

Article

An NLP-Based Exploration of Variance in Student Writing and Syntax: Implications for Automated Writing Evaluation

Maria Goldshtein ^{1,*}, Amin G. Alhashim ²  and Rod D. Roscoe ^{1,*} ¹ Human Systems Engineering, Arizona State University, Mesa, AZ 85212, USA² Mathematics, Statistics, and Computer Science, Macalester College, Saint Paul, MN 55105, USA; aalhashi@macalester.edu

* Correspondence: maria.goldshtein@asu.edu (M.G.); rod.roscoe@asu.edu (R.D.R.)

Abstract: In writing assessment, expert human evaluators ideally judge individual essays with attention to variance among writers' syntactic patterns. There are many ways to compose text successfully or less successfully. For automated writing evaluation (AWE) systems to provide accurate assessment and relevant feedback, they must be able to consider similar kinds of variance. The current study employed natural language processing (NLP) to explore variance in syntactic complexity and sophistication across clusters characterized in a large corpus ($n = 36,207$) of middle school and high school argumentative essays. Using NLP tools, k-means clustering, and discriminant function analysis (DFA), we observed that student writers employed four distinct syntactic patterns: (1) familiar and descriptive language, (2) consistently simple noun phrases, (3) variably complex noun phrases, and (4) moderate complexity with less familiar language. Importantly, each pattern spanned the full range of writing quality; there were no syntactic patterns consistently evaluated as "good" or "bad". These findings support the need for nuanced approaches in automated writing assessment while informing ways that AWE can participate in that process. Future AWE research can and should explore similar variability across other detectable elements of writing (e.g., vocabulary, cohesion, discursive cues, and sentiment) via diverse modeling methods.

Keywords: automated writing evaluation; natural language processing; student writing variability; syntax; writing styles



Citation: Goldshtein, M.; Alhashim, A.G.; Roscoe, R.D. An NLP-Based Exploration of Variance in Student Writing and Syntax: Implications for Automated Writing Evaluation. *Computers* **2024**, *13*, 160. <https://doi.org/10.3390/computers13070160>

Academic Editor: Ming Liu

Received: 24 May 2024

Revised: 12 June 2024

Accepted: 19 June 2024

Published: 25 June 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

Writing and written expression are almost infinitely variable. There are numerous techniques for communicating our ideas [1,2], and authors may demonstrate flexibility in meeting their discursive goals [3]. Importantly, variance in writing is not merely the product of rhetorical decision making, but also emerges from the conscious and unconscious knowledge, styles, preferences, and cultures of the authors [4–6]. Such variations complicate writing assessment because there are many ways to “succeed” [7]. Training and well-defined rubrics offer structure that draws expert human evaluators' attention to key features and variations [8,9], although evaluators may also possess implicit biases that color their perceptions of student writing [10,11]. These demands exacerbate the already substantial workload of writing assessment. Educators understand that offering frequent writing assignments, deliberate practice, and formative feedback are crucial for writing and intellectual development [12], but enacting these goals stresses constrained instructor resources.

Numerous automated writing evaluation (AWE) technologies now exist to facilitate the aforementioned assessment tasks. AWE tools ostensibly make educators' jobs easier by automating assignment management, summative and formative assessment and feedback, and more [13,14]. Indeed, these technologies can evaluate large numbers of essays with

speed, consistency, and accuracy [15–19], and deliver actionable recommendations for students to revise and improve [20–25]. Decades of research have shown that these admittedly fallible tools have value in writing instruction [24,26–29].

We contend that a valuable and necessary opportunity to improve AWE technologies is to further explore the issue of variance. Automation relies upon predetermined (i.e., algorithmic) evaluative processes and metrics, which are driven by similarly predetermined expectations about “good” versus “poor” writing. AWE systems can only “reward” and provide feedback on aspects of writing that they have been designed to detect and recognize as worthy. Valid critiques of AWE have thus noted that AWE tools may promote constrained writing norms, contexts, and processes [30–33]. Compared to human evaluators, automated systems have limited access to contextual information about students as whole persons. In classrooms, teachers might possess a deeper understanding of their students’ diverse assets and needs, which they could flexibly consider when teaching or assessing writing [14,34]. AWE technologies may provide less appropriate or personalized assessment and assessment because they lack human empathy, inferencing, and interpersonal knowledge [35–37]. AWE systems must be (re)designed to examine and account for variance in student writing.

In this paper, we attend to the variability of syntactic sophistication and complexity within student writing to (a) affirm the reality of variance and (b) demonstrate an approach for addressing this variance in AWE using natural language processing (NLP). We acknowledge that syntax is only one component of writing [38]. However, a focused inspection of one component is useful for encouraging others to explore similar and more expansive lines of work. In the following sections, we further discuss the importance of acknowledging variance in student writing and AWE. We then explore clusters of syntactic variation within a large corpus of student essays using NLP indices. Finally, we consider the implications of this approach and findings for AWE.

1.1. Recognizing Writing Variance in Automated Writing Evaluation

The development of AWE algorithms employs an ever-expanding toolbox of methods spanning simple correlations, linear regressions, machine learning, and neural networks [19]. Regardless of the specific methodology, the process typically begins with “training data” texts that have been assessed by human raters. Such ratings may be holistic (e.g., overall quality), specific subscales (e.g., organization and content, register, and genre), and include annotated features (e.g., rhetorical moves). Next, NLP tools extract linguistic properties of the texts, ranging from descriptive features (e.g., number of words and average sentence length) to more fine-grained calculation (e.g., average number of adjectives per noun phrase). Finally, statistical methods (e.g., regression and machine learning) are implemented to map human-assigned ratings to sets of NLP metrics. These predictive relationships form the basis for AWE algorithms; patterns of NLP-derived features are interpreted as reliable and valid indicators of writing characteristics.

Although these methods have generated numerous accurate algorithms, they might neglect variance in several ways. First, algorithms can only explicitly attend to features when (a) detectors are present and (b) those indices are included in assessment models. For instance, to examine vocabulary, NLP tools might rely on measures of average word concreteness, age-of-acquisition, specificity, familiarity, and more [39]. However, other properties may be inaccessible (e.g., personal emotional associations) and thus unusable. Likewise, metrics might be excluded from algorithms if initial analyses reveal “no statistically significant relationship” to human-assigned ratings. When metrics are missing or excluded, resulting algorithms cannot be readily sensitive to variance associated with those features.

Another neglect of variance may occur when algorithms do not account for nested or contextualized patterns. For example, essays naturally vary in length, but the meaning of length may depend on the task, environment, or writer. When prompts ask for “a brief explanation”, then short essays are perfectly reasonable; prompts that request “detailed

exploration” might warrant a longer essay. Similarly, students may write more when given ample time but write less under artificial time constraints. Students’ prior knowledge, motivation, life experiences, and strategies also influence how much they write about a topic regardless of their actual ability to produce text [4]. Finally, optimal text length may vary based on other features like vocabulary, syntax, and cohesion. Skilled writers may use precise word choices to convey ideas concisely. By contrast, knowledgeable writers may include ample details and elaboration that make an essay longer [40]. Thus, the interpretation of any given feature as an indicator of “quality” may be nested within the variance of other features.

In the current paper, we focus attention on third aspect of variance: student writers can enact their skills in different ways [7,40–42]. In AWE, typical assumptions conceptualize the variance of holistic quality or multiple dimensions on linear continua from “poor” (or “low” or “weak”) to “good” (or “high” or “strong”). Thus, student essays might be rated as having “poor logical flow” versus demonstrating “a clear flow of ideas and arguments”, or may be described as showing “unsophisticated word choices” versus “skilled command of vocabulary”. However, as noted above, writers might achieve success by leveraging very different kinds or combinations of rhetorical or vocabulary strategies. Which essays are “better”? Which writers are “more skilled”?

Several prior studies have quantitatively documented writing variance along multiple lexical and grammatical dimensions (e.g., Refs. [1,30,41,43–46]). Via multidimensional analysis (MDA), Deane & Quinlan [30] observed three distinct styles in a large corpus of student essays, including (1) fluent writing, appropriate syntactic complexity, and appropriate elaboration of text structure; (2) appropriately adapting vocabulary to academic style, frequent grammatical patterns; and (3) maintaining conventional patterns of grammar, mechanics, and usage. Friginal & Weigle [44] observed four stylistic dimensions within a corpus of 207 L2-English essays. These dimensions were characterized as (1) Involved vs. Informational Focus, (2) Addressee-Focused Description vs. Personal Narrative, (3) Simplified vs. Elaborated Description, and (4) Personal Opinion vs. Impersonal Evaluation/Assessment. In research on writing across disciplines, Gardner and colleagues [41] characterized four distinct clusters based on lexical and grammatical features. Overall, quantitative analyses have clearly established significant variance in syntactic structure, register, vocabulary use, and more in writing.

Crossley and colleagues [7] similarly employed cluster analysis and discriminant function analysis (DFA) to identify distinct profiles of “successful” student writing using diverse NLP indices. Specifically, they first constructed a corpus of 148 “successful” essays (i.e., human-assigned scores of 4.5 or better on a linear 6-point scale). Next, nearly 200 NLP indices were extracted via Coh-Metrix and related tools [47,48] spanning lexical, syntactic, cohesive, structural, semantic, and rhetorical features. Hierarchical cluster analyses were conducted to reveal distinct groupings and DFA was used to characterize those groups based on NLP measures. Within this small sample, the researchers observed four patterns of successful writing: (1) action and depiction, (2) academic, (3) accessible, and (4) lexical. Essays in the sample were able to achieve high scores via more descriptive language, academic language, accessible and cohesive language, or more skillful vocabulary usage, respectively. Such findings—derived from ostensibly linear human ratings—argue against straightforward or linear mapping between writing features, styles, and quality.

An important implication of [7] and similar work [30,41] is that AWE can feasibly address greater variance in student writing. Crossley and colleagues utilized NLP metrics to characterize essays—the same indices and tools that underlie several AWE systems (e.g., Writing Pal; [49–53]). Additional human corpus judgments or annotations were not required. However, one limitation was that this work focused on a small sample of only high-scoring essays; their analyses characterized only a few successful writers. There is value in extending that work by considering a larger pool of student authors and wider range of quality. Scores are also only one window into the variance of student writing. We

argue that it is valuable to first examine variance in how students write before constraining such patterns within specific “quality” expectations.

1.2. A Focus on Syntax

The current exploration focuses on syntax, which refers to how words (and word units) are combined, structured, and sequenced to produce larger units of meaning (e.g., clauses), and eventually entire sentences [54–56]. Syntax is often linked to grammar, which reflects the rules by which linguistic units are “allowed” to be combined or transformed (e.g., verb conjugation). Notably, the current study is not concerned with grammatical errors or “typos”, but rather the overall sophistication and complexity with which students construct their sentences. We acknowledge that syntax is only one component of writing, which also comprises lexical, semantic, rhetorical, pragmatic, and other dimensions. Conceptually, however, syntax operates at a level of language (see [57–59]) that connects lexical and discursive features, thus making it a meaningful and feasible target for this work.

Syntax is one of the central components of language in general and writing in particular. Syntactic behaviors and patterns have been shown to be related with writing quality as measured by academic evaluation and scoring (e.g., [3,60,61]). A focus on syntax is also motivated by prior research demonstrating the capacity for assessing syntax via NLP. In the current study, we employ the Tool for the Automated Assessment of Syntactic Sophistication and Complexity (TAASC) developed and validated by [46,62,63]. In that work, [64] (p. 8) has defined syntactic complexity and sophistication as follows:

“Syntactic complexity refers to the formal characteristics of syntax (e.g., the amount of subordination) [. . .]. In contrast, syntactic sophistication refers to the relative difficulty of learning particular syntactic structures [. . .], which (from a usage-based perspective) is related to input frequency and contingency. The term sophistication [. . .] refers to less frequent words as more sophisticated because they tend to be produced by more proficient writers”

More generally, syntax is a central component in efforts to automate writing evaluation through NLP (e.g., [65]). For example, Jagaiah and colleagues [54] examined 36 studies on syntactic complexity measures and found variance in syntactic complexity measures across genres, but also variance across individuals within those groupings. Similarly, Kyle and Crossley [62] observed that incorporating usage-based measures (e.g., frequency of verb argument constructions) helped explain variance in L2 authors’ writing quality scores. The authors proposed that these measures should be incorporated into the automated assessment of syntactic complexity, as part of the automation of writing evaluation as a whole. This research was expanded upon in [64], which focuses on the developmental trajectories of L2 writers from the same usage-based perspective, through indices of verb argument construction sophistication. Findings show a trajectory of improvement in writing (as measured by scores) over the course of two years, which is correlated with changes in syntactic complexity and verb argument construction sophistication measures (e.g., number of dependent clauses per clause and main verb frequency). These findings illustrate the importance of syntactic complexity and sophistication to writing outcomes.

Prior studies with TAASC have observed meaningful relationships between syntactic complexity and L1/L2 writing quality [54,61], lexical diversity [62], and writing development [64]. In general, syntax is a central component in efforts to automate academic writing evaluation through NLP [65]. NLP algorithms map various writing properties of existing essay corpora (e.g., structures, word frequency, meaning, and relationship to prompt) to human evaluations. Through this process of inference, algorithms produce mappings of writing patterns to evaluate behaviors of human evaluators. For example, [66] used a combination of syntactic, semantic, and sentiment related features of essay writing to estimate essay quality. Syntactic features (e.g., unique parts-of-speech used, sentence length, and words ending with “-ing”), helped the reported model achieve significant agreement with human raters (QWP (Quadrating Weighted Kappa) = 0.793).

1.3. Research Questions

The current study is driven by three research questions embedded within overarching considerations for AWE development and implementation. To answer these questions, we took direct inspiration from [7] to explore potential patterns using cluster analysis that are then characterized using DFA.

1. What variance do student writers display with regards to syntactic sophistication and complexity? AWE algorithms typically capture syntactic variance by a single dimension that varies linearly from “lower” to “higher” sophistication and complexity. In truth, students may enact complexity in different ways revealed by variance in syntactic features or constructions;
2. What are the primary linguistic features (i.e., NLP metrics) that characterize patterns of syntactic sophistication and complexity? If there are multiple patterns or profiles, these patterns should exhibit distinct defining characteristics. Subsequently, these patterns of NLP metrics could directly inform future AWE algorithms;
3. How does variance in syntactic sophistication and complexity relate to writing quality (i.e., human-assigned scores)? The purpose of this research is neither to analyze how syntactic measures predict writing quality nor to create AWE scoring algorithms. Nonetheless, it is useful to consider how variance in syntactic patterns is associated with variance in writing quality, if at all. Greater sophistication and complexity may be predictive of higher quality. Similarly, if distinct patterns are observed, certain patterns may be associated with higher quality, whereas others are associated with lower quality. Alternatively, variance in writing quality may be observed across all patterns. That is, observed patterns may represent truly distinct ways of writing that can each be navigated successfully or unsuccessfully. These generalizations may be applicable to development AWE algorithms that are responsive or personalized to students’ variability in writing.

2. Method

2.1. Essay Corpus and Preparation

2.1.1. Initial Corpus

The initial corpus comprised 39,511 argument essays collected by a state-level education agency in the United States for standardized testing [67]. Argument essays are a common format for assessing students’ writing and rhetorical skills wherein students craft a persuasive response to a prompt (e.g., [63,68]). All essays were composed in response to one of five topics (i.e., driverless cars, exploring Venus, facial action coding, the face on Mars, or seagoing cowboys). Essays were assigned holistic scores by trained human raters on a scale of “1” (lowest quality) to “6” (highest). The writers were 6th-, 8th-, and 10th grade students. Limited demographic data included “gender” (reported as binary “female” or “male”), “race/ethnicity” (reported as American Indian/Alaska Native, Asian/Pacific Islander, Black/African American, Hispanic/Latino, White, Two or More Race/Other), “economic disadvantage” (reported as binary “no” or “yes”), “disability status” (reported as binary “no” or “yes”), and “English language learner” (reported as binary “no” or “yes”).

2.1.2. Syntactic NLP Features

Linguistic features related to syntax were extracted using the *Tool for the Automatic Analysis of Syntactic Sophistication and Complexity* (TAASSC; Version 1.3.8) [62,63]. TAASSC included 355 indices and component scores pertaining to *clause complexity*, *noun phrase complexity*, and *syntactic sophistication*.

First, clauses are sentence components that comprise a subject and predicate but may not constitute a complete sentence on their own. Specifically, independent clauses may stand alone as complete sentences (e.g., “The red carpet added color to the room”) that vary in complexity based on additional nouns, adjectives, and other details (i.e., complements). Dependent clauses modify other components within a sentence; they depend on the presence of another independent clause (e.g., “The interior decorator decided

that the red carpet added color to the room”). Dependent clauses add complexity. An *adjective complement* modifies or adds information to an adjective within the clause. Similarly, a *nominal complement* modifies or adds information to noun within the clause. Both kinds of complements can increase specificity and clarity, but too many can contribute to sentence processing difficulties. Examples (1) and (2) below represent adjective complements (in brackets, with the adjective) that increase clarity or make the sentence harder to parse, respectively:

1. Anna is [delighted with her new job].
2. Anna is [delighted that the people who interviewed her last week have made an offer and the salary is what she had hoped for].

Examples (3) and (4) below demonstrate nominal complements that make a sentence more specific or harder to parse, respectively:

3. Ryan is a [teacher of Portuguese].
4. Ryan is a [teacher who really likes doing fun activities and creating fun lesson plans for his students every semester].

Second, noun phrases or “nominals” are linguistic units wherein a focal noun (e.g., “carpet”) is described or modified by other words (e.g., “red” or “on the floor”), but the entire phrase serves the same grammatical role as the noun (e.g., “the red carpet on the floor”). The simplest noun phrases may comprise only the noun; more complex noun phrases may also incorporate objects, adjectives, adverbs, dependents, prepositional relations, and other details that add information, nuance, and context.

Finally, syntactic sophistication may also be developed based on the use of less common vocabulary, phrases, and sentence constructions. Uncommon words (e.g., “vermillion” or “carmine”) are harder to understand and parse than familiar words (e.g., “red”) and thus add complexity. The same is true for phrases and sentence constructions. In TAASSC, the typicality of word roots (i.e., lemmas) and sentence constructions are assessed based on their frequency ratings in the Corpus of Contemporary American English (COCA, [69]). Higher ratings indicate that a lemma or construction occurs more frequently in the English language.

TAASSC has been productively implemented in numerous studies (e.g., [54,61,70,71]). For instance, [70] studied how neural networks and NLP tools (e.g., TAAASC) can reveal the contribution of linguistic features to rubric scores, and to explore what features are important in effective rubric scoring models. The researchers found that it was possible to train a model to produce a transparent grading rubric where the most predictive NLP properties were similar to human judgments. In a systematic analysis on measures of syntactic complexity, writing ability, and writing quality, Jagaiah and colleagues [54] observed a lack of straightforward connections between these constructs, in part due to a lack of research using the same metrics and the variance associated with these measures. Outside of the field of writing, Clarke and colleagues [72] explored the potential for syntax (and other metrics) to provide early indicators of Alzheimer’s disease. In sum, a growing body of literature has documented that TAASSC offers a reliable, valid, and meaningful tool for exploring the syntactic features and impact of text.

Notably, the more than 350 indices available through TAASSC include numerous redundant or highly correlated metrics—many metrics capture the same information in different ways. To reduce the number of indices used in the current analysis, we (a) reviewed the literature to identify metrics that demonstrated meaningful effects in prior studies and (b) examined metrics for multicollinearity (i.e., Pearson’s $r > 0.70$). This theory-driven and data-driven process identified 18 concrete indices to be used in the current study. Table 1 summarizes these metrics.

Table 1. List of 18 TAASSC indices analyzed in the current study.

Index Category (TAASSC Software Label)
Clause Complexity
1. Adjective Complements (avg) average number of adjective complements per clause (acomp_per_cl)
2. Nominal Complements (avg) average number of nominal complements per clause (ncomp_per_cl)
Noun Phrase Complexity
3. Dependents per Nominal (avg) average number of dependents per nominal (av_nominal_deps)
4. Dependents per Object (avg) average number of dependents per direct object (av_dobj_deps)
5. Dependents per Preposition (avg) average number of dependents per object of the preposition (av_pobj_deps)
6. Dependents per Nominal (stdev) dependents per nominal, standard deviation (nominal_deps_stdev)
7. Dependents per Subject (stdev) dependents per nominal subject, standard deviation (nsubj_stdev)
8. Dependents per Object (stdev) dependents per direct object, standard deviation (dobj_stdev)
9. Dependents per Preposition (stdev) dependents per object of the preposition, standard deviation (pobj_stdev)
10. Determiners per Nominal (avg) average number of determiners per nominal (det_all_nominal_deps_struct)
11. Prepositions per Nominal (avg) average number of prepositions per nominal (prep_all_nominal_deps_struct)
12. Adjectival Modifiers (avg) average number of adjectival modifiers per direct object (amod_dobj_deps_struct)
13. Prepositions per Preposition (avg) average number of prepositions per object of the preposition (prep_pobj_deps_struct)
Syntactic Sophistication
14. Lemmas (avg) average frequency of lemmas (all_av_lemma_freq)
15. Lemma Construction Combinations (avg) average frequency of lemma construction combinations (all_av_lemma_construction_freq)
16. Constructions (log) average frequency of constructions, log transform (all_av_construction_freq_log)
17. Constructions in Reference (prop) proportion of constructions in reference corpus (all_construction_attested)
18. Combinations in Reference (prop) prop. of lemma construction combinations in reference (all_lemma_construction_attested)

For measures of clause complexity and noun phrase complexity, higher values indicate a more complex structure (i.e., higher average frequency or larger variation). Measures of syntactic sophistication captured the use of common words and constructions. Thus, higher values on these metrics indicate simpler syntax and more familiar language.

2.1.3. Corpus Filtering and Analysis

Several steps were implemented to “clean” the corpus for analysis. First, essays that lacked accompanying demographic author data were excluded ($n = 1491$). Second, essays that generated two or more “0” scores on NLP indices were excluded ($n = 1119$). Inspection of these essays revealed textual details or errors (e.g., use of nonstandard notation or punctuation) that caused errors in the NLP tools and prevented analysis. Finally, a qualitative review of the data observed that many essays assigned a score of “1” by human raters were not valid attempts at authoring an essay (e.g., they comprised a single repeated word or highly off-topic commentary). To avoid skewed measurements and analyses,

we excluded essays with a score of “1” ($n = 694$). The final analysis corpus comprised 36,207 essays. Summary details for the analysis corpus are provided in Tables 2 and 3.

Table 2. Student information for the analyzed corpus ($n = 36,207$).

Student Information	<i>n</i> (% of Essays)
Grade	
6th	6212 (17.2%)
8th	8072 (22.3%)
10th	21,923 (60.5%)
Gender	
Male	17,659 (48.8%)
Female	18,548 (51.2%)
Race_Ethnicity	
American Indian/Alaskan Native	102 (0.3%)
Asian/Pacific Islander	689 (1.9%)
Black/African American	3353 (9.3%)
Hispanic/Latino	3073 (8.5%)
Two or more races/Other	1512 (4.2%)
White	27,478 (75.9%)
ELL	
Yes	1104 (3.0%)
No	35,103 (97.0%)
Disability	
Not identified as having disability	34,073 (94.1%)
Identified as having disability	2134 (5.9%)
Economic disadvantage	
Economically disadvantaged	15,305 (42.3%)
Not economically disadvantaged	20,902 (57.7%)

Table 3. Number of essays per score level in the analyzed corpus ($n = 36,207$).

Essay Score	<i>n</i> (% of Essays)
2	5003 (13.8%)
3	13,847 (38.2%)
4	13,356 (36.9%)
5	3453 (9.5%)
6	548 (1.5%)

2.2. Analysis

To address the primary research questions, two analytical methods were implemented. K-means clustering was used to identify potential syntactic patterns of student writing based on syntactic sophistication and complexity. Discrimination function analysis (DFA) was used to characterize resulting clusters based on patterns of predictive variables. Other approaches (e.g., MDA, [1,30,41,44,46]) are similarly informative for capturing variance in writing. MDA has been specifically applied to studying variance in register and genre, school writing [73], and in certain AWE settings [30], and writing evaluation [44]. More complex clustering methodologies such as hierarchical clustering (e.g., [74,75]) and random forest analysis (e.g., [76,77]) can also shed light and add nuance to analyses of variance in student writing behaviors. For the current work, we selected k-mean clustering and DFA due to their relative simplicity, accessibility, and speed. These methods enable exploration of clear patterns that can then drive more precise and detailed analyses. In addition, in taking inspiration from [7], we mirrored their methodology to facilitate comparison and connections to AWE. We have selected to use K-means clustering, although other clustering and analysis methods like hierarchical clustering [78] are equally valid for grouping items by relative similarity.

2.2.1. K-Means Clustering

K-means clustering is an algorithm that classifies data into a certain number of groupings based on variance among the input variables (e.g., [79,80]). Specifically, algorithms identify clusters of cases that are most similar to each other (i.e., within-cluster variance) while distinct from other clusters (i.e., between-cluster variance) across input variables. Importantly, the number of clusters generated in the analysis is prespecified (i.e., k = number of clusters). K-means clustering is a commonly used clustering method used to identify categories in language research [78,81].

To identify the optimal number of clusters, the outputs of for each k from 2 to 20 (in this case) are plotted and compared. The scree plot illustrating the sum of squared errors (SSE) for each cluster can be inspected to identify the “elbow”-shaped curve in the plot—the inflection point indicating that additional clusters contribute minimal additional variance (i.e., increasing k clusters results in only minor shifts in SSE). Similar cluster number selection processes are attested in other work on linguistic data [81].

The analysis was conducted using the k-means function in the stats package in R [81]. Input data included the 18 TAASSC syntax metrics identified in Table 1. Thus, this analysis reveals clusters of student writers characterized by different “patterns” or “profiles” of syntactic sophistication and complexity.

2.2.2. Discriminant Function Analysis

Discriminant function analysis (DFA) is a statistical process that classifies cases into distinct categories based on patterns of input variables (e.g., [82,83]). The categories are prespecified; multivariate analyses reveal the input variables that most discriminate between these groups. DFA produces a number of outputs that enable characterization of the predicted clusters, including (a) descriptive statistics for target cluster and input variables, (b) the functions (similar to linear regression equations) that determine cluster membership, (c) eigenvalues and tests of statistical significance for each function, (d) a structure matrix that reports the loadings of input variables on each function, and (e) group centroids (i.e., mean values computed from each function for each cluster).

The DFA was conducted using IBM SPSS 29.0. The target categories were the clusters identified by the k-means clustering algorithm (see Results). Input data included the 18 TAASSC syntax metrics identified in Table 1. This analysis describes the syntactic variables that best define or describe observed syntactic clusters, if any.

2.2.3. Analysis of Variance (ANOVA) and Linear Regression

To examine associations between observed clusters and writing quality, ANOVAs were conducted to test whether clusters differed in human-assigned scores. Subsequently, linear regression analyses were conducted to reveal the variables that most predicted variance in scores (a) across the corpus and (b) within each cluster.

3. Results

3.1. Variance in Syntactic Sophistication and Complexity among Student Writers

A four-cluster solution was the most optimal and parsimonious; five or more clusters offered minimal further impact on observed SSE. Table 4 reports the number of essays per cluster, along with the means and standard deviations for syntactic variables (see [78,81] for similar K-Means analyses). All indices demonstrated statistically significant differences across clusters. Effects sizes were also generally large, although several metrics were notable, and include the following: average number of dependents per nominal ($\eta^2 = 0.56$), average lemma frequency ($\eta^2 = 0.38$), average lemma constructions combinations frequency ($\eta^2 = 0.35$), average number of nominal complements per clause ($\eta^2 = 0.34$), average number of dependents per nominal ($\eta^2 = 0.33$), and average number of prepositions per nominal ($\eta^2 = 0.33$). Thus, aspects of clause complexity, noun phrase complexity, and sophistication all contributed to differences between clusters.

Table 4. Mean TAASSC index values (and SDs) for the four clusters identified in the k-means analysis.

	Clusters						
	Cluster 1 (<i>n</i> = 8730)	Cluster 2 (<i>n</i> = 9608)	Cluster 3 (<i>n</i> = 6306)	Cluster 4 (<i>n</i> = 11,563)	<i>F</i> (3,36,203)	<i>p</i>	η ²
Clausal Complexity							
Adjective Complements (avg)	0.12 (0.06)	0.07 (0.04)	0.07 (0.05)	0.08 (0.04)	2061.28	<0.001	0.15
Nominal Complements (avg)	0.12 (0.06)	0.07 (0.04)	0.17 (0.07)	0.08 (0.04)	6251.43	<0.001	0.34
Noun Phrase Complexity							
Dependents per Nominal (avg)	0.95 (0.11)	0.80 (0.11)	1.19 (0.14)	1.06 (0.11)	15,611.47	<0.001	0.56
Dependents per Object (avg)	1.15 (0.29)	1.04 (0.25)	1.48 (0.37)	1.30 (0.27)	3244.24	<0.001	0.21
Dependents per Preposition (avg)	1.07 (0.21)	1.06 (0.22)	1.26 (0.24)	1.27 (0.19)	2817.51	<0.001	0.19
Dependents per Nominal (stdev)	1.07 (0.12)	1.01 (0.12)	1.29 (0.18)	1.14 (0.12)	6010.14	<0.001	0.33
Dependents per Subject (stdev)	0.83 (0.18)	0.70 (0.20)	0.95 (0.26)	0.88 (0.19)	2311.62	<0.001	0.16
Dependents per Object (stdev)	0.95 (0.22)	0.96 (0.20)	1.14 (0.29)	1.07 (0.22)	1165.04	<0.001	0.09
Dependents per Preposition (stdev)	0.92 (0.17)	0.94 (0.19)	1.13 (0.20)	1.06 (0.19)	2207.78	<0.001	0.15
Determiners per Nominal (avg)	0.32 (0.07)	0.25 (0.07)	0.34 (0.08)	0.33 (0.07)	4056.27	<0.001	0.25
Prepositions per Nominal (avg)	0.11 (0.04)	0.09 (0.04)	0.18 (0.05)	0.14 (0.04)	5953.60	<0.001	0.33
Adjectival Modifiers (avg)	0.21 (0.14)	0.18 (0.11)	0.28 (0.17)	0.24 (0.13)	848.93	<0.001	0.07
Prepositions per Preposition (avg)	0.10 (0.06)	0.09 (0.06)	0.16 (0.08)	0.15 (0.07)	2113.29	<0.001	0.15
Syntactic Sophistication							
Lemma Frequency (avg)	2,189,536.92 (441,103.99)	1,499,780.58 (393,182.18)	2,254,996.81 (526,383.97)	1,572,929.61 (397,670.37)	7278.08	<0.001	0.38
Combinations Frequency (avg)	219,078.92 (71,150.95)	116,342.16 (52,686.20)	219,198.11 (84,868.54)	127,725.66 (54,329.51)	6633.29	<0.001	0.35
Constructions Frequency (log)	4.79 (0.21)	4.67 (0.22)	4.82 (0.22)	4.61 (0.20)	1844.56	<0.001	0.13
Lemmas in Reference (prop)	0.95 (0.03)	0.94 (0.04)	0.95 (0.04)	0.92 (0.04)	1399.68	<0.001	0.10
Combinations in Reference (prop)	0.86 (0.05)	0.83 (0.06)	0.86 (0.06)	0.80 (0.06)	2408.76	<0.001	0.17

Three statistically significant discriminant functions were reported. Function 1 accounted for 68.7% of the variance in the clusters (eigenvalue = 2.37), Function 2 accounted for 27.3% of the variance (eigenvalue = 0.94), and Function 3 accounted for 4.0% of the variance (eigenvalue = 0.14). Multivariate tests were statistically significant for tests of Functions 1 through 3, Wilks' $\lambda = 0.13$, $\chi^2(54) = 72,787.50$, $p < 0.001$; Functions 2 through 3, Wilks' $\lambda = 0.45$, $\chi^2(34) = 28,765.54$, $p < 0.001$; and Function 3, Wilks' $\lambda = 0.88$, $\chi^2(16) = 4703.70$, $p < 0.001$. In simpler terms, Function 1 was the primary driver of category membership; many cases could be sorted based on this function alone. Function 2 also contributed substantively to classifying cases. The contribution of Function 3 was very small yet statistically significant due to the large sample size (i.e., high power).

The DFA structure matrix (Table 5) summarizes the sophistication and complexity variables that loaded most strongly on (i.e., correlated with) each function. For readability, correlations below 0.30 are not reported.

Function 1 was characterized by variations in noun phrase complexity. Specifically, influential components of Function 1 included phrases with more dependents, prepositions, and determiners per noun phrase, on average. This function was driven by complicated noun phrases (e.g., “the angry dog on the short leash”) in contrast to simpler noun phrases (e.g., “the dog”). This function also included higher variation in dependents per nominal. Instead of uniformly complex noun phrases, there could be a mix of simpler and complex phrasing.

Function 2 was characterized by variations in familiar language. Specifically, components of Function 2 are primarily related to the use of more frequent, and thus more common and familiar words and sentence constructions. In addition, this pattern demonstrated fewer dependents per preposition (e.g., “on the leash”) instead of more complex phrases with more dependents (e.g., “on the long leash loosely held by the owner”).

Table 5. DFA function loading using TAASSC indices, ordered by function and magnitude.

Syntactic Variable	Function		
	1	2	3
Dependents per Nominal (avg)	0.706	−0.327	
Dependents per Nominal (stdev)	0.444		0.306
Prepositions per Nominal (avg)	0.434		
Determiners per Nominal (avg)	0.364		−0.345
Dependents per Direct Object (avg)	0.319		
Lemma Frequency (avg)	0.350	0.573	
Lemma Construction Combinations (avg)	0.321	0.563	
Lemmas in Reference (prop)		0.432	0.322
Dependents per Preposition		−0.357	
Construction Frequency (log)		0.340	
Combinations in Reference (prop)		0.328	
Adjective Complements (avg)			−0.785
Nominal Complements (avg)			0.461

Finally, Function 3 was characterized by variations in clause complexity. Components of Function 3 negatively related to the number of adjectival complements per clause and positively related to nominal complements per clause. Thus, this function captured clauses and sentences with more nouns but fewer adjectives (e.g., “the owner held a leash as she walked her dog” compared to “the nervous owner tightly held the frayed leash as she walked her energetic dog”).

Similar to linear regression, linear discriminant functions can be used to calculate mean values for each function based on their constituent variables. These “group centroids” reveal discriminating patterns across the clusters (Table 6). Function 1 (noun phrase complexity) strongly discriminated between Clusters 2 and 3. A larger positive Function 1 value was associated with Cluster 3, whereas a larger negative value was associated with Cluster 2. Function 2 (familiar language) further discriminated between Clusters 1 and 4. A larger positive Function 2 value was associated with Cluster 1, whereas a larger negative value was associated with Cluster 4. Thus, given Functions 1 and 2, many essays might be classified within one of four distinct clusters. Function 3 (clause complexity) provided additional nuance to further discriminate and characterize the clusters. For example, Clusters 1 and 4 both exhibited somewhat negative values for Function 4, whereas Clusters 2 and 3 displayed somewhat positive values. The four clusters are further described in the following sections.

Table 6. Group centroids based on eigenvalues for each function by each cluster.

Cluster	Group Centroid		
	Function 1	Function 2	Function 3
Cluster 1	0.02	1.40	−0.39
Cluster 2	−2.05	0.03	0.37
Cluster 3	2.67	0.22	0.48
Cluster 4	0.24	−1.20	−0.28

3.2. Summary of Clusters

Cluster 1 was distinguished by a use of familiar language (i.e., Function 2), defined as words and sentence constructions that occur more frequently in English. These essays also exhibited relatively higher use of adjectival complements per clause than other clusters (i.e., negative value for Function 3). Thus, adjective structures were more structurally complex and potentially more descriptive. For example, compare the adjective (in brackets) in “the student was [happy]” versus the more complex adjectival clause in “the student was [happy that he passed all his math and engineering tests]”. Cluster 1 can be tentatively named *Familiar and Descriptive Language*.

Cluster 2 was characterized by simpler noun phrases with fewer dependents per nominal, per direct object, per preposition, and so on (i.e., Function 1). These essays also demonstrated the lowest variance in these metrics. Thus, authors of these essays consistently employed simpler syntax at the noun phrase level. Cluster 2 can be named *Consistently Simple Noun Phrases*.

Cluster 3 demonstrated many of the highest values for measures of noun phrase complexity (i.e., Function 1) along with the highest variance in these indicators. Thus, essays in this cluster employed more complex sentence structures but also varied in levels of complexity. In addition, essays in this cluster demonstrated the highest mean value for average number of nominal complements per clause (i.e., Function 3), further adding to overall complexity. These dual patterns of complexity and variability are often noted as hallmarks of “skillful” syntax in writing [84,85]. Notably, these essays also tended to use more familiar words and sentence constructions (i.e., Function 2). Cluster 3 can be named *Variably Complex Noun Phrases*.

Cluster 4 was distinguished by words and sentence constructions that are less frequent in the English language (i.e., Function 2). These essays also demonstrated moderately complex noun phrases (i.e., Function 1) and more adjective complements per nominal (i.e., negative value for Function 3). Taken together, these patterns suggest that authors perhaps displayed a more extensive or sophisticated vocabulary, which was implemented descriptively and via moderately complex sentences. Cluster 4 might be named *Moderate Complexity with Less Familiar Language*.

3.3. Relationships between Clusters, Syntactic Sophistication, and Writing Quality

Although writing assessment encompasses more than “quality”, the ability to assign valid “scores” to student writing remains an important goal for instructors and AWE [19,83,84]. Thus, it is meaningful to consider how the observed syntactic clusters were associated with variations in writing quality, and whether distinct clusters achieved successful writing in different ways.

Mean holistic scores were computed and compared for each cluster (ANOVA), revealing the following significant main effect of cluster: $F(3,36,203) = 326.16$, $p < 0.001$, $\eta^2 = 0.03$. Specifically, Cluster 4 (Moderate Complexity with Less Familiar Language) reported the highest score ($M = 3.64$, $SD = 0.90$), followed by Cluster 3 (Variably Complex Noun Phrases) ($M = 3.52$, $SD = 0.97$), Cluster 1 (Familiar and Descriptive Language), and then Cluster 2 (Consistently Simple Noun Phrases). All pair-wise comparisons were significant (i.e., all $p < 0.001$). Superficially, Clusters 4 and 3 both exhibited signs of syntactic complexity that is often rewarded in assessment, whereas Clusters 1 and 2 perhaps align with simpler writing. Thus, this statistically significant “ordering” of clusters may seem to confirm expectations about “good” writing. However, the overall main effect and differences between clusters were quite small.

Figure 1 provides the following revealing illustration: every possible score (i.e., from 2 to 6) was observed in every possible cluster. In other words, each of the four clusters encompassed a range of writing quality. Although not equally likely, student writers who demonstrated “familiar and descriptive language” (Cluster 1) could achieve the same levels of success as students who exhibited “consistently complex noun phrases” (Cluster 3), and so on. These patterns provide evidence that observed clusters were not merely incremental manifestations of linear syntactic sophistication (i.e., from “less” to “more”). In other words, distinct writing patterns are not inherently “good” or “bad” but can be enacted in varying ways that receive better or worse evaluations.

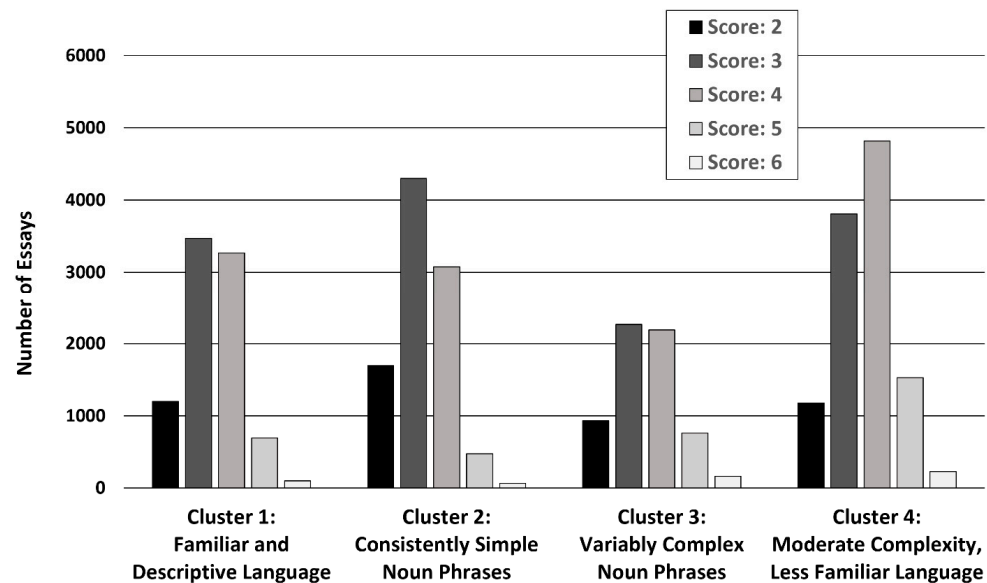


Figure 1. Distribution of essays by score (2 through 6) and cluster (1 through 4).

Linear regression analyses (Table 7) were conducted to explore how well the 18 syntactic and sophistication variables might predict holistic essay scores. We first conducted a linear regression for the entire corpus to examine how syntax predicted quality overall. We then investigated each cluster to explore how and whether within-cluster estimates differed from whole-corpus estimates. Importantly, we recognize that syntax alone should not account for much variance in writing quality. Nonetheless, syntax contributes to perceived writing quality because syntax is a part of holistic writing skills [7,58]. “Bad grammar” results in lower perceived quality (e.g., Johnson and colleagues [86]). For brevity, we omit correlation matrices for each analysis. However, for any given analysis, correlations for individual metrics were small ($r < |0.20|$) but nearly all were statistically significant (i.e., $p \leq 0.001$).

For the entire corpus, the linear regression was significant, $F(18,36,206) = 256.38$, $p < 0.001$, $R^2 = 0.11$. Thus, a model based on a small number of syntactic indices accounted for about 11% of the variance in scores. Standardized beta coefficients suggest that a variety of factors influenced scores, such as noun phrase complexity (e.g., standard deviation for dependents per nominal, and standard deviation for dependents per direct object) and sophistication (e.g., average frequency of lemmas and average proportion of lemma construction combinations appearing the reference corpus). Essays attained higher scores when they demonstrated variable complexity (i.e., a mix of simpler and complex structures) and used recognizable but less common vocabulary and language.

The linear regression for Cluster 1 Only was statistically significant, $F(18,8729) = 79.13$, $p < 0.001$, $R^2 = 0.14$; as were Cluster 3 Only, $F(18,6305) = 55.58$, $p < 0.001$, $R^2 = 0.14$; and Cluster 4 Only, $F(18,11,562) = 70.48$, $p < 0.001$, $R^2 = 0.13$. Although these models were based on only a fraction of the entire dataset, they accounted for 14%, 14%, and 13% of the variance in individual cluster scores, respectively. In other words, estimating scores within clusters was perhaps more accurate than estimating scores for the whole corpus. The one exception was Cluster 2 Only, $F(18,9607) = 53.42$, $p < 0.001$, $R^2 = 0.09$, which exhibited a decrease.

These analyses further revealed that the variables contributing to score variations were similar but not identical across clusters. For Cluster 1 (Familiar and Descriptive Language), higher scores were most associated with (i.e., the largest β coefficients) the use of less frequent vocabulary ($\beta = -0.18$), higher variability in average dependents per direct object ($\beta = 0.17$), and higher variability in average dependents per object of the preposition ($\beta = 0.16$). Notably, in comparison to the whole corpus, measures of clausal complexity (e.g., average number of adjectival complements per clause) and noun phrase complexity

(e.g., average number of dependents per nominal or preposition) mattered less. Thus, when writers adopted a more familiar and descriptive style, they were more successful when using sophisticated vocabulary (e.g., precise and meaningful wording) and variable syntax, but increased complexity by itself was less meaningful.

Table 7. Linear regression analyses for predicting essay score across all clusters and within individual clusters.

Syntactic Measure	β Coefficients				
	Entire Corpus	Cluster 1 (<i>n</i> = 8730)	Cluster 2 (<i>n</i> = 9608)	Cluster 3 (<i>n</i> = 6306)	Cluster 4 (<i>n</i> = 11,563)
Clausal Complexity					
Adjective Complements (avg)	0.11	0.01	0.08	0.15	0.14
Nominal Complements (avg)	0.04	0.02	0.07	−0.01	0.07
Noun Phrase Complexity					
Dependents per Nominal (avg)	0.13	0.07	0.07	0.01	0.13
Dependents per Object (avg)	−0.10	−0.05	−0.03	−0.12	−0.12
Dependents per Preposition (avg)	−0.12	−0.07	−0.02	−0.20	−0.15
Dependents per Nominal (stdev)	−0.17	−0.08	−0.06	−0.19	−0.14
Dependents per Subject (stdev)	0.12	0.11	0.13	0.10	0.08
Dependents per Object (stdev)	0.15	0.17	0.15	0.06	0.13
Dependents per Preposition (stdev)	0.13	0.16	0.10	0.12	0.10
Determiners per Nominal (avg)	0.12	0.09	0.09	0.13	0.08
Prepositions per Nominal (avg)	0.05	0.05	0.04	−0.02	0.07
Adjectival Modifiers (avg)	0.05	0.06	0.05	0.00	0.05
Prepositions per Preposition (avg)	0.06	0.08	0.07	0.05	0.02
Syntactic Sophistication					
Lemma Frequency (avg)	−0.19	−0.18	−0.10	−0.20	−0.11
Combinations Frequency (avg)	0.03	−0.01	0.08	−0.04	0.09
Constructions Frequency (log)	−0.08	−0.12	−0.03	−0.10	−0.04
Lemmas in Reference (prop)	−0.09	−0.11	−0.10	−0.07	−0.05
Combinations in Reference (prop)	0.16	0.12	0.18	0.07	0.19
Model Statistics					
	$R^2 = 0.11$	$R^2 = 0.14$	$R^2 = 0.09$	$R^2 = 0.14$	$R^2 = 0.13$
	$F(18,36,206) =$	$F(18,8729) =$	$F(18,9607) =$	$F(18,6305) =$	$F(18,11,562) =$
	256.38	79.13	53.42	55.58	70.48
	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$

Cluster 1 might be exemplified by sentences (5) and (6) below. These sentences (and later examples) illustrate properties exhibited in real student writing. However, none of the examples are direct quotes as per nondisclosure agreements. Sentence (5) is modeled after a sentence from a higher scoring essay. This sentence uses familiar yet meaningful words to convey ideas with precision. In contrast, sentence (6) demonstrates that familiarity can coincide with a lack of clarity and sophistication. The words are highly familiar, yet tend to be vague in meaning (e.g., “old” and “kinds of things”). Sentence (6) is modeled after a sentence from a lower scoring essay.

5. *Venus is the most comparable planet to earth, and sometimes, the closest in distance.*
6. *The Earth is old and has many different kinds of things living on Earth.*

For Cluster 2 (Consistently Simple Noun Phrases), higher scores were most associated with higher variability in the average number of dependents per nominal subject ($\beta = 0.13$), variability in the average number of dependents per direct object ($\beta = 0.15$), and higher proportion of lemma construction combinations appearing in the reference corpus ($\beta = 0.18$). In comparison to the whole corpus, overall noun phrase complexity and the use of less frequent words were less important. However, unlike Cluster 1, the impact of clausal complexity was similar to the corpus mean. Overall, student writers whose syntactic pattern demonstrated simplicity attained better scores when they used recognizable language

(e.g., fewer spelling and grammatical errors) and demonstrated syntactic variability. When syntax is generally more simple, occasional instances of complexity likely “stand out”. Indeed, skillful writers may even strategically rely on simpler writing to communicate most ideas, but then use greater complexity only when necessary for the topics at hand.

Sentence (7) illustrates two sentences with varying complexity of noun phrase structure from a higher scoring essay, whereas sentence (8) emulates sentences with similar properties from a lower scoring essay. In both cases, the constituent noun phrases are simple, yet writers display varying degrees of skill in communicating ideas coherently. Example (7) communicates in a relatively straightforward manner. In contrast, example (8) strings together multiple ideas and noun phrases in a more tangled structure. Example (8) is more complex in a less effective way.

7. *Each time a person gets into a car, they put themselves at the risk of being killed or severely injured in a car accident from the second they turn the ignition to the moment they put the car back in “park”. Traffic accidents claim the lives of countless innocent people each and every day.*
8. *driverless cars should not be made or thought about personal. Also in the reading it states that driverless cars arent fully driverless some of them need to have the hands on the sensors on the steering wheel and the seats will vibrate when something is wrong and the car cant take control of it and you have to control the car yourself.*

For Cluster 3 (Variably Complex Noun Phrases), higher ratings were associated with a less common vocabulary ($\beta = -0.20$), lower average number of dependents per object of the preposition ($\beta = -0.20$), less variability in the of dependents per nominal ($\beta = -0.19$), and more adjectival complements per clause ($\beta = 0.15$). Essays in this cluster received higher scores when writers used more sophisticated vocabulary and when noun phrase complexity did not involve overly complicated prepositions and prepositional phrases. Thus, when writers used more advanced syntax, it was perhaps important not to “overdo it”. The incorporation of more descriptive detail or precision was also beneficial (i.e., adjectival complements).

Sentences (9) and (10) illustrate ways in which complex noun phrases manifested in higher and lower scored essays, respectively. In sentence (9), higher complexity serves to establish the writer’s stance and contribute meaningful information. In sentence (10), similar properties result in a sentence that is less well organized and harder to parse.

9. *With car companies such as [company name] already planning the release of these self-driving cars, this future of transportation will increase safety, efficiency, and entertainment for humans going from one place to another and eventually make standard automobiles obsolete.*
10. *I never want there to be flying cars because thats when people get lazy and the cars would be useless i want to be able to hop in my cars and go race around and not hop in it and read a book and watch the car drive.*

For Cluster 4 (Moderate Complexity with Less Familiar Language), higher essay scores were associated with more recognizable word and construction combinations appearing in the reference corpus ($\beta = 0.19$), lower average number of dependents per object of the preposition ($\beta = -0.15$), less variability in the of dependents per nominal ($\beta = -0.14$), and more adjectival complements per clause ($\beta = 0.14$). Similar to Cluster 3, essays in this cluster received higher scores when noun phrase complexity did not rely overmuch on complicated prepositions and prepositional phrases. Given that Cluster 4 tended to exhibit more prepositional complexity, it seemed particularly worthwhile for writers to moderate that tendency. The incorporation of more descriptive detail or precision was again beneficial (i.e., adjectival complements). However, the factors that contributed to higher scores in Cluster 4 differed from Cluster 3 in a few ways. The use of sophisticated vocabulary was less important than using recognizable language (e.g., fewer typos, grammar errors, or slang terms). In addition, a higher average number of dependents per nominal ($\beta = 0.13$) and higher variability in the number of dependents per direct object ($\beta = 0.13$) somewhat contributed to higher scores for Cluster 4.

Sentences (11) and (12) below are somewhat lengthy and complex, thus requiring some attention to parse. However, although sentence (11) outlines its message supported by the complexity, sentence (12) presents many ideas in one sentence in a way that is harder to follow.

11. *Since automobiles were first invented, they have been continuously updated in all aspects of the car, it's design, how aerodynamic it is, the amount of cylinders an engine can have, the fuel efficiency, and a large variety of other properties.*
12. *Self-driving cars could be a more productive way for transportation and could also save a lot of lives in the process, a long with making the common person's life just a bit easier in this hard world.*

4. Discussion

Appreciating variance in writing is an important component of valid assessment because students express themselves and achieve their writing goals in diverse ways. Consequently, for AWE systems to optimally facilitate appropriate writing assessment, these technologies must be designed to also recognize variability. In the current paper, we explored how variance in writing variables pertaining to syntactic complexity and sophistication could be captured in a large corpus of high school argumentative essays using natural language processing (NLP) tools. Specifically, the Tool for the Automated Assessment of Syntactic Sophistication and Complexity (TAASCC, [63,68]) was used to detect clausal complexity, noun phrase complexity, and syntactic sophistication. We then conducted clustering and DFA analyses to characterize possible “syntactic styles” in student essays. To the extent that NLP tools and quantitative analyses can perform these tasks, they demonstrate how AWE tools might implement similar approaches.

4.1. Syntactic Variance in Writing

Our primary research questions considered syntactic variance displayed by student writers (RQ1) and how such variations could be characterized via NLP indices (RQ2). Our analyses demonstrated four possible clusters representing different syntactic complexity and sophistication profiles. Inspection of means and DFA allowed us to define these clusters. Observed patterns also partially corroborated styles reported in prior NLP-based analyses (e.g., [7]; see also [44,87,88]).

Cluster 1 was defined by *descriptive and familiar language*. For this cluster, overall noun phrase complexity was moderate, but essays tended to use more frequently occurring words and constructions (COCA, [69]). Clausal complexity was somewhat higher in this cluster than in others, particularly with respect to adjectival complements. Cluster 1 essays tended to include more adjectives in clauses or information that elaborated the meaning of adjectives. Although not perfectly aligned, this cluster perhaps captured syntactic elements of the “action and depiction” style described by [7], which was characterized by an increased number of adjectives, adverbs, rhetorical devices, and words overall.

Cluster 2 was defined by *consistently simple noun phrases*. Cluster 2 was the most syntactically simple of all four clusters, with fewer dependents (e.g., per nominal, direct object, and preposition) and low variability. Essays also tended to use less recognizable and frequently occurring language. The syntactic patterns displayed in this cluster potentially resemble the “accessible” style displayed in [7], which was characterized by the use of more common words and constructions, lower syntactic complexity, higher cohesion, and higher lexical and semantic overlap. Thus, as above, our focus on syntax may have captured a portion of that style.

Cluster 3 was defined by *variably complex noun phrases*. Essays in this cluster demonstrated high (or the highest) mean values for noun phrase complexity. In addition, these essays exhibited high variability in these measures—ranging from moderate to very high complexity. Such complexity was perhaps balanced by using more frequent and familiar words and constructions. Our Cluster 3 perhaps displays some similarity to the “lexical” style shown in [7], which also featured words and constructions that are less common.

In addition, their lexical cluster was described as having greater lexical diversity, more imageable words, and more specific words. Our analyses did not include word-based measures, but we can speculate that a more sophisticated vocabulary might lead to more detailed and complex noun phrases.

Finally, Cluster 4 was defined by *moderately complex noun phrases and clauses* overall, along with the use of less frequently occurring words and sentence constructions. The syntactic properties displayed in this cluster may resemble the “academic” style present in [7], which was similarly characterized by syntactic complexity and less frequent lemma and construction patterns. Their “academic” style also included strong structural components and rhetorical choices, which were beyond the scope of TAASSC to detect in this study.

4.2. Style and Score

Given the fundamental work of evaluating student writing, this research also considered the associations between observed styles and essay ratings (RQ3). One possibility was that clusters might be ordered linearly by quality—representing a range from “good” or “skilled” syntax to “poor” or “unskilled” syntax (e.g., see research on writing assessment rubrics, [89–91]). Indeed, findings showed that Cluster 4 (Moderate Complexity with Less Familiar Language) earned the highest ratings, followed by Cluster 3 (Variably Complex Noun Phrases), Cluster 1 (Descriptive and Familiar Language), and then Cluster 2 (Consistently Simple Noun Phrases). The higher scoring clusters demonstrated moderate-to-high syntactic complexity balanced by use of familiar language—these patterns align with prior research on syntax and writing quality [54,62,92,93].

Crucially, differences in average cluster scores exhibited very small effect sizes. Statistical significance was likely due to the large number of essays analyzed. Most importantly, all possible scores were distributed across all observed clusters—successful writing was possible regardless of syntactic pattern. In addition, the pathway to success differed somewhat across clusters. When writers implemented greater syntactic complexity, it was worthwhile to moderate such complexity with more familiar language, avoid overly convoluted syntax, and perhaps interweave more and less complex sentences. However, when writers favored syntactic simplicity, it was perhaps worthwhile to demonstrate meaningful and sophisticated word choices, and occasional sentence complexity, for more precise communication. These findings corroborate broad guidance for students to improve their syntax, diction, and varied sentence structure, but also underscore that not all students may equally benefit from generalized feedback recommendations.

4.3. Implications for Automated Writing Evaluation

AWE systems employ diverse computational and machine learning processes to detect linguistic features (e.g., vocabulary, syntax, cohesion, and semantics) and perform writing evaluation based on statistical generalizations derived from prior ratings [17,19,84]. Once an AWE algorithm is developed and deployed, all essays and writers can be evaluated in a rapid, consistent, and scalable manner. However, this approach may neglect critical variance because students can navigate the same tasks and goals of writing in different ways (e.g., [7,44,93,94]) that may defy uniform assessment. Although such variance is perhaps understood by expert human writing instructors, many or most AWE systems are not equipped to detect or respond to the different ways that students write.

4.4. AWE Development

Current findings offer evidence that greater algorithmic sensitivity to syntactic variance, one dimension of a more complex variance including other linguistic properties (e.g., lexicon, cohesion), in writing is both possible and necessary for AWE. Future systems need the capacity to automatically detect and respond to distinct writing styles, patterns, behaviors, or strategies exhibited by different students. Importantly, this approach specifically avoids prescriptive (and potentially biased) notions of what constitutes “good” or “desirable” writing. Moreover, this formulation avoids linear assumptions that student writing

only varies from “less” (or “poor”) to “more” (or “good”) on given features of language. Sensitivity to variance emphasizes acknowledging *how* students write before evaluating how *well* students write, because assessment may need to differ based on students’ pattern or approach.

Improved algorithmic sensitivity is attainable through at least two advancements. First, the field should continue to develop expanded automated indices that capture a broad range of writing features, behaviors, and more. For example, Kyle, Crossley, and colleagues have contributed an impressive variety of tools to the *Suite of Automatic Linguistic Analysis Tools* (SALAT) (e.g., [39,63,95,96]). Other teams have innovated methods for detecting and assessing student revisions [97–99], use of rhetorical moves [92,100], writing behaviors and keystrokes [94,98,99], and more. Syntax indices alone can already reveal multiple distinct profiles—almost certainly an understatement of the true variation among student writers that can be captured via rich toolkit of NLP packages.

Importantly, metrics need not be limited to writing features and processes. For instance, the current study did not formally analyze variations across different demographic backgrounds, but preliminary inspection found that all four syntactic clusters were observed across all reported grades, races and ethnicities, genders, and language backgrounds (i.e., English language learner and native English speakers). Future work will need to consider how nonlinguistic variables and demographic data could further enrich our understanding of variability, context, and nuances in student writing when paired with NLP metrics (see [101]). Writers’ motivations and cultural experiences shape the knowledge and experiences they bring to writing [4–6], which should be respected throughout the assessment of writing.

Enhanced NLP detection (and additional variables) is only the initial step towards improved AWE sensitivity to variance in writing. The second necessary advancement is to develop alternative approaches for operationalizing that variance. The current study employed simple but accessible methods for clustering and characterization (i.e., k-means clustering and DFA) that can be readily replicated. These accessible analytical methods were intentionally chosen to conduct a coarse analysis, but nonetheless revealed meaningful clusters of student writing. Moreover, for three out of four clusters, estimations of essay scores were better within-cluster than when derived from the entire corpus. More sophisticated clustering and profiling methods (for example, MDA, e.g., [30]; latent profile analysis, e.g., [93]) will almost certainly contribute to even more nuanced understanding of student writing. Future approaches may also benefit from deeper examination of interdependencies between writing features, such as combinations of nested variables that account for the changing influences of metrics in context (e.g., the impact of clausal complexity may depend on vocabulary usage).

Large Language Models (LLMs) are another technological advancement that have already begun transforming educational technologies (see [102] for review of recent LLM use in education). LLMs are machine learning algorithms that use deep learning [103] to develop generalizations from training data. Although this process can occur without human intervention, fine-tuning is often required at a later stage (i.e., supervision).

A recent study [104] tested the performance of several LLMs (i.e., Google’s PaLM 2, Anthropic’s Claude 2, and OpenAI’s GPT-3.5 and GPT-4) compared to humans in essay scoring. Findings showed that GPT-4 had the best performance as measured by intra-rater reliability and validity. However, GPT-4 performance worsened over time. In another study [105], researchers conducted several experiments testing fine-tuned GPT-3.5 and GPT-4 performance in conducting essay scoring and aiding instructors. Findings showed that the fine-tuned GPT-3.5 produced consistent and accurate scoring. An additional finding was that the LLM model helped human scorers perform better. Specifically, novice raters were able to learn faster, and experienced raters were able to become more consistent and efficient. To conclude, LLMs are a promising avenue in automated scoring and evaluation, which can allow for nuanced generalizations and variance. At the same time, LLMs must be supervised and fine-tuned to avoid the perpetuation of biases in the training data [101].

4.5. Implications for Instruction with AWE

Research on AWE has observed mixed but encouraging findings for the effectiveness of these systems [13,14], with critiques arising due to the formulaic, decontextualized, and/or impersonal way that AWE systems assess writing. Improving algorithmic sensitivity may address these challenges.

Instead of assessing student writing from a singular perspective (i.e., the same algorithms applied to all essays), our findings suggest an approach in which AWE systems might (a) first detect the approach(es) exhibited within an essay, and then (b) provide assessment and feedback attuned to those patterns. Variance-sensitive models and systems might operate in a multi-stage or “nested” fashion. Our data showed that essays in all four syntactic clusters could attain high scores, but the pathway to success may differ somewhat. Instead of encouraging all writers to use more sophisticated syntax and vocabulary, some writers might benefit from guidance in varying or even reducing overall complexity. Other writers may benefit from strategies for leveraging familiar and accessible language to complement complex syntax. Importantly, the implication here is not to sort or lock students into a handful of stagnant profiles that determine their destiny. Rather, the purpose is to recognize and appreciate variance in written expression, which then enables instruction and support that aligns with writers’ current strengths and needs in context (e.g., scholarship on asset-based and student-centered assessment, [106,107]). AWE systems that are sensitive and responsive to variance may be better able to provide feedback that is centered on the students and their writing rather than the software.

From the increasingly popular perspective of explainable AI (e.g., [108]), it might also be pedagogically worthwhile to explain to students how AWE systems define, detect, and assess different “patterns”. Students might be invited to reflect on their own preferred patterns and/or practice new patterns. Instead of learning to write in a single formulaic manner to “get a good grade from the computer”, students might learn to purposefully explore or enact different patterns that will be assessed on their own merits. A more nuanced AWE system may serve to reinforce that there are multiple ways to write successfully and express ideas. Scoring and feedback need not happen in the same way for every student, and thus, diverse patterns (and students) can achieve comparable success. Through transparency and explainability regarding variance, students might (a) better understand, discuss, or debate how “scores” are determined, and (b) gain greater awareness for how intentional writing choices can influence writing quality and audiences.

Author Contributions: Conceptualization, M.G., R.D.R. and A.G.A.; methodology, M.G., R.D.R. and A.G.A.; software, A.G.A.; validation, M.G., R.D.R. and A.G.A.; formal analysis, A.G.A., R.D.R. and M.G.; investigation, M.G., R.D.R. and A.G.A.; data curation, M.G. and A.G.A.; writing—original draft preparation, M.G.; writing—review and editing, M.G. and R.D.R.; visualization, M.G. and R.D.R.; supervision, R.D.R.; funding acquisition, R.D.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Gates Foundation (INV-006213).

Data Availability Statement: A description of the data and a downloadable version are available via [67].

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

References

1. Biber, D.; Conrad, S. *Register, Genre, and Style*; Cambridge University Press: Cambridge, UK, 2019. [\[CrossRef\]](#)
2. Nesi, H.; Gardner, S. *Genres across the Disciplines: Student Writing in Higher Education*; Cambridge University Press: Cambridge, UK, 2012.
3. Allen, L.K.; Likens, A.D.; McNamara, D.S. Writing flexibility in argumentative essays: A multidimensional analysis. *Read. Writ.* **2019**, *32*, 1607–1634. [\[CrossRef\]](#)
4. Collins, P.; Tate, T.P.; Won Lee, J.; Krishnan, J.A.; Warschauer, M. A multi-dimensional examination of adolescent writing: Considering the writer, genre and task demands. *Read. Writ.* **2021**, *34*, 2151–2173. [\[CrossRef\]](#)

5. Graham, S.; Harris, K.R.; Fishman, E.; Houston, J.; Wijekumar, K.; Lei, P.W.; Ray, A.B. Writing skills, knowledge, motivation, and strategic behavior predict students' persuasive writing performance in the context of robust writing instruction. *Elem. Sch. J.* **2019**, *119*, 487–510. [\[CrossRef\]](#)
6. Wijekumar, K.; Graham, S.; Harris, K.R.; Lei, P.W.; Barkel, A.; Aitken, A.; Ray, A.; Houston, J. The roles of writing knowledge, motivation, strategic behaviors, and skills in predicting elementary students' persuasive writing from source material. *Read. Writ.* **2019**, *32*, 1431–1457. [\[CrossRef\]](#)
7. Crossley, S.A.; Roscoe, R.D.; McNamara, D.S. What is successful writing? An investigation into the multiple ways writers can write successful essays. *Writ. Commun.* **2014**, *31*, 184–214. [\[CrossRef\]](#)
8. Attali, Y. A comparison of newly-trained and experienced raters on a standardized writing assessment. *Lang. Test.* **2016**, *33*, 99–115. [\[CrossRef\]](#)
9. Raczynski, K.R.; Cohen, A.S.; Engelhard, G., Jr.; Lu, Z. Comparing the effectiveness of self-paced and collaborative frame-of-reference training on rater accuracy in a large-scale writing assessment. *J. Educ. Meas.* **2015**, *52*, 301–318. [\[CrossRef\]](#)
10. Denessen, E.; Hornstra, L.; van den Bergh, L.; Bijlstra, G. Implicit measures of teachers' attitudes and stereotypes, and their effects on teacher practice and student outcomes: A review. *Learn. Instr.* **2022**, *78*, 101437. [\[CrossRef\]](#)
11. Quinn, D.M. Experimental evidence on teachers' racial bias in student evaluation: The role of grading scales. *Educ. Eval. Policy Anal.* **2020**, *42*, 375–392. [\[CrossRef\]](#)
12. Kellogg, R.T.; Whitehead, A.P. Training advanced writing skills: The case for deliberate practice. *Educ. Psychol.* **2009**, *44*, 250–266. [\[CrossRef\]](#)
13. Stevenson, M.; Phakiti, A. Automated feedback and second language writing. In *Feedback in Second Language Writing: Contexts and Issues*; Hyland, K., Hyland, F., Eds.; Cambridge University Press: Cambridge, UK, 2019; pp. 125–142. [\[CrossRef\]](#)
14. Wilson, J.; Myers, M.C.; Potter, A. Investigating the promise of automated writing evaluation for supporting formative writing assessment at scale. *Assess. Educ. Princ. Policy Pract.* **2022**, *29*, 183–199. [\[CrossRef\]](#)
15. Dodigovic, M.; Tovmasyan, A. Automated writing evaluation: The accuracy of Grammarly's feedback on form. *Int. J. TESOL Stud.* **2021**, *3*, 71–88. [\[CrossRef\]](#)
16. Ferrara, S.; Qunbar, S. Validity arguments for AI-based automated scores: Essay scoring as an illustration. *J. Educ. Meas.* **2022**, *59*, 288–313. [\[CrossRef\]](#)
17. Shermis, M.D.; Burstein, J. *Handbook of Automated Essay Evaluation: Current Applications and New Direction*; Routledge: New York, NY, USA, 2013. [\[CrossRef\]](#)
18. Strobl, C.; Ailhaud, E.; Benetos, K.; Devitt, A.; Kruse, O.; Proske, A.; Rapp, C. Digital support for academic writing: A review of technologies and pedagogies. *Comput. Educ.* **2019**, *131*, 33–48. [\[CrossRef\]](#)
19. Yan, D.; Rupp, A.A.; Foltz, P.W. (Eds.) *Handbook of Automated Scoring: Theory into Practice*, 1st ed.; CRC Press: Boca Raton, FL, USA, 2020. [\[CrossRef\]](#)
20. Chen, C.E.; Cheng, W. Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Lang. Learn. Technol.* **2008**, *12*, 94–112.
21. Fu, Q.K.; Zou, D.; Xie, H.; Cheng, G. A review of AWE feedback: Types, learning outcomes, and implications. *Comput. Assist. Lang. Learn.* **2022**, *37*, 179–221. [\[CrossRef\]](#)
22. Li, Z.; Feng, H.H.; Saricaoglu, A. The short-term and long-term effects of AWE feedback on ESL students' development of grammatical accuracy. *CALICO J.* **2017**, *34*, 355–375. [\[CrossRef\]](#)
23. Link, S.; Mehrzad, M.; Rahimi, M. Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Comput. Assist. Lang. Learn.* **2022**, *35*, 605–634. [\[CrossRef\]](#)
24. Stevenson, M.; Phakiti, A. The effects of computer-generated feedback on the quality of writing. *Assess. Writ.* **2014**, *19*, 51–65. [\[CrossRef\]](#)
25. Zhang, Z.V.; Hyland, K. Fostering student engagement with feedback: An integrated approach. *Assess. Writ.* **2022**, *51*, 100586. [\[CrossRef\]](#)
26. Grimes, D.; Warschauer, M. Utility in a fallible tool: A multi-site case study of automated writing evaluation. *J. Technol. Learn. Assess.* **2010**, *8*, 4–42.
27. Lv, X.; Ren, W.; Xie, Y. The effects of online feedback on ESL/EFL writing: A meta-analysis. *Asia-Pac. Educ. Res.* **2021**, *30*, 643–653. [\[CrossRef\]](#)
28. Nunes, A.; Cordeiro, C.; Limpo, T.; Castro, S.L. Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020. *J. Comput. Assist. Learn.* **2022**, *38*, 599–620. [\[CrossRef\]](#)
29. Zhang, Z.V. Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assess. Writ.* **2020**, *43*, 100439. [\[CrossRef\]](#)
30. Deane, P.; Quinlan, T. What automated analyses of corpora can tell us about students' writing skills. *J. Writ. Res.* **2010**, *2*, 151–177. [\[CrossRef\]](#)
31. Hyland, K.; Hyland, F. Feedback on second language students' writing. *Lang. Teach.* **2006**, *39*, 83–101. [\[CrossRef\]](#)
32. McCaffrey, D.F.; Zhang, M.; Burstein, J. Across performance contexts: Using automated writing evaluation to explore student writing. *J. Writ. Anal.* **2022**, *6*, 167–199. [\[CrossRef\]](#)
33. Perelman, L. When “the state of the art” is counting words. *Assess. Writ.* **2014**, *21*, 104–111. [\[CrossRef\]](#)

34. Hoang, G.T.L.; Kunnan, A.J. Automated essay evaluation for English language learners: A case study of MY Access. *Lang. Assess. Q.* **2016**, *13*, 359–376. [\[CrossRef\]](#)
35. Anson, C.M. Assessing writing in cross-curricular programs: Determining the locus of activity. *Assess. Writ.* **2006**, *11*, 100–112. [\[CrossRef\]](#)
36. Bai, L.; Hu, G. In the face of fallible AWE feedback: How do students respond? *Educ. Psychol.* **2017**, *37*, 67–81. [\[CrossRef\]](#)
37. Chen, D.; Hebert, M.; Wilson, J. Examining human and automated ratings of elementary students' writing quality: A multivariate generalizability theory application. *Am. Educ. Res. J.* **2022**, *59*, 1122–1156. [\[CrossRef\]](#)
38. Crossley, S.A. Linguistic features in writing quality and development: An overview. *J. Writ. Res.* **2020**, *11*, 415–443. [\[CrossRef\]](#)
39. Crossley, S.A.; Kyle, K. Assessing writing with the tool for the automatic analysis of lexical sophistication (TAALES). *Assess. Writ.* **2018**, *38*, 46–50. [\[CrossRef\]](#)
40. Crossley, S.A.; McNamara, D.S. Say more and be more coherent: How text elaboration and cohesion can increase writing quality. *J. Writ. Res.* **2016**, *7*, 351–370. [\[CrossRef\]](#)
41. Gardner, S.; Nesi, H.; Biber, D. Discipline, level, genre: Integrating situational perspectives in a new MD analysis of university student writing. *Appl. Linguist.* **2019**, *40*, 646–674. [\[CrossRef\]](#)
42. Graham, S. A revised writer(s)-within-community model of writing. *Educ. Psychol.* **2018**, *53*, 258–279. [\[CrossRef\]](#)
43. Biber, D.; Conrad, S. *Variation in English: Multi-Dimensional Studies*; Routledge: London, UK, 2014. [\[CrossRef\]](#)
44. Friginal, E.; Weigle, S.C. Exploring multiple profiles of L2 writing using multi-dimensional analysis. *J. Second Lang. Writ.* **2014**, *26*, 80–95. [\[CrossRef\]](#)
45. Goulart, L. Register variation in L1 and L2 student writing: A multidimensional analysis. *Regist. Stud.* **2021**, *3*, 115–143. [\[CrossRef\]](#)
46. Goulart, L.; Staples, S. Multidimensional analysis. In *Conducting Genre-Based Research in Applied Linguistics*; Kessler, M., Polio, C., Eds.; Routledge: New York, NY, USA, 2023; pp. 127–148.
47. McNamara, D.S.; Graesser, A.C. Coh-Metrix: An automated tool for theoretical and applied natural language processing. In *Applied Natural Language Processing and Content Analysis: Identification, Investigation, and Resolution*; McCarthy, P.M., Boonthum-Denecke, C., Eds.; IGI Global: Hershey, PA, USA, 2012; pp. 188–205. [\[CrossRef\]](#)
48. McNamara, D.S.; Graesser, A.C.; McCarthy, P.M.; Cai, Z. *Automated Evaluation of Text and Discourse with Coh-Metrix*; Cambridge University Press: New York, NY, USA, 2014. [\[CrossRef\]](#)
49. Butterfuss, R.; Roscoe, R.D.; Allen, L.K.; McCarthy, K.S.; McNamara, D.S. Strategy uptake in Writing Pal: Adaptive feedback and instruction. *J. Educ. Comput. Res.* **2022**, *60*, 696–721. [\[CrossRef\]](#)
50. McCarthy, K.S.; Roscoe, R.D.; Allen, L.K.; Likens, A.D.; McNamara, D.S. Automated writing evaluation: Does spelling and grammar feedback support high-quality writing and revision? *Assess. Writ.* **2022**, *52*, 100608. [\[CrossRef\]](#)
51. 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. [\[CrossRef\]](#)
52. Roscoe, R.D.; Allen, L.K.; Weston, J.L.; Crossley, S.A.; McNamara, D.S. The Writing Pal intelligent tutoring system: Usability testing and development. *Comput. Compos.* **2014**, *34*, 39–59. [\[CrossRef\]](#)
53. Weston-Sementelli, J.L.; Allen, L.K.; McNamara, D.S. Comprehension and writing strategy training improves performance on content-specific source-based writing tasks. *Int. J. Artif. Intell. Educ.* **2018**, *28*, 106–137. [\[CrossRef\]](#)
54. Jagaiah, T.; Olinghouse, N.G.; Kearns, D.M. Syntactic complexity measures: Variation by genre, grade-level, students' writing abilities, and writing quality. *Read. Writ.* **2020**, *33*, 2577–2638. [\[CrossRef\]](#)
55. Song, R. A scientometric review of syntactic complexity in L2 writing based on Web of Science (2010–2022). *Int. J. Linguist. Lit. Transl.* **2022**, *5*, 18–27. [\[CrossRef\]](#)
56. Staples, S.; Egbert, J.; Biber, D.; Gray, B. Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Writ. Commun.* **2016**, *33*, 149–183. [\[CrossRef\]](#)
57. Abbott, R.D.; Berninger, V.W.; Fayol, M. Longitudinal relationships of levels of language in writing and between writing and reading in grades 1 to 7. *J. Educ. Psychol.* **2010**, *102*, 281–298. [\[CrossRef\]](#)
58. Berninger, V.W.; Mizokawa, D.T.; Bragg, R.; Cartwright, A.; Yates, C. Intraindividual differences in levels of written language. *Read. Writ. Q.* **1994**, *10*, 259–275. [\[CrossRef\]](#)
59. Wilson, J.; Roscoe, R.D.; Ahmed, Y. Automated formative writing assessment using a levels of language framework. *Assess. Writ.* **2017**, *34*, 16–36. [\[CrossRef\]](#)
60. Lu, X. Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Lang. Test.* **2017**, *34*, 493–511. [\[CrossRef\]](#)
61. Mostafa, T.; Crossley, S.A. Verb argument construction complexity indices and L2 writing quality: Effects of writing tasks and prompts. *J. Second Lang. Writ.* **2020**, *49*, 100730. [\[CrossRef\]](#)
62. Kyle, K.; Crossley, S.; Verspoor, M. Measuring longitudinal writing development using indices of syntactic complexity and sophistication. *Stud. Second Lang. Acquis.* **2021**, *43*, 781–812. [\[CrossRef\]](#)
63. 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. [\[CrossRef\]](#)
64. Wang, S.; Xu, T.; Li, H.; Zhang, C.; Liang, J.; Tang, J.; Yu, P.S.; Wen, Q. Large language models for education: A survey and outlook. *arXiv* **2024**, arXiv:2403.18105. [\[CrossRef\]](#)

65. Ramesh, D.; Sanampudi, S.K. An automated essay scoring systems: A systematic literature review. *Artif. Intell. Rev.* **2022**, *55*, 2495–2527. [\[CrossRef\]](#)
66. Janda, H.K.; Pawar, A.; Du, S.; Mago, V. Syntactic, semantic and sentiment analysis: The joint effect on automated essay evaluation. *IEEE Access* **2019**, *7*, 108486–108503. [\[CrossRef\]](#)
67. Crossley, S.A.; Baffour, P.; Tian, Y.; Picou, A.; Benner, M.; Boser, U. The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (PERSUADE) corpus 1.0. *Assess. Writ.* **2022**, *54*, 100667. [\[CrossRef\]](#)
68. Uccelli, P.; Dobbs, C.L.; Scott, J. Mastering academic language: Organization and stance in the persuasive writing of high school students. *Writ. Commun.* **2013**, *30*, 36–62. [\[CrossRef\]](#)
69. Davies, M. *The Corpus of Contemporary American English (COCA): 560 Million Words, 1990-Present*; Brigham Young University: Provo, UT, USA, 2008.
70. Larsson, T.; Kaatari, H. Syntactic complexity across registers: Investigating (in) formality in second-language writing. *J. Engl. Acad. Purp.* **2020**, *45*, 100850. [\[CrossRef\]](#)
71. Deane, P.; Wilson, J.; Zhang, M.; Li, C.; van Rijn, P.; Guo, H.; Roth, A.; Winchester, E.; Richter, T. The sensitivity of a scenario-based assessment of written argumentation to school differences in curriculum and instruction. *Int. J. Artif. Intell. Educ.* **2021**, *31*, 57–98. [\[CrossRef\]](#)
72. Clarke, N.; Foltz, P.; Garrard, P. How to do things with (thousands of) words: Computational approaches to discourse analysis in Alzheimer’s disease. *Cortex* **2020**, *129*, 446–463. [\[CrossRef\]](#)
73. Bernius, J.P.; Krusche, S.; Bruegge, B. Machine learning based feedback on textual student answers in large courses. *Comput. Educ. Artif. Intell.* **2022**, *3*, 100081. [\[CrossRef\]](#)
74. Whitelock-Wainwright, A.; Laan, N.; Wen, D.; Gašević, D. Exploring student information problem solving behaviour using fine-grained concept map and search tool data. *Comput. Educ.* **2020**, *145*, 103731. [\[CrossRef\]](#)
75. Mizumoto, A. Calculating the relative importance of multiple regression predictor variables using dominance analysis and random forests. *Lang. Learn.* **2023**, *73*, 161–196. [\[CrossRef\]](#)
76. Sinharay, S.; Zhang, M.; Deane, P. Prediction of essay scores from writing process and product features using data mining methods. *Appl. Meas. Educ.* **2019**, *32*, 116–137. [\[CrossRef\]](#)
77. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A K-means clustering algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.* **1979**, *28*, 100–108. [\[CrossRef\]](#)
78. Crowther, D.; Kim, S.; Lee, J.; Lim, J.; Loewen, S. Methodological synthesis of cluster analysis in second language research. *Lang. Learn.* **2021**, *71*, 99–130. [\[CrossRef\]](#)
79. Wu, J. *Advances in K-Means Cluster: A Data Mining Thinking*; Springer: Berlin/Heidelberg, Germany, 2012.
80. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013. Available online: <https://www.R-project.org/> (accessed on 31 May 2024).
81. Talebinamvar, M.; Zarrabi, F. Clustering students’ writing behaviors using keystroke logging: A learning analytic approach in EFL writing. *Lang. Test. Asia* **2022**, *12*, 6. [\[CrossRef\]](#)
82. Meyers, L.S.; Gamst, G.; Guarino, A.J. *Applied Multivariate Research: Design and Interpretation*, 3rd ed.; Sage: Thousand Oaks, CA, USA, 2016.
83. Tabachnick, B.G.; Fidell, L.S. *Using Multivariate Statistics*, 7th ed.; Pearson: Boston, MA, USA, 2018.
84. Dikli, S. An overview of automated scoring of essays. *J. Technol. Learn. Assess.* **2006**, *5*, 4–34.
85. Shermis, M.D.; Burstein, J.; Elliot, N.; Miel, S.; Foltz, P.W. Automated writing evaluation: An expanding body of knowledge. In *Handbook of Writing Research*; MacArthur, C.A., Graham, S., Fitzgerald, J., Eds.; Guilford Press: New York, NY, USA, 2016; pp. 395–409.
86. Johnson, A.C.; Wilson, J.; Roscoe, R.D. College student perceptions of writing errors, text quality, and author characteristics. *Assess. Writ.* **2017**, *34*, 72–87. [\[CrossRef\]](#)
87. Jarvis, S.; Grant, L.; Bikowski, D.; Ferris, D. Exploring multiple profiles of highly rated learner compositions. *J. Second Lang. Writ.* **2003**, *12*, 377–403. [\[CrossRef\]](#)
88. Tywoniw, R.; Crossley, S. The Effect of Cohesive Features in Integrated and Independent L2 Writing Quality and Text Classification. *Lang. Educ. Assess.* **2019**, *2*, 110–134. [\[CrossRef\]](#)
89. Andrade, H.L.; Du, Y.; Mycek, K. Rubric-referenced self-assessment and middle school students’ writing. *Assess. Educ. Princ. Policy Pract.* **2010**, *17*, 199–214. [\[CrossRef\]](#)
90. Ghaffar, M.A.; Khairallah, M.; Salloum, S. Co-constructed rubrics and assessment for learning: The impact on middle school students’ attitudes and writing skills. *Assess. Writ.* **2020**, *45*, 100468. [\[CrossRef\]](#)
91. Panadero, E.; Jonsson, A. The use of scoring rubrics for formative assessment purposes revisited: A review. *Educ. Res. Rev.* **2013**, *9*, 129–144. [\[CrossRef\]](#)
92. Knight, S.; Shibani, A.; Abel, S.; Gibson, A.; Ryan, P.; Sutton, N.; Shum, S. AcaWriter: A learning analytics tool for formative feedback on academic writing. *J. Writ. Res.* **2020**, *12*, 141–186. [\[CrossRef\]](#)
93. McNamara, D.S.; Crossley, S.A.; McCarthy, P.M. Linguistic features of writing quality. *Writ. Commun.* **2010**, *27*, 57–86. [\[CrossRef\]](#)
94. Kim, H. Profiles of undergraduate student writers: Differences in writing strategy and impacts on text quality. *Learn. Individ. Differ.* **2020**, *78*, 101823. [\[CrossRef\]](#)

95. Van Steendam, E.; Vandermeulen, N.; De Maeyer, S.; Lesterhuis, M.; Van den Bergh, H.; Rijlaarsdam, G. How students perform synthesis tasks: An empirical study into dynamic process configurations. *J. Educ. Psychol.* **2022**, *114*, 1773–1800. [\[CrossRef\]](#)
96. Kyle, K.; Crossley, S. Assessing syntactic sophistication in L2 writing: A usage-based approach. *Lang. Test.* **2017**, *34*, 513–535. [\[CrossRef\]](#)
97. 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. [\[CrossRef\]](#) [\[PubMed\]](#)
98. Benetos, K.; Bétrancourt, M. Digital authoring support for argumentative writing: What does it change? *J. Writ. Res.* **2020**, *12*, 263–290. [\[CrossRef\]](#)
99. Bowen, N.E.J.A.; Thomas, N.; Vandermeulen, N. Exploring feedback and regulation in online writing classes with keystroke logging. *Comput. Compos.* **2022**, *63*, 102692. [\[CrossRef\]](#)
100. Correnti, R.; Matsumura, L.C.; Wang, E.L.; Litman, D.; Zhang, H. Building a validity argument for an automated writing evaluation system (eRevise) as a formative assessment. *Comput. Educ. Open* **2022**, *3*, 100084. [\[CrossRef\]](#)
101. Goldshtein, M.; Alhashim, A.G.; Roscoe, R.D. Automating Bias in Writing Evaluation: Sources, Barriers, and Recommendations. In *The Routledge International Handbook of Automated Essay Evaluation*; Shermis, M.D., Wilson, J., Eds.; Routledge: London, UK, 2024; pp. 421–445.
102. Yan, L.; Sha, L.; Zhao, L.; Li, Y.; Martinez-Maldonado, R.; Chen, G.; Li, X.; Jin, Y.; Gašević, D. Practical and ethical challenges of large language models in education: A systematic scoping review. *Br. J. Educ. Technol.* **2024**, *55*, 90–112. [\[CrossRef\]](#)
103. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [\[CrossRef\]](#) [\[PubMed\]](#)
104. Pack, A.; Barrett, A.; Escalante, J. Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Comput. Educ. Artif. Intell.* **2024**, *6*, 100234. [\[CrossRef\]](#)
105. Xiao, C.; Ma, W.; Xu, S.X.; Zhang, K.; Wang, Y.; Fu, Q. From Automation to Augmentation: Large Language Models Elevating Essay Scoring Landscape. *arXiv* **2024**, arXiv:2401.06431.
106. Aull, L. Student-centered assessment and online writing feedback: Technology in a time of crisis. *Assess. Writ.* **2020**, *46*, 100483. [\[CrossRef\]](#)
107. MacSwan, J. Academic English as standard language ideology: A renewed research agency for asset-based language education. *Lang. Teach. Res.* **2020**, *24*, 28–36. [\[CrossRef\]](#)
108. Kay, J. Foundations for human-AI teaming for self-regulated learning with explainable AI (XAI). *Comput. Hum. Behav.* **2023**, *147*, 107848. [\[CrossRef\]](#)

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.