment rubrics: Thinking inside the boxes. *Learn. Teach.* **2016**, *9*, 50–72. [CrossRef]
36. Saxton, E.; Belanger, S.; Becker, W. The Critical Thinking Analytic Rubric (CTAR): Investigating intra-rater and inter-rater reliability of a scoring mechanism for critical thinking performance assessments. *Assess. Writ.* **2012**, *17*, 251–270. [CrossRef]
37. Chan, Z.; Ho, S. Good and bad practices in rubrics: The perspectives of students and educators. *Assess. Eval. High. Educ.* **2019**, *44*, 533–545. [CrossRef]

38. Hodges, T.S.; Wright, K.L.; Wind, S.A.; Matthews, S.D.; Zimmer, W.K.; McTigue, E. Developing and examining validity evidence for the Writing Rubric to Inform Teacher Educators (WRITE). *Assess. Writ.* **2019**, *40*, 1–13. [CrossRef]
39. Knoch, U. Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assess. Writ.* **2011**, *16*, 81–96. [CrossRef]
40. Li, H.; He, L. A comparison of EFL raters' essay-rating processes across two types of rating scales. *Lang. Assess. Q.* **2015**, *12*, 178–212. [CrossRef]
41. Carr, N.T. A comparison of the effects of analytic and holistic rating scale types in the context of composition tests. *Issues Appl. Linguist.* **2000**, *11* . [CrossRef]
42. Cumming, A. Expertise in evaluating second language compositions. *Lang. Test.* **1990**, *7*, 31–51. [CrossRef]
43. Olinghouse, N.G.; Santangelo, T.; Wilson, J. Examining the validity of single-occasion, single-genre, holistically scored writing assessments. In *Measuring Writing: Recent Insights into Theory, Methodology and Practice*; Brill: Leiden, The Netherlands, 2012; pp. 55–82.
44. White, E.M. Holisticism. *Coll. Compos. Commun.* **1984**, *35*, 400–409. [CrossRef]
45. Harsch, C.; Martin, G. Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assess. Educ. Princ. Policy Pract.* **2013**, *20*, 281–307. [CrossRef]
46. Huot, B. The literature of direct writing assessment: Major concerns and prevailing trends. *Rev. Educ. Res.* **1990**, *60*, 237–263. [CrossRef]
47. Hyland, K. *Second Language Writing*; Cambridge University Press: Cambridge, UK, 2019.
48. Wind, S.A. Do raters use rating scale categories consistently across analytic rubric domains in writing assessment? *Assess. Writ.* **2020**, *43*, 100416. [CrossRef]
49. Liu, Y.; Huang, J. The quality assurance of a national English writing assessment: Policy implications for quality improvement. *Stud. Educ. Eval.* **2020**, *67*, 100941. [CrossRef]
50. Golparvar, S.E.; Abolhasani, H. Unpacking the contribution of linguistic features to graph writing quality: An analytic scoring approach. *Assess. Writ.* **2022**, *53*, 100644. [CrossRef]
51. Imbler, A.C.; Clark, S.K.; Young, T.A.; Feinauer, E. Teaching second-grade students to write science expository text: Does a holistic or analytic rubric provide more meaningful results? *Assess. Writ.* **2023**, *55*, 100676. [CrossRef]
52. Ohta, R.; Plakans, L.M.; Gebril, A. Integrated writing scores based on holistic and multi-trait scales: A generalizability analysis. *Assess. Writ.* **2018**, *38*, 21–36. [CrossRef]
53. Crossley, S.A.; Kyle, K.; Dascalu, M. The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behav. Res. Methods* **2019**, *51*, 14–27. [CrossRef]
54. Eid, S.M.; Wanas, N.M. Automated essay scoring linguistic feature: Comparative study. In Proceedings of the 2017 Intl Conf on Advanced Control Circuits Systems (ACCS) Systems & 2017 Intl Conf on New Paradigms in Electronics & Information Technology (PEIT), Alexandria, Egypt, 5–8 November 2017; pp. 212–217.
55. Guo, L.; Crossley, S.A.; McNamara, D.S. Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assess. Writ.* **2013**, *18*, 218–238. [CrossRef]
56. Kyle, K.; Crossley, S.A. Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *Mod. Lang. J.* **2018**, *102*, 333–349. [CrossRef]
57. Tian, Y.; Kim, M.; Crossley, S.; Wan, Q. Cohesive devices as an indicator of L2 students' writing fluency. *Read. Writ.* **2021**, *37*, 1–23. [CrossRef]
58. Weigle, S.C. English language learners and automated scoring of essays: Critical considerations. *Assess. Writ.* **2013**, *18*, 85–99. [CrossRef]
59. Crossley, S.A. Linguistic features in writing quality and development: An overview. *J. Writ. Res.* **2020**, *11*, 415–443. [CrossRef]
60. McNamara, D.S.; Crossley, S.A.; McCarthy, P.M. Linguistic features of writing quality. *Writ. Commun.* **2010**, *27*, 57–86. [CrossRef]
61. Crossley, S.A.; Weston, J.L.; McLain Sullivan, S.T.; McNamara, D.S. The development of writing proficiency as a function of grade level: A linguistic analysis. *Writ. Commun.* **2011**, *28*, 282–311. [CrossRef]
62. Crossley, S.; Cai, Z.; Mcnamara, D.S. Syntagmatic, paradigmatic, and automatic N-gram approaches to assessing essay quality. In Proceedings of the Twenty-Fifth International FLAIRS Conference, Marco Island, FL, USA, 23–25 May 2012.
63. Douglas, S.R. The Lexical Breadth of Undergraduate Novice Level Writing Competency. *Can. J. Appl. Linguist. Can. Linguist. Appliquée* **2013**, *16*, 152–170.
64. Goh, T.T.; Sun, H.; Yang, B. Microfeatures influencing writing quality: The case of Chinese students' SAT essays. *Comput. Assist. Lang. Learn.* **2020**, *33*, 455–481. [CrossRef]
65. Kettunen, K. Can type-token ratio be used to show morphological complexity of languages? *J. Quant. Linguist.* **2014**, *21*, 223–245. [CrossRef]
66. Kim, M.; Crossley, S.A.; Kyle, K. Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *Mod. Lang. J.* **2018**, *102*, 120–141. [CrossRef]
67. McCarthy, P.M.; Jarvis, S. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behav. Res. Methods* **2010**, *42*, 381–392. [CrossRef] [PubMed]
68. Crossley, S.A.; McNamara, D.S. Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *J. Res. Read.* **2012**, *35*, 115–135. [CrossRef]

69. Eckstein, G.; Ferris, D. Comparing L1 and L2 texts and writers in first-year composition. *Tesol Q.* **2018**, *52*, 137–162. [CrossRef]
70. Ferris, D.R. Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *Tesol Q.* **1994**, *28*, 414–420. [CrossRef]
71. Gómez Vera, G.; Sotomayor, C.; Bedwell, P.; Domínguez, A.M.; Jéldrez, E. Analysis of lexical quality and its relation to writing quality for 4th grade, primary school students in Chile. *Read. Writ.* **2016**, *29*, 1317–1336. [CrossRef]
72. Grant, L.; Ginther, A. Using computer-tagged linguistic features to describe L2 writing differences. *J. Second. Lang. Writ.* **2000**, *9*, 123–145. [CrossRef]
73. Jarvis, S.; Grant, L.; Bikowski, D.; Ferris, D. Exploring multiple profiles of highly rated learner compositions. *J. Second. Lang. Writ.* **2003**, *12*, 377–403. [CrossRef]
74. Yang, Y.; Yap, N.T.; Ali, A.M. Predicting EFL expository writing quality with measures of lexical richness. *Assess. Writ.* **2023**, *57*, 100762. [CrossRef]
75. Benson, B.J.; Campbell, H.M. Assessment of student writing with curriculum-based measurement. In *Instruction and Assessment for Struggling Writers: Evidence-Based Practices*; Guilford Press: New York, NY, USA, 2009; pp. 337–353.
76. Myhill, D. Towards a linguistic model of sentence development in writing. *Lang. Educ.* **2008**, *22*, 271–288. [CrossRef]
77. Crossley, S.A.; Roscoe, R.; McNamara, D.S. Predicting human scores of essay quality using computational indices of linguistic and textual features. In Proceedings of the International Conference on Artificial Intelligence in Education, Auckland, New Zealand, 28 June–2 July 2011; Springer: Berlin/Heidelberg, Germany, 2011; pp. 438–440.
78. Perin, D.; Lauterbach, M.; Raufman, J.; Kalamkarian, H.S. Text-based writing of low-skilled postsecondary students: Relation to comprehension, self-efficacy and teacher judgments. *Read. Writ.* **2017**, *30*, 887–915. [CrossRef]
79. Connor, U. Linguistic/rhetorical measures for international persuasive student writing. *Res. Teach. Engl.* **1990**, *24*, 67–87.
80. Aull, L.L.; Lancaster, Z. Linguistic markers of stance in early and advanced academic writing: A corpus-based comparison. *Writ. Commun.* **2014**, *31*, 151–183. [CrossRef]
81. León, J.A.; Olmos, R.; Escudero, I.; Cañas, J.J.; Salmerón, L. Assessing short summaries with human judgments procedure and latent semantic analysis in narrative and expository texts. *Behav. Res. Methods* **2006**, *38*, 616–627. [CrossRef]
82. Olmos, R.; León, J.A.; Jorge-Botana, G.; Escudero, I. New algorithms assessing short summaries in expository texts using latent semantic analysis. *Behav. Res. Methods* **2009**, *41*, 944–950. [CrossRef]
83. Hussein, M.A.; Hassan, H.; Nassef, M. Automated language essay scoring systems: A literature review. *PeerJ Comput. Sci.* **2019**, *5*, e208. [CrossRef] [PubMed]
84. Rupp, A.A.; Casabianca, J.M.; Krüger, M.; Keller, S.; Köller, O. Automated essay scoring at scale: A case study in Switzerland and Germany. *ETS Res. Rep. Ser.* **2019**, *2019*, 1–23. [CrossRef]
85. Shermis, M.D. State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assess. Writ.* **2014**, *20*, 53–76. [CrossRef]
86. Page, E.B. The imminence of... grading essays by computer. *Phi Delta Kappan* **1966**, *47*, 238–243.
87. Page, E.B. The use of the computer in analyzing student essays. *Int. Rev. Educ.* **1968**, *14*, 210–225. [CrossRef]
88. Kukich, K. Beyond automated essay scoring, the debate on automated essay grading. *IEEE Intell. Syst.* **2000**, *15*, 22–27.
89. Shermis, M.D.; Koch, C.M.; Page, E.B.; Keith, T.Z.; Harrington, S. Trait ratings for automated essay grading. *Educ. Psychol. Meas.* **2002**, *62*, 5–18. [CrossRef]
90. Valenti, S.; Neri, F.; Cucchiarelli, A. An overview of current research on automated essay grading. *J. Inf. Technol. Educ. Res.* **2003**, *2*, 319–330. [CrossRef] [PubMed]
91. Landauer, T.K.; Laham, D.; Foltz, P.W. Automated scoring and annotation of essays with the Intelligent Essay Assessor. In *Automated Essay Scoring: A Cross-Disciplinary Perspective*; Shermis, M.D., Burstein, J.C., Eds.; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2003; pp. 87–112
92. Foltz, P.W.; Kintsch, W.; Landauer, T.K. The measurement of textual coherence with latent semantic analysis. *Discourse Process.* **1998**, *25*, 285–307. [CrossRef]
93. Landauer, T.K.; McNamara, D.S.; Dennis, S.; Kintsch, W. *Handbook of Latent Semantic Analysis*; Psychology Press: London, UK, 2013.
94. Li, H.; Cai, Z.; Graesser, A.C. Computerized summary scoring: Crowdsourcing-based latent semantic analysis. *Behav. Res. Methods* **2018**, *50*, 2144–2161. [CrossRef]
95. Attali, Y.; Burstein, J. Automated essay scoring with e-rater® V. 2. *J. Technol. Learn. Assess.* **2006**, *4*, 1–17. [CrossRef]
96. Ramineni, C.; Williamson, D. Understanding mean score differences between the e-rater® automated scoring engine and humans for demographically based groups in the GRE® general test. *ETS Res. Rep. Ser.* **2018**, *2018*, 1–31. [CrossRef]
97. Enright, M.K.; Quinlan, T. Complementing human judgment of essays written by English language learners with e-rater® scoring. *Lang. Test.* **2010**, *27*, 317–334. [CrossRef]
98. Elliot, S. IntelliMetric: From here to validity. In *Automated Essay Scoring: A Cross-Disciplinary Perspective*; Shermis, M.D., Burstein, J.C., Eds.; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2002; pp. 71–86.
99. Crossley, S.; Kyle, K.; Davenport, J.; McNamara, D.S. Automatic Assessment of Constructed Response Data in a Chemistry Tutor. In Proceedings of the 9th International Conference on Educational Data Mining, EDM 2016, Raleigh, NC, USA, 29 June–2 July 2016; pp. 336–340.

100. Kyle, K.; Crossley, S.A.; Jarvis, S. Assessing the validity of lexical diversity indices using direct judgements. *Lang. Assess. Q.* **2021**, *18*, 154–170. [CrossRef]
101. Kyle, K. Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication. Ph.D. Thesis, Georgia State University, Atlanta, GA, USA, 2016.
102. Crossley, S.A.; Kyle, K.; McNamara, D.S. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behav. Res. Methods* **2016**, *48*, 1227–1237. [CrossRef] [PubMed]
103. Crossley, S.A.; Bradfield, F.; Bustamante, A. Using human judgments to examine the validity of automated grammar, syntax, and mechanical errors in writing. *J. Writ. Res.* **2019**, *11*, 251–270. [CrossRef]
104. Landauer, T.K.; Foltz, P.W.; Laham, D. An introduction to latent semantic analysis. *Discourse Process.* **1998**, *25*, 259–284. [CrossRef]
105. Covington, M.A.; McFall, J.D. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *J. Quant. Linguist.* **2010**, *17*, 94–100. [CrossRef]
106. McCarthy, P.M.; Jarvis, S. vocd: A theoretical and empirical evaluation. *Lang. Test.* **2007**, *24*, 459–488. [CrossRef]
107. Lu, X. Automatic analysis of syntactic complexity in second language writing. *Int. J. Corpus Linguist.* **2010**, *15*, 474–496. [CrossRef]
108. Harris, C.W.; Kaiser, H.F. Oblique factor analytic solutions by orthogonal transformations. *Psychometrika* **1964**, *29*, 347–362. [CrossRef]
109. Cattell, R.B. The scree test for the number of factors. *Multivar. Behav. Res.* **1966**, *1*, 245–276. [CrossRef]
110. Hayton, J.C.; Allen, D.G.; Scarpello, V. Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organ. Res. Methods* **2004**, *7*, 191–205. [CrossRef]
111. MacKay, D.J. A practical Bayesian framework for backpropagation networks. *Neural Comput.* **1992**, *4*, 448–472. [CrossRef]
112. Phandi, P.; Chai, K.M.A.; Ng, H.T. Flexible domain adaptation for automated essay scoring using correlated linear regression. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 431–439.
113. Cook, R.D.; Weisberg, S. Linear and nonlinear regression. In *Statistical Methodology in the Pharmacological Sciences*; Marcel Dekker: New York, NY, USA, 1990; pp. 163–199.
114. Bach, F. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *J. Mach. Learn. Res.* **2014**, *15*, 595–627.
115. Roy, M.H.; Larocque, D. Robustness of random forests for regression. *J. Nonparametr. Stat.* **2012**, *24*, 993–1006. [CrossRef]
116. Xu, M.; Watanachaturaporn, P.; Varshney, P.K.; Arora, M.K. Decision tree regression for soft classification of remote sensing data. *Remote. Sens. Environ.* **2005**, *97*, 322–336. [CrossRef]
117. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. [CrossRef]
118. Chen, H.; He, B. Automated essay scoring by maximizing human-machine agreement. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Washington, DC, USA, 18–21 October 2013; pp. 1741–1752.
119. Zhao, S.; Zhang, Y.; Xiong, X.; Botelho, A.; Heffernan, N. A memory-augmented neural model for automated grading. In Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale, Cambridge, MA, USA, 20–21 April 2017; pp. 189–192.
120. Cozma, M.; Butnaru, A.M.; Ionescu, R.T. Automated essay scoring with string kernels and word embeddings. *arXiv* **2018**, arXiv:1804.07954.
121. Liu, J.; Xu, Y.; Zhu, Y. Automated essay scoring based on two-stage learning. *arXiv* **2019**, arXiv:1901.07744.
122. Dasgupta, T.; Naskar, A.; Dey, L.; Saha, R. Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, Melbourne, Australia, 19 July 2018; pp. 93–102.
123. Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **2023**, *103*, 102274. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.