Research Article

Discovering the Predictive Power of Five Baseline Writing Competences

Vivekanandan Kumar, *Athabasca University*
Shawn N. Fraser, *Athabasca University*
David Boulanger, *Athabasca University*

Structured Abstract

- **Background**: A shift of focus has been marked in recent years in the development of automated essay scoring systems (AES) passing from merely assigning a holistic score to an essay to providing constructive feedback over it. Despite all the major advances in the domain, many objections persist concerning their credibility and readiness to replace human scoring in high-stakes writing assessments. The purpose of this study is to shed light on how to build a relatively simple AES system based on five baseline writing features. The study shows that the proposed AES system compares very well with other state-of-the-art systems despite its obvious limitations.

- **Literature Review**: In 2012, ASAP (Automated Student Assessment Prize) launched a demonstration to benchmark the performance of state-of-the-art AES systems using eight hand-graded essay datasets originating from state writing assessments. These datasets are still used today to measure the accuracy of new AES systems. Recently, Zupanc and Bosnic (2017) developed and evaluated another state-of-the-art AES system, called SAGE, which enclosed new semantic and consistency features and provided for the first time an automatic semantic feedback. SAGE's agreement level between machine and human scores for ASAP dataset #8 (the dataset also of interest in this study) was measured and had a quadratic weighted kappa of 0.81, while it

DOI: 10.37514/JWA-J.2017.1.1.08

ranged for 10 other state-of-the-art systems between 0.60 and 0.73 (Chen et al., 2012; Shermis, 2014). Finally, this section discusses the limitations of AES, which come mainly from its omission to assess higher-order thinking skills that all writing constructs are ultimately designed to assess.

- **Research Questions**: The research questions that guide this study are as follows:

  RQ1: What is the power of the writing analytics tool's five-variable model (spelling accuracy, grammatical accuracy, semantic similarity, connectivity, lexical diversity) to predict the holistic scores of Grade 10 narrative essays (ASAP dataset #8)?

  RQ2: What is the agreement level between the computer rater based on the regression model obtained in RQ1 and the human raters who scored the 723 narrative essays written by Grade 10 students (ASAP dataset #8)?

- **Methodology**: ASAP dataset #8 was used to train the predictive model of the writing analytics tool introduced in this study. Each essay was graded by two teachers. In case of disagreement between the two raters, the scoring was resolved by a third rater. Basically, essay scores were the weighted sums of four rubric scores. A multiple linear regression analysis was conducted to determine the extent to which a five-variable model (selected from a set of 86 writing features) was effective to predict essay scores.

- **Results**: The regression model in this study accounted for 57% of the essay score variability. The correlation (Pearson), the percentage of perfect matches, the percentage of adjacent matches (± 2), and the quadratic weighted kappa between the resolved scores and predicted essay scores were 0.76, 10%, 49%, and 0.73, respectively. The results were measured on an integer scale of resolved essay scores between 10-60.

- **Discussion**: When measuring the accuracy of an AES system, it is important to take into account several metrics to better understand how predicted essay scores are distributed along the distribution of

human scores. Using average ranking over correlation, exact/adjacent agreement, quadratic weighted kappa, and distributional characteristics such as standard deviation and mean, this study's regression model ranks 4th out of 10 AES systems. Despite its relatively good rank, the predictions of the proposed AES system remain imprecise and do not even look optimal to identify poor-quality essays (binary condition) smaller than or equal to a 65% threshold (71% precision and 92% recall).

- **Conclusions**: This study sheds light on the implementation process and the evaluation of a new simple AES system comparable to the state of the art and reveals that the generally obscure state-of-the-art AES system is most likely concerned only with shallow assessment of text production features. Consequently, the authors advocate greater transparency in the development and publication of AES systems. In addition, the relationship between the explanation of essay score variability and the inter-rater agreement level should be further investigated to better represent the changes in terms of level of agreement when a new variable is added to a regression model. This study should also be replicated at a larger scale in several different writing settings for more robust results.

Keywords: automated essay scoring, connectivity, grammatical accuracy, inter-rater agreement, lexical diversity, regression, semantic similarity, spelling accuracy, writing analytics

1.0 Background

The ambition of scoring essays through a computer program dates back to the 1960s (Aluthman, 2016; Zupanc & Bosnic, 2017). With the soaring development of natural language processing techniques (NLP) and computational methods from the 1990s (Aluthman, 2016; Zupanc & Bosnic, 2017), the objective and the path to reach the quality and consistency of human graders started to take shape. The hope offered through the use of automated essay scoring systems includes improved scoring consistency, reduction of teachers' workload, shortened time to score and deliver grades to students, minimized scoring costs, and the now-possible provision of real-time formative feedback to students for better results on student writing proficiency (Aluthman, 2016; Deane, 2013; Latifi et al., 2016; Shermis, 2014; Zupanc & Bosnic, 2017). It is, however, important to acknowledge that machine scoring is mainly concerned with text production

features (e.g., spelling and grammatical accuracy, writing speed, cohesion, coherence) rather than with critical thinking (e.g., strength of argumentation, rhetorical effectiveness, attention to audience). Therefore, many object that machine scoring is not yet mature enough to replace human scoring, especially in high-stakes writing assessments (Deane, 2013; Perelman, 2013, 2014).

This notion of "automated essay scoring" (AES) has also been coined as automated essay grading (AEG), automated essay evaluation (AEE), and automated writing evaluation (AWE) (Zupanc & Bosnic, 2017). It should be noted that as the prospects of having such systems closely mimicking human graders, the term "evaluation" instead of "grading" or "scoring" has been used to reflect the changing priority from the assigning of a grade to providing instructive formative feedback to students (Zupanc & Bosnic, 2017). Nevertheless, the term AES will be mainly used in this study given that the authors propose a writing analytics tool that focuses on giving scores to a set of narrative essays (however, all these terms could be employed interchangeably). Thus, the AES process basically consists of evaluating and scoring essays through software and providing feedback to students based on the results of the evaluation (Zupanc & Bosnic, 2017). The types of feedback can be categorized as follows: 1) essay score (also known as holistic scoring), 2) rubric scores (also called trait scoring), and 3) identification and classification of good and bad writing practices according to the scoring rubrics along with the provision of feedback to help maintain or correct those practices (Aluthman, 2016; Fazal et al., 2013; Zupanc & Bosnic, 2017).

This study showcases in a transparent manner how a relatively simple writing analytics tool (an AES system) based on multiple linear regression can achieve relatively good results in terms of levels of agreement with human scores and can compare very well with other AES systems. Unfortunately, most literature on such systems discusses only at a high-level the key writing features they measure and the broad families of algorithms and statistical techniques used behind the scene to automate essay scoring. Although there exist studies in the literature that report the predictive power of specific writing features or of a subset of these features within the AES system they describe, the implementation details remain most often unknown, hiding their actual predictive power behind imperfect algorithms. This study advocates that more transparency on the implementation process would greatly benefit the writing analytics research community by helping researchers to understand what is the actual predictive power of a writing metric (the terms metrics, features, attributes, and factors are interchangeably used in this study) in the context of writing assessments performed by human readers and the estimated predictive power of that same metric using a particular software or algorithm.

This study addresses this gap by showcasing how to build, out of a set of 86 writing metrics, a regression model that will consist of only five baseline writing features (spelling accuracy, grammatical accuracy, semantic similarity, connectivity, lexical diversity) identified in the literature (Crossley et al., 2016; Fazal et al., 2013; McNamara et al., 2015), how to estimate its effectiveness in predicting the holistic scores of a batch of 723 Grade 10 narrative essays, and how to measure quantitatively its strength of agreement with the actual human scores and compare it with the performance of other state-of-the-art AES systems. The fairly "slim" design of the proposed writing analytics tool will underscore how so-called sophisticated AES systems superficially mimic human scoring (Deane, 2013; Perelman, 2013, 2014).

Given that AES systems enclose a very prized intellectual property, they are most often proprietary solutions sold as software as a service (SaaS), reducing opportunities for greater personalization to better meet educational institutions' goals and impeding optimal adaptation with other learning analytics initiatives undertaken by these same educational organizations. Some of the possibilities that may result from the personalized development of an AES system and its integration within an educational institution's system include:

- Tracking and evaluating the writing process in real time or in a customizable time frame instead of evaluating only the writing artifact at the end or at a few milestones during the writing process.

- Detecting students at risk of failing both at the short term (e.g., a specific formative activity or assessment) and at the long term (e.g., writing competences measured summatively that grow below the average classroom's rate).

- Providing a student with feedback over his/her effort in current essay. This feedback could include holistic scoring, trait scoring, identification per rubric of spots in the essay that require the student's attention, and provision of hints to remediate to the pointed issues.

- Providing a student with feedback over his/her achievement in relation to the curriculum's learning outcomes and the targeted writing competences. This will imply compiling evidence from all the writing activities performed by a specific student and reporting strengths and weaknesses in general.

- A self-regulated learning module where students will be able to set short-term and long-term goals to improve their writing proficiency. Students will be supplied with insights on the performance of top, average, and at-risk student groups to better situate their effort.

- Merging the writing analytics tool's dataset with other datasets collected through other learning analytics software implemented within the learning institution to acquire a better background knowledge of every student (e.g., socioeconomic status, talent, performance in previous years, etc.) and measure the impact of these factors over competence growth.

As a by-product of this study's investigation, this study shows how the proposed writing analytics tool could contribute to the implementation and evaluation of one of these use cases in a classroom setting; that is, how the tool can be useful in automatically identifying students scoring below a certain threshold when performing a writing activity. By being notified of students' poor performance, teachers would be able to provide remedial interventions in time. Moreover, this mechanism could be extended to multiple writing activities within the curriculum (outside formal assessments) so that teachers can receive formative feedback throughout a school year and, ahead of time, detect students who are struggling with text production skills. This study will, therefore, assess whether the proposed AES model is accurate enough for profitable usage in a classroom setting. In addition, by using AES as an assistant instead of a substitute to human teachers, researchers avoid the sensitive issue of replacing human scoring with machine scoring, as described later.

Central to this study is a set of eight essay datasets formed in the setting of the 2012 Automated Student Assessment Prize (ASAP) contest hosted by Kaggle and sponsored by the William and Flora Hewlett Foundation. However, the predictive model developed in this research is based only on one essay dataset (#8) as described in the next sections. It is important to note that these datasets are widely used to benchmark the performance of state-of-the-art AES systems (Chen et al., 2012; Shermis, 2014; Zupanc & Bosnic, 2017). This study will, therefore, compare the performance of the proposed writing analytics tool against the performance of previously benchmarked AES systems.

2.0 Literature Review

2.1 State-of-the-Art AES Systems

Zupanc and Bosnic (2017) proposed a new AES system called SAGE, trained and tested with the datasets provided in the setting of the ASAP

demonstration, that they claim surpasses the performance of previous state-of-the-art AES systems as described in Shermis (2014). Given that the techniques and efficacy of commercial AES systems are rarely published while they are still in the market, a large swath of literature on AES only discusses obsolete AES systems. Hence, the contribution by Zupanc and Bosnic is significant in terms of roughly indicating the state of research on AES today. SAGE distinguishes itself from other state-of-the-art AES systems in that it includes extra writing features on *semantic coherence* and *statement consistency* to counter the "predominant focus on vocabulary and text syntax, and limited consideration of text semantics" (Zupanc & Bosnic, 2017, p. 1). Essentially, semantic coherence is measured by mapping sequential parts (e.g., two consecutive sentences) of an essay to a semantic space where changes between the two parts are analyzed. Consistency metrics are based on information extraction and logic reasoning, where the goal is to compare the knowledge embedded in essays against a database of common sense knowledge to detect any semantic error. Zupanc and Bosnic show that the average quadratic weighted kappa of SAGE over the eight ASAP datasets surpasses 10 of the most popular AES systems benchmarked with the same essay datasets, many of which are commercial solutions (Chen et al., 2012; Shermis, 2014).

This enhanced accuracy, along with the fact that it is one of the first AES systems to analyze in more depth the consistency and the semantic coherence of essays, allows SAGE to provide faster, finer-grained semantic feedback to students. In contrast, Zupanc and Bosnic (2017) assert that the main fault of AES systems is that they focus mainly on aspects of vocabulary and text syntax, while ignoring or just superficially addressing text semantics. Moreover, many of the state-of-the-art AES systems are proprietary software, and little information is provided regarding the types of linguistic features they extract and the computational methods they employ (McNamara et al., 2014). Popular commercial solutions include PEG (Project Essay Grade), e-rater (developed by Educational Testing Service), Intelligent Essay Assessor (IEA; developed by Pearson Inc.), and IntelliMetric (Vantage) (Aluthman, 2016; Zupanc & Bosnic, 2017). Among open-source solutions, only LightSIDE is reported in the literature (Latifi et al., 2016; Zupanc & Bosnic, 2017). The reported average level of agreement for each of the 10 state-of-the-art systems as measured by the quadratic weighted kappa ranged from 0.63 to 0.79. These agreement measurements have all been obtained by testing every AES system with the eight essay datasets provided by ASAP. These 10 AES systems represented 97% of the US market in 2013 (Chen et al., 2012; Shermis, 2014; Zupanc & Bosnic, 2017).

Fazal et al. (2013) have identified four types of AES methods. The first type of method is called the hybrid method and employs a mix of NLP and

statistical techniques. AES systems that implement this method include PEG, e-rater, Criterion, SEAR, IEMS, PS-ME, Intellimetric, My! Access, and an AES system for CET4. The second type of method is based on latent semantic analysis (LSA) and includes all AES systems that employ "a statistical model of word usage that permits comparisons of the semantic similarity between pieces of textual information" (Foltz, 1996, p. 198). IEA, AEA, Jess, MarkIT, and AES systems based on Generalized LSA are considered LSA-based. The third type of AES method employs text categorization techniques (TCT) "to train binary classifiers to distinguish good from bad essays, and use the scores output by the classifiers to rank essays and assign grades to them" (Larkey, 1998, p. 90). AES systems such as BETSY, CarmelTC, and AES systems employing the k-nearest neighbor algorithm and TCT techniques are considered TCT-based. Finally, the fourth type of method employs miscellaneous techniques, which enclose less conventional algorithms (e.g., unsupervised learning, assessment of connections between paragraphs, usage of literary sememes, modified BLEU algorithm, etc.). In all, only four AES systems were classified by Fazal et al. as scoring essays based on rubrics, while the remaining 20 AES systems gave only a holistic score to essays.

2.2 Types of Writing Metrics

Zupanc and Bosnic (2017) state that AES systems can be divided into three core components: the writing attributes used to assess the quality of an essay; the methodology to extract the writing features from essays (e.g., the four methods identified by Fazal et al. (2013) as hybrid, LSA-based, TCT-based, and miscellaneous); and the predictive model to score the essays. Hence, adequately identifying those features of a text that are representative of essay quality meeting all their associated learning outcomes is crucial. Zupanc and Bosnic categorize writing attributes as style, content, and semantic. The essay style is assessed through elements such as lexical sophistication, grammar, and mechanics (spelling, capitalization, and punctuation). Content analysis generally implies only a high-level semantic analysis and comparison with source text and graded essays, while semantic metrics assess the correctness of content connotation. The reader should note that the selection of the proper set of writing features will depend on the type of writings students will be engaged in, the rubrics used to score the essays, and the type of feedback that must be provided to the users of the system (teachers and students).

For instance, PEG is particularly interested with the rubrics fluency, diction, and complexity. E-rater, on the other side, is more oriented toward grammar, usage, mechanics, style, and organization, while IEA focuses on content, mechanics, and style. IntelliMetric takes into account cohesion,

coherence, content, discourse, syntactic complexity, variety, and accuracy (Aluthman, 2016; El Ebyary & Windeatt, 2010; Warschauer, 2006). As for SAGE, it focuses on linguistics and content, semantic coherence, and semantic consistency. For a detailed list of possible writing features, the reader may consult McNamara et al. (2014) and Zupanc and Bosnic (2017).

2.3 Predictive Models

In contrast to Zupanc and Bosnic (2017), Latifi et al. (2016) have adopted the following three phases in the development of their AES system: 1) extraction of writing features, 2) creation of predictive models, and 3) actual essay scoring and provision of feedback. To create and assess the accuracy of a predictive model, a set of hand-graded essays needs first to be acquired and then separated into a training set and a testing set. The training set will tune the predictive model through the extracted writing features, while the validation (testing) set will measure the accuracy of the developed model by comparing the predicted scores against the human ratings (McNamara et al., 2014).

Automated essay scoring (AES) can be classified into two categories: holistic scoring or trait scoring (Aluthman, 2016; Fazal et al., 2013). Holistic scoring involves assigning an essay grade directly from the extracted writing features. Trait scoring provides a score for each scoring rubric as used by a teacher and derives the holistic score as a weighted sum of the trait scores. Trait scoring has been found as a solution to reduce variability among holistic scores assigned by several raters (McNamara et al., 2014). Hence, to implement trait scoring, a model based on its own specific set of writing attributes must be developed for each trait (rubric). For instance, to assess the score of a vocabulary rubric, a predictive model based on the following five writing features could be built (Gebril & Plakans, 2016; Gregori-Signes & Clavel-Arroitia, 2015): lexical diversity (number of unique words); lexical sophistication (sophistication of the words); lexical density (number of content words which include nouns, verbs, adjectives, and adverbs); lexical accuracy (number of errors); and lexical originality (the number of words unique to a writer divided by the total number of words in a corpus of a group of writers). Although some of these metrics will most likely be re-used to assess the score of another rubric (e.g., the content rubric), the remaining metrics will usually contribute solely to the assessment of the vocabulary rubric score. Thus, each rubric will have a unique predictive model. A trait-scoring AES system will, therefore, consist of a set of models, while holistic-scoring AES systems will typically consist of a single predictive model. This, of course, depends on the context in which an AES system operates (e.g., type of writing, grade, etc.).

AES systems are generally implemented as a combination of *computational linguistics* ("the study of computer processing, understanding, and generation of human languages"[1]), *statistical modeling* ("a simplified, mathematically-formalized way to approximate reality and optionally to make predictions from this approximation"[2]), and *natural language processing* ("a branch of artificial intelligence that deals with analyzing, understanding and generating the languages that humans use naturally in order to interface with computers in both written and spoken contexts"[3]) (McNamara et al., 2014). More specifically, the extraction of writing features occurs through natural language processing (NLP) techniques or through a variant of latent semantic analysis (LSA) when the features are more content-oriented (Zupanc & Bosnic, 2017). As for predictive models, different statistical approaches are used depending on the type of rubric scored. The most popular approaches are regression analyses, latent semantic analysis regression, neural networks, Bayesian networks, cosine similarity, and random forests (Aluthman, 2016; El Ebyary & Windeatt, 2010; McNamara et al., 2014; Warschauer, 2006; Zupanc & Bosnic, 2017). These artificial intelligence techniques are usually implemented through machine learning code libraries.

2.4 Accuracy of AES

It is important to understand how accuracy is measured and that accuracy of essay scoring is highly contextual depending on several factors such as the grade, whether English is the first or second language of the writer, the mode of writing (e.g., exposition, description, narrative, and persuasion), whether the writing activity is source-based or not, to name a few. A model suited for one context will not necessarily be optimal in another one. For instance, Aluthman (2016) asserted that AES designed for Grade 6-10 students with English as their native language might not apply in the same manner to English as a second language students.

---

[1] University of Toronto, Department of Computer Science (2017). Computational linguistics. Retrieved June 27, 2017, from http://www.cs.toronto.edu/compling/

[2] XLSTAT (2015, September 29). What is statistical modeling? Retrieved June 27, 2017, from https://help.xlstat.com/customer/en/portal/articles/2062460-what-is-statistical-modeling-?b_id=9283

[3] Webopedia (2017). NLP - natural language processing. Retrieved June 27, 2017, from http://www.webopedia.com/TERM/N/NLP.html

Determining the level of agreement between AES computer programs and human raters can be accomplished in a number of ways. The techniques used by Chen et al. (2012), Shermis (2014), and Zupanc and Bosnic (2017) to compare AES systems include, among others, exact agreement, adjacent agreement, the quadratic weighted kappa, and the Pearson correlation. Exact agreement computes the percentage of perfect matches between the computer rater and the human rater (e.g., both a computer rater and a human rater give a 90% to an essay). Adjacent agreement is the percentage of adjacent marks (e.g., an AES system gives a mark of 90% to an essay, while a human rater gives a 91%). The quadratic weighted kappa takes into account the size of the differences between the computer rater's ratings and the human rater's ratings and penalizes differences with a greater magnitude (Williamson et al., 2012). The Pearson correlation coefficient provides some estimate of the variability of essay scores explained by the computer rater. This study has used these four techniques to measure the accuracy of the proposed AES model to better situate its results with those of previously reported studies (Chen et al., 2012; Shermis, 2014; Zupanc & Bosnic, 2017).

Studies investigating the predictive power of specific groups of writing features and assessing the agreement level between these models and human raters (as the current study) have already been conducted in the past. For instance, a study by McNamara et al. (2013) performed a multiple regression analysis on a variety of text difficulty, cohesion, and rhetoric indices (more than 40). The results from the regression analysis showed that the following combination of eight writing attributes explained approximately 46% of the variability in scores of essay quality (from a corpus of 313 essays written by college freshmen): 1) the number of different (unique) words, 2) average givenness of each sentence (number of words that are new or given, such as initial noun referents versus noun referents referred to pronominally), 3) narrativity reading ease score, 4) noun hypernymy (word specificity), 5) LSA essay to prompt (lexical and semantic overlap between the entire essay and the essay prompt) (Crossley & McNamara, 2011), 6) conclusion paragraph n-grams (words and phrases common in high-quality conclusion paragraphs), 7) body paragraph n-grams (words and phrases common in high-quality body paragraphs), and 8) word frequency. The study reported a perfect agreement of 44% between human scores of essay quality and the model's predicted scores and an adjacent agreement of 94%. The weighted Cohen's kappa for adjacent agreement was 0.40, which was considered a moderate agreement. Several studies (Attali & Burstein, 2006; McNamara et al., 2015; Rudner et al., 2006) have reported that the correlation between human and computer-based essay scores ranged between 0.60 and 0.85, the perfect agreement was between 30% and 60%, and the adjacent agreement was between 85% and 100%.

Zupanc and Bosnic (2017) evaluated the accuracy of SAGE, a very recent AES system that they developed and tested, by building three distinct models. SAGE basically consists of 104 writing metrics subdivided as follows: 72 linguistic and content metrics, 29 coherence metrics, and 3 consistency metrics. The first model, which was called AGE, consisted exclusively of the 72 linguistic and content metrics (first type of writing attributes: style). The second model, called AGE+, included the 29 coherence metrics (second type of writing attributes: content) in addition to the AGE model's metrics. The third model, in addition to the AGE+ metrics, consisted of the three consistency metrics (third type of writing attributes: semantic). The third model is actually the full-fledged version of SAGE and is therefore referred to as such. Zupanc and Bosnic report that 22% of AGE's predicted essay scores and 16% of AGE+'s predicted scores perfectly matched with the human essay scores for the set of 723 narrative essays written by Grade 10 students, the eighth dataset supplied by ASAP. Their quadratic weighted kappas for the same essay dataset were 0.785 and 0.805, respectively, while SAGE had a kappa value of 0.81.

2.5 Limitations of AES Systems

While promising, automated essay scoring systems have been subjected to criticism, and stakeholders take polarized stands on the matter (Deane, 2013). On one side, there are those who advocate "unrestricted use of AES to replace human scoring," claiming that AES provides quick and almost real-time feedback to students; on the other side, there are those who recommend "complete avoidance of automated methods," claiming that writing is social by nature and that humans write for social purposes, that human writings deserve human readers, and that writing for a machine devalues human communications (Deane, 2013). Deane believes that neither of these polarized views is correct and that it is fundamental to understand where AES succeeds well and where human interventions remain necessary. Essentially, AES systems are considered relatively effective when assessing text production skills, but they are rather poor when it comes to evaluating the strength of argumentation or the rhetorical effectiveness of a piece of writing. Deane summarizes the three types of objections to automated essay scoring systems as follows: 1) students may change their writing behavior if they know that their writing will be assessed by a machine (e.g., temptation to game the system, lower quality resulting from demotivation in not communicating with a human); 2) the inability of the system to "interpret meaning, infer the communicative intent, evaluate factual correctness, and quality of argumentation, or take the writing process into account"; 3) the inaccuracy resulting from the inability of the AES system to deal with text semantics (as explained in #2),

which generates false positives and false negatives risking to wreck any trust in the overall system.

A study by Bennett (2011) with an essay dataset that originated from a persuasive writing prompt showed that text production was correlated to critical thinking. The essay set was scored by two sets of raters with the first set of raters assessing effective argumentation and attention to audience and the second set of raters using a text production rubric (fluency, coherence of expression, effective word choice, adherence to conventions, etc.). It was found that the correlation between the scores of both rubrics was 0.80. Given that there exists some relationship between text production proficiency and the level of cognitive engagement in addressing "rhetorical and conceptual problems" pertaining to the intended writing construct (Slomp, 2012), according to Deane (2013), AES systems can be highly correlated to overall performance despite the lack of consideration of elements related to critical-thinking skills. However, the omission of these higher-order elements can undermine the credibility of the AES system's predicted scores and potentially the users' trust in the overall system since the system only takes into account a subset of what human teachers consider as important. In addition, the absence of genuine critical-thinking evaluation compromises the quality of the formative feedback supplied to students by ignoring more advanced concepts. While neuroscience investigates brain dynamism to measure the presence of cognitive traits such as critical thinking, for now, one can remain convinced that AES systems are not yet ready to replace humans in the assessment of high-stakes tests. Yet, AES systems are well positioned to offer valuable supplementary information to human raters.

2.6 ASAP Studies

In order to investigate the capability of AES systems to meet the growing requirements of the Common Core State Standards to better prepare United States students for college in terms of writing proficiency and this despite significant educational budget cuts, the William and Flora Hewlett Foundation funded a demonstration organized by the Automated Student Assessment Prize (ASAP) to verify whether AES systems would be mature enough to shift from multiple-choice question assessments to essay assessments to measure actually how close a student's performance is to the intended writing construct and whether they could score a larger volume of essays (Perelman, 2013; Shermis, 2014). The purpose of that demonstration was to compare the performance of the state-of-the-art AES systems against human scorers and determine whether machine scoring was as reliable as human scoring. Hence, eight essay datasets totaling 22,029 essays hand-graded by teachers were collected from Grades 7, 8, and 10 from six different states in the United States. Each essay dataset responded to a different

writing prompt representing "a range of typical writing tasks … that might be similar to those anticipated in … new assessments" (Shermis, 2014). In addition, although Shermis (2014) warned against generalizing the results of the studies conducted in the context of the ASAP demonstration to larger student populations, the essays were collected to represent as much as possible gender and a range of ethnicities. These sample datasets were created to take into account a diversity of scoring policies, demographic compositions, geographic regions, and writing prompt types. Approximately 60% of these essays, in each dataset, was randomly set aside for training the scoring models of AES systems, while 20% of each essay dataset was randomly assigned to a validation set and 20% to a test set. Two studies were conducted in the setting of the ASAP demonstration. The first study examined the performance of nine AES systems from eight commercial vendors and one university laboratory (the open source LightSIDE). The goal was to examine the reliability of the current state-of-the-art systems against the reliability between human scorers. In the second study, the same essay datasets were made available to the public, and data scientists all over the world were invited to develop new algorithms and compare their performance against the industry solutions. The only difference between the two studies was that the essay datasets in the second study (public contest) were anonymized to protect students' privacy. To ensure that the anonymization would not impact the accuracy measurements of the AES systems developed by data scientists, one of the nine AES systems (LightSIDE) in the first study was selected to compare its accuracy with both the original datasets and the anonymized datasets. The difference in terms of agreement level (as measured by the quadratic weighted kappa) was found to be statistically not significant. The outcomes of the two studies were to be measured as the level of agreement between machine scores and human scores. Five metrics were used to measure agreement level: exact agreement, adjacent agreement, kappa, quadratic weighted kappa, and Pearson correlation.

According to Perelman (2013), Shermis (2014) initially presented his conclusions of the ASAP studies in an unpublished paper at the annual meeting of the National Council on Measurement in Education in the following terms:

> "The results demonstrated that overall, automated essay scoring was capable of producing scores similar to human scores for extended-response writing items with equal performance for both source-based and traditional writing genre [sic]."

Perelman (2013) reported that Shermis' study results were echoed and exaggerated in the press. This induced skepticism and discomfort to stakeholders

who claimed that the state-of-the-art AES systems merely counted words when predicting essay scores (Perelman, 2014).

Following the publication of the study results by Shermis, Perelman published his criticism against the validity of the studies' conclusions claiming above all that the experiments were conducted with "the absence of any *articulated* construct for writing" pointing out that half of the datasets were short responses (whose eligibility to be called "essay" was even questioned) that were oriented toward assessing reading comprehension instead of writing ability. Perelman extended his criticism to state that without explicitly defined writing constructs, it is simply "impossible to judge the validity of any measurement." In addition, Perelman indicated that all multi-paragraph essays were converted to single paragraphs (paragraph markings were removed) preventing machine scorers from considering key essay features such as paragraph coherence. Besides these remarks, Perelman highlighted the danger of misinterpreting the outcomes of both machine and human scoring by comparing "apples with oranges." He denounced the employment of two different scales (for four out of the eight datasets) when measuring the reliability of the scores assigned by humans and machines, which in this situation granted an advantage to machine scoring. Finally, Perelman exposed the lack of statistical methods to judge whether the differences between the reliability of machines and humans were significant. For all of these reasons and others, he invited Shermis to formally retract the study results.

A year later, Shermis re-published the study results, providing some of the justifications that Perelman was looking for. For example, he nuanced the previous claims as follows: "With additional validity studies, it appears that automated essay scoring holds the potential to play a viable role in high-stakes writing assessments [sic]." He also made available the testing dataset as requested by Perelman for external verification. In addition, he explained why disparate scales were used to compare reliability among human and machine scorers. Since the scores from the two human raters contributed to determine the resolved scores of the essays (for several essay datasets), comparing a human rater's ratings against the resolved scores would have introduced a bias caused by the whole-part correlation that existed between the human scores and resolved scores. Hence, the ratings of the first human rater were directly compared against the ratings of the second rater, while machine scores were directly compared to the resolved scores. This did not satisfy Perelman as it can be seen in his next publication. However, it is important to say that despite all of these critiques, Shermis did a colossal work by laying the foundation for determining the state of machine scoring, that is, how it compares against the reliability of traditional human scoring.

In a more recent publication (2014), Perelman exposed the results of his own analysis of the dataset previously made available by Shermis. Interestingly, he brought out the fact that AES overvalued certain essay features, while devaluing other features. For example, he demonstrated that the correlation between word counts and machine scores was significantly higher than the correlation between word counts and human scores. He also made the point that the state-of-the-art AES systems seemingly approximated the level of agreement between two human raters by overvaluing the shared variance between the number of words in an essay and the machine score. He re-asserted the fact that Shermis' data analysis provided an advantage to machines by measuring agreement level between the machine scores and resolved scores. He suggested that the comparison should have been made using the mean of human raters' ratings.

3.0 Research Questions

The goal of this study is to assess the accuracy of the writing analytics tool's model as previously described and compare it against AES systems that currently exist on the market. Given that commercial products do not reveal much about their implementation details, they can only provide limited clues about the importance of each writing feature in predicting holistic scores. Although the proposed writing analytics tool's model consists of only five variables (spelling accuracy, grammatical accuracy, semantic similarity, connectivity, and lexical diversity), it is hoped that it will enrich the reader's comprehension regarding the impact and role that each variable plays in predicting holistic scores. For that purpose, this study attempted to answer these two research questions:

> RQ1: What is the power of the writing analytics tool's five-variable model (spelling accuracy, grammatical accuracy, semantic similarity, connectivity, lexical diversity) to predict the holistic scores of Grade 10 narrative essays (ASAP dataset #8)?

It was initially hypothesized that given the limited number of variables in the model, the accuracy of the predicted essay scores would be lower or equivalent to the average state-of-the-art AES system. However, it was wondered whether using some of the state-of-the-art analytics techniques could compensate the limited number of variables and improve the ability of the proposed writing analytics tool to predict essay scores. At a minimum, it was speculated that the writing analytics tool's model would accurately detect poor-quality essays. Simplistically, an essay will be considered poor-quality if it gets a score smaller

than or equal to 65%.

> RQ2: What is the agreement level between the computer rater based on the regression model obtained in RQ1 and the human raters who scored the 723 narrative essays written by Grade 10 students (ASAP dataset #8)?

It was expected that the accuracy of the regression model, developed as part of the effort to answer the first research question, would generate agreement levels among the computer and human raters in the range of what is reported in the literature, that is, a correlation coefficient between 0.60 and 0.85, a percentage of perfect agreement between 30% and 60%, a percentage of adjacent agreement beyond 85%, and a quadratic weighted kappa between 0.60 and 0.73. The main question is how much of the predictive power, as assessed in a regression analysis, translates in better agreement levels as per the dimensions listed above.

This study replicated much of the AES development process as described in the literature, hopefully providing a more transparent process on one way to conduct a systematic regression analysis.

4.0 Research Methodology

4.1 Dataset & Participants

This study is based on the eighth essay dataset supplied in the setting of the ASAP studies. The broader context of these studies has been depicted in the previous section. However, this section provides the details of the selected dataset and justifications for that selection and reports how reliability of machine scoring has been measured in the ASAP studies and how it has driven the analysis plan of this study.

Eight essay sample datasets were collected from six states in the United States: three states from the Northeast, two from the Midwest, and one from the West Coast. Each essay dataset originated from a state assessment. Given that state writing assessments significantly differed from state to state, each sample dataset came from a single state. Each sample dataset was then randomly selected from the student population of that state. As previously mentioned, the formation of these datasets was conducted with the purpose of representing the diversity of scoring policies among the states, demographic compositions in the United States, geographic regions, and types of essay prompts, etc. The demographic information about the student participants of these studies was not revealed by the participating states. Hence, the demographic features of the students participating in these studies had to be estimated from the entire student population for the selected grade of the state concerned. Because of that, the results of the ASAP

studies as well as of the current study could not be generalized to the entire student populations. Table 1 displays the estimated demographic characteristics of the student samples for each dataset. It can be seen from the table that there were 44,289 Grade 10 students in the dataset #8's state and that 1,527 essays were sampled from the Grade 10 student population (3.45%). In the Grade 10 student population, 48.7% were male and 51.3% were female; 66.3% were white, while 33.7% were non-white; and 41.3% of the Grade 10 student population came from a more precarious financial situation. Given that the sample of essays resulted from a random selection from the student population, this provided a rough estimate of the demographic characteristics of the participants in the studies.

Table 1

Characteristics of Sample Essay Datasets and Estimated Sample Demographic Characteristics from Reported State Demographics (Shermis, 2014)

	Data set #							
	1	2	3	4	5	6	7	8
State	State 1	State 2	State 3	State 3	State 4	State 4	State 5	State 6
Grade	8	10	10	10	8	10	7	10
Grade level *N*	42,992	80,905	68,025	68,025	71,588	73,101	115,626	44,289
n	2,968	3,000	2,858	2,948	3,006	3,000	2,722	1,527
Training *n*	1,785	1,800	1,726	1,772	1,805	1,800	1,730	918
Test *n*	589	600	568	586	601	600	495	304
Validation *n*	594	600	564	590	600	600	497	305
Mean # of words	366.40	381.19	108.69	94.39	122.29	153.64	171.28	622.13
SD # of words	120.40	156.44	53.30	51.68	57.37	55.92	85.20	197.08
Gender M% / F%	51.2 / 48.8	51.4 / 48.6	51.0 / 49.0	51.0 / 49.0	49.6 / 50.4	49.2 / 50.8	51.2 / 48.8	48.7 / 51.3
Race W% / N%	63.8 / 36.2	77.8 / 22.2	42.9 / 57.1	42.9 / 57.1	70.2 / 29.8	69.5 / 30.5	70.2 / 29.8	66.3 / 33.7
Free/reduced lunch %	32.9	40.0	32.2	32.2	34.2	34.2	46.6	41.3

Following Perelman's critiques on the experimental design of the ASAP studies, the current study focused only on the eighth essay dataset since many of the critiques particularly concerned datasets 3-6 and, at a lesser extent, datasets 1, 2, 7, and 8. For example, Perelman indicated that although there was no articulated writing construct for any of the essay datasets, the writing "constructs" for datasets 1, 2, 7, and 8 at least assessed the writing ability of students instead of assessing only the essay contents. Moreover, only three datasets (1, 2, and 8) had a mean number of words per essay greater than 350 words. The five other datasets had a mean number of words ranging from 94.39 to 171.28 words, so small that it was questionable to consider them as essays. Dataset #8 was selected since it had a significantly higher mean of words per essay than the other datasets and contained (most likely) the greatest number of multi-paragraph essays. Besides, the average quadratic weighted kappa of the tested AES systems for this dataset

was the lowest among the eight datasets (0.67). Finally, the way resolved scores were calculated for dataset #8 did not introduce bias in favor of machine scoring as was the case with the resolved scores for many of the other datasets.

The prompt and scoring rubrics of the writing construct for essay dataset #8 are displayed in Table 2. The scoring of the essays was subcontracted by the states to commercial testing vendors who guaranteed quality assurance by recruiting, training, and staffing professional graders and checking the reliability and validity of the scorings (Shermis, 2014). Each essay was scored by two human raters. Each essay was given a score by each rater for each rubric on an integer scale from one to six (1-6) with six being the best score. Only four of the six rubrics contributed to the final essay score (rubrics 1, 2, 5, and 6). Thus, the resolved score of each rubric was in most cases calculated as being the sum of the two raters' scores, implying that the range of the resolved scores was 2-12. The adjudication rules in case of disagreement among the two human raters were defined as 1) if on any single rubric the raters' ratings are not adjacent (differ by more than one), then a resolution will be required; 2) if all four scores for rubrics 1, 2, 5, and 6 of one of the two raters are identical and that the other rater's scores for the same rubrics are also identical to the first rater's scores except for one rubric score that is an adjacent mark, then a resolution will also be required (e.g., rater 1 gives 4 to all four rubrics and rater 2 gives 4 to three rubrics and 3 to the fourth rubric). When a resolution was required, the concerned essay was rated by a third human rater, whose scores served as the resolved scores (RS) for all rubrics. Each rubric score was then multiplied by two to fit the range 2-12. The holistic score of the essay (after resolution) was then calculated to be:

$$\mathrm{RS}_{\mathrm{essay}} = \mathrm{RS}_1 + \mathrm{RS}_2 + \mathrm{RS}_5 + 2 \cdot \mathrm{RS}_6$$

where RS_i is the resolved score for rubric i. Thus, the range of resolved scores for the holistic scores was 10-60 (an integer scale).

Table 2

Essay Prompt and Scoring Rubrics for the Writing Construct of ASAP Dataset #8

Prompt & rubrics	Description
Essay prompt	We all understand the benefits of laughter. For example, someone once said, "Laughter is the shortest distance between two people." Many other people believe that laughter is an important part of any relationship. Tell a true story in which laughter was one element or part.
Rubric 1: Ideas and Content	Is the writing exceptionally clear, focused, and interesting? Does it hold the reader's attention throughout? Do main ideas stand out, are they well developed by strong support and rich details suitable to audience and purpose?
Rubric 2: Organization	Does the organization of the writing enhance the central idea(s) and its/their development? Are the order and structure compelling, and do they move the reader through the text easily?
Rubric 3: Voice	Has the writer chosen a voice appropriate for the topic, purpose, and audience? Does the writer demonstrate deep commitment to the topic, and is there an exceptional sense of "writing to be read?" Is the writing expressive, engaging, or sincere?
Rubric 4: Word Choice	Do the words convey the intended message in an exceptionally interesting, precise, and natural way appropriate to audience and purpose? Does the writer employ a rich, broad range of words, which have been carefully chosen and thoughtfully placed for impact?
Rubric 5: Sentence Fluency	Has the writing an effective flow and rhythm? Do sentences show a high degree of craftsmanship, with consistently strong and varied structure that makes expressive oral reading easy and enjoyable?
Rubric 6: Conventions	Does the writing demonstrate exceptionally strong control of standard writing conventions (e.g., punctuation, spelling, capitalization, grammar, and usage) and use them effectively to enhance communication? Are errors so few and so minor that the reader can easily skim right over them unless specifically searching for them?

Source: Automated Student Assessment Prize (2012). The Hewlett Foundation: Automated essay scoring. Retrieved June 27, 2017, from https://www.kaggle.com/c/asap-aes/data

All the original handwritten documents of the essays of dataset #8 were handed over by the states to ASAP to be converted to the ASCII format. The

essay documents were firstly scanned using high-quality digital scanners and then transcribed by transcription companies. The guidelines used by the transcription companies are further elaborated in Shermis (2014). The reader should note here that some essays were filtered out after random selection from the original population. Some essays were not included in the analysis because they were either illegible (e.g., undecipherable handwriting, smudged original documents, handwriting too light to be reproduced well), off-topic, or inappropriate and could not be scored by human raters. The scanning and transcription processes introduced some errors. The transcription companies measured the accuracy of their transcription, whose results are reported in Shermis (2014). Besides, multi-paragraph essays were converted to single-paragraph ones during the conversion process to ASCII format. The removal of paragraph markings potentially constituted an impediment to the performance of certain AES systems. Finally, the essays were anonymized for the second ASAP study to protect students' privacy. For example, names of people, organizations, and locations (countries, states, cities) as well as dates, times, amounts of money, percentages, names of months, email addresses, numbers, etc. were anonymized as follows: @PERSON, @ORGANIZATION, @LOCATION, @DATE, @TIME, @MONEY, @PERCENT, @MONTH, @EMAIL, @NUM, @CAPS, @DR, @CITY, @STATE[4]. As described previously, a test was performed with the LightSIDE AES system on both the anonymized and the non-anonymized datasets. It was found that the difference between the quadratic weighted kappas generated from the analyses of both datasets was not significant. This study was performed on the anonymized dataset.

The reliability of the machine scores in relation to the resolved scores (RS) was measured through the following metrics in the ASAP studies: exact agreement, adjacent agreement, kappa, quadratic weighted kappa, Pearson correlation, and the distributional characteristics standard deviation and mean. Given the wide range of the holistic scores (10-60) for essay dataset #8, adjacent marks (between machine scores and resolved scores) were defined as two marks, whose difference was zero, one, or two. The kappa metric used in the ASAP studies measures the level of agreement between two datasets when there is no ordinality in the data. The quadratic weighted kappa on the other side is appropriate in presence of ordinal data. However, Perelman (2013) warned that quadratically weighted kappas tend to increase with larger scales and that "variation of the quadratically weighted kappa coefficient with the number of categories appears to be strongest in the range from two to five categories

[4] Automated Student Assessment Prize (2012). The Hewlett Foundation: Automated essay scoring. Retrieved June 27, 2017, from https://www.kaggle.com/c/asap-aes/data

(Brenner & Kliebsch, 1996).” As for the Pearson correlation, Shermis (2014) stated that coefficient values usually mirror those of the quadratic weighted kappas.

4.2 Instruments

The proposed writing analytics tool processed further the dataset of 723 narrative essays. Eighty-six writing features were extracted by means of natural language processing libraries (Stanford CoreNLP and Apache OpenNLP), a spellchecker (LanguageTool enhanced by the Google n-gram corpus), latent semantic analysis (Text Mining Library – TML), and a technique of information retrieval (term frequency–inverse document frequency – TF-IDF) to measure connectivity in a text.

The text of every essay was parsed using Stanford CoreNLP 3.6.0 and Apache OpenNLP 1.6.0. The purpose was to build tables of words and n-grams (unigrams to five-grams using OpenNLP) for every essay. Each word was then tagged with its part of speech (using Stanford CoreNLP), lemma (Stanford CoreNLP), and stem (OpenNLP). According to Manning et al. (2008), the stem of a word is generally the word resulting from dropping its end according to a crude heuristic in order to remove derivational affixes. On the other side, the lemma of a word is its base or dictionary form without any inflectional ending. Table 3 displays the unigram table for the following sentence:

> “I believe that with all people laughter, and having a sense of humor, is something that generally everyone has in common, everyone loves to share with others.” – Essay #20725, Sentence 1

Table 3

Example of a Unigram Table

Index	Token	Part of speech	Stem	Lemma
1	I	Pronoun (PRP)	I	I
2	believe	Verb (VBP)	believ	believe
3	that	Preposition (IN)	that	that
4	with	Preposition (IN)	with	with
5	all	Article (DT)	all	all
6	people	Noun (NNS)	peopl	people
7	laughter	Noun (NN)	laughter	laughter
8	,	(,)	,	,

Table 3 (continued)

Example of a Unigram Table

Index	Token	Part of speech	Stem	Lemma
9	and	Conjunction (CC)	and	and
10	having	Verb (VBG)	have	have
11	a	Article (DT)	a	a
12	sense	Noun (NN)	sens	sense
13	of	Preposition (IN)	of	of
14	humor	Noun (NN)	humor	humor
15	,	(,)	,	,
16	is	Verb (VBZ)	is	be
17	something	Noun (NN)	someth	something
18	that	Determiner (WDT)	that	that
19	generally	Adverb (RB)	general	generally
20	everyone	Noun (NN)	everyon	everyone
21	has	Verb (VBZ)	has	have
22	in	Preposition (IN)	in	in
23	common	Adjective (JJ)	common	common
24	,	(,)	,	,
25	everyone	Noun (NN)	everyon	everyone
26	loves	Verb (VBZ)	love	love
27	to	(TO)	to	to
28	share	Verb (VB)	share	share
29	with	Preposition (IN)	with	with
30	others	Noun (NNS)	other	other
31	.	(.)	.	.

The next step was to indicate whether each word was either a spelling mistake or engaged in a grammatical error. The writing analytics tool used the LanguageTool (LT) spellchecker (version 3.4) (Miłkowski, 2010; Naber, 2003) to identify these two types of errors. It is important to note that the spellchecker's task to identify spelling errors is tricky because the spellchecker never knows with certainty whether an unknown word is actually a misspelling or just a word that is not an entry in its dictionary (e.g., a technical term). Fortunately, LT allows to retrieve 1) words that LT does not know (unknown), and 2) words that LT considers as misspellings. Misspelled words are identified as such when they are found as an entry in a bank of explicit misspellings. As for unknown words, LT

defines an unknown word when its tagger is unable to assign it a part of speech. The writing analytics tool took also some extra measures to ensure that any anonymized word or name of organization, location, or person (any named entity recognized as such by Stanford Named Entity Recognizer [NER]) was not considered as a spelling error. In addition to the core set of rules to identify spelling and grammatical errors, the writing analytics tool used an enhanced version of LT using Google n-gram data (8GB dataset including unigrams, bigrams, and trigrams) to look at the context of each word to determine whether the phrase has already been used in the literature and if so what is the probability that it has been correctly used in the current context. It is particularly useful to detect homophone errors. Table 4 shows how each word in the following sentence (the sentence was preceded by the title since it was the first sentence of the essay) has been tagged whether it was an unknown word, a misspelling, or a word engaged in a grammatical error:

> "Laugher – Laughter is to express delight, fun, or a object of a joke." – Essay #20727, Sentence 1

Table 4

Example of a Set of Words (One Sentence) with Tags Indicating if They Are Unknown Words, Have a Spelling Error, and if They Are Engaged in a Grammatical Error

Index	Token	Lemma	Unknown	Spelling error	Grammatical error
1	Laugher	Laugher	FALSE	TRUE	FALSE
2	Laughter	Laughter	FALSE	FALSE	FALSE
3	is	be	FALSE	FALSE	FALSE
4	to	to	FALSE	FALSE	FALSE
5	express	express	FALSE	FALSE	FALSE
6	delight	delight	FALSE	FALSE	FALSE
7	,	,	FALSE	FALSE	FALSE
8	fun	fun	FALSE	FALSE	FALSE
9	,	,	FALSE	FALSE	FALSE
10	or	or	FALSE	FALSE	FALSE
11	a	a	FALSE	FALSE	TRUE
12	object	object	FALSE	FALSE	FALSE

Table 4 (continued)

Example of a Set of Words (One Sentence) with Tags Indicating if They Are Unknown Words, Have a Spelling Error, and if They Are Engaged in a Grammatical Error

Index	Token	Lemma	Unknown	Spelling error	Grammatical error
13	of	of	FALSE	FALSE	FALSE
14	a	a	FALSE	FALSE	FALSE
15	joke	joke	FALSE	FALSE	FALSE
16	.	.	FALSE	FALSE	FALSE

Finally, the writing analytics tool computed the ratio of misspelled words over the total number of words and the ratio of words engaged in grammatical errors over the total number of words for every essay. In the example sentence above, there are 16 words (every punctuation symbol is considered as a "word"), and there is one grammatical error and one spelling error. Therefore, the ratios of spelling errors and grammatical errors are 1/16 = 0.0625.

Once unknown words and spelling and grammatical errors can be identified, it is then possible to assess the lexical diversity of each essay, that is, in that case counting the number of unique lemmas in every essay. For example, "continuous" and "continuously" have the same lemma "continuous" and are therefore considered as only one unique word despite their different parts of speech. Furthermore, to be part of the count, a word must neither be engaged in a grammatical error nor be a spelling mistake nor be an unknown word. Hence, the first sentence of Essay #20727 has 13 unique words and its lexical diversity (ratio of unique words over total number of words) is 13/16 = 0.8125.

To assess the baseline ability of students to employ effective transition phrases in their essays, a bank of 341 connecting words/phrases was constructed[5]. Since connectors may consist of one or more words (n-grams) and their

[5] University of Manitoba, Academic Learning Centre (2016). Connection words. Retrieved June 27, 2017, from https://umanitoba.ca/student/academiclearning/media/Connection_Words_NEW.pdf

Possel, H. Linking words. Retrieved June 27, 2017, from.http://www.smart-words.org/linking-words/

constituent words may have multiple parts of speech (e.g., prepositions, articles, adverbs, conjunctions), n-gram tables were created as previously explained. To count the number of connectors within an essay, the writing analytics tool iterated over the lists of unigrams, bigrams, trigrams, four-grams, and five-grams pertaining to that essay and searched for connecting words (phrases) present in the student's text by comparing each entry in those lists against the bank of connecting words. It is important to note that some inaccuracy was introduced due to the fact that the system counted the number of connectors by adding up the numbers of one-word, two-word, …, five-word connectors. For instance, "since then" is a two-word connector from which the connectors "since then", "since", and "then" can be derived. Thus, the overlap between the one-word, two-word, …, five-word groups was not subtracted from the total. For example, Table 5 lists the unigrams, bigrams, and trigrams of the following paragraph:

> "The purpose of this journal article is to show how a simple AES system can be built. In addition, it demonstrates that it is comparable to other state-of-the-art systems in spite of some of its limitations. Finally, it also provides an application in the classroom setting."

Table 5

An Example of Unigram, Bigram, and Trigram Tables Combined Together with Each Entry Indicating Whether It Is a Connecting Phrase

Unigram	Connector	Bigram	Connector	Trigram	Connector
The	FALSE	The purpose	FALSE	The purpose of	FALSE
purpose	FALSE	purpose of	FALSE	purpose of this	FALSE
of	FALSE	of this	FALSE	of this journal	FALSE
this	FALSE	this journal	FALSE	this journal article	FALSE
journal	FALSE	journal article	FALSE	journal article is	FALSE
article	FALSE	article is	FALSE	article is to	FALSE
is	FALSE	is to	FALSE	is to show	FALSE
to	TRUE	to show	FALSE	to show how	FALSE
show	FALSE	show how	FALSE	show how a	FALSE
how	TRUE	how a	FALSE	how a simple	FALSE
a	FALSE	a simple	FALSE	a simple AES	FALSE
simple	FALSE	simple AES	FALSE	simple AES system	FALSE
AES	FALSE	AES system	FALSE	AES system can	FALSE
system	FALSE	system can	FALSE	system can be	FALSE
can	FALSE	can be	FALSE	can be built	FALSE
be	FALSE	be built	FALSE	be built .	FALSE
built	FALSE	built .	FALSE	built . In	FALSE
.	FALSE	. In	FALSE	. In addition	FALSE

Table 5 (continued)

An Example of Unigram, Bigram, and Trigram Tables Combined Together with Each Entry Indicating Whether It Is a Connecting Phrase

Unigram	Connector	Bigram	Connector	Trigram	Connector
In	TRUE	In addition	TRUE	In addition ,	FALSE
addition	FALSE	addition ,	FALSE	addition , it	FALSE
,	FALSE	, it	FALSE	, it demonstrates	FALSE
it	FALSE	it demonstrates	FALSE	it demonstrates that	FALSE
demonstrates	FALSE	demonstrates that	FALSE	demonstrates that it	FALSE
that	TRUE	that it	FALSE	that it is	FALSE
it	FALSE	it is	FALSE	it is comparable	FALSE
is	FALSE	is comparable	FALSE	is comparable to	FALSE
comparable	FALSE	comparable to	FALSE	comparable to other	FALSE
to	TRUE	to other	FALSE	to other state-of-the-art	FALSE
other	FALSE	other state-of-the-art	FALSE	other state-of-the-art systems	FALSE
state-of-the-art	FALSE	state-of-the-art systems	FALSE	state-of-the-art systems in	FALSE
systems	FALSE	systems in	FALSE	systems in spite	FALSE
in	TRUE	in spite	FALSE	in spite of	TRUE
spite	FALSE	spite of	FALSE	spite of some	FALSE
of	FALSE	of some	FALSE	of some of	FALSE
some	FALSE	some of	FALSE	some of its	FALSE
of	FALSE	of its	FALSE	of its limitations	FALSE
its	FALSE	its limitations	FALSE	its limitations .	FALSE
limitations	FALSE	limitations .	FALSE	limitations . Finally	FALSE
.	FALSE	. Finally	FALSE	. Finally ,	FALSE
Finally	TRUE	Finally ,	FALSE	Finally , it	FALSE
,	FALSE	, it	FALSE	, it also	FALSE
it	FALSE	it also	FALSE	it also provides	FALSE
also	TRUE	also provides	FALSE	also provides an	FALSE
provides	FALSE	provides an	FALSE	provides an application	FALSE
an	FALSE	an application	FALSE	an application in	FALSE
application	FALSE	application in	FALSE	application in the	FALSE
in	TRUE	in the	FALSE	in the classroom	FALSE
the	FALSE	the classroom	FALSE	the classroom setting	FALSE
classroom	FALSE	classroom setting	FALSE	classroom setting .	FALSE
setting	FALSE	setting .	FALSE		
.	FALSE				
Total	9		1		1

From Table 5, it can be seen that the system counted 11 connectors, that there was one bigram connector (“in addition”) and a trigram connector (“in spite of”), and that the preposition “in” inside these connectors was counted as two extra connectors.

To compensate that, a connectivity index was developed in which a transformation was applied on the list of unique connectors extracted from the analyzed essay. The number of occurrences per unique connector (including one/two/…/five-word connectors) was counted for every sentence. TF-IDF was then applied to balance overuse of connectors such as the prepositions and those connectors that were derived from longer ones (such as the "in" in the "in addition" and "in spite of" connectors). The term frequency is the number of times a specific connector appears in a specific sentence, while the inverse document frequency is the logarithm of the total number of sentences divided by the number of sentences containing that connector (plus one for smoothing). Thus, each connector is assigned a weight that determines how much it contributes to the overall connectivity of the essay. The formula to compute the connectivity index of a text is:

$$\mathrm{CI} = \frac{\sum_{i,j} \mathrm{tf}(\text{connector i},\ \text{sentence j}) \times \mathrm{idf}(\text{connector i},\ \text{all sentences})}{\#\ \text{of sentences}}$$

where

$$\mathrm{tf}(\text{connector i}, \text{sentence j}) = \text{number of times connector i occurs in sentence j}$$

and

$$\mathrm{idf}(\text{connector i}, \text{all sentences}) = \log\left(\frac{\#\ \text{of sentences}}{1 + \text{number of sentences containing connector i}}\right)$$

The current scheme encourages using a variety of different connectors instead of repeating the same ones. Moreover, it also values having connectors in every sentence instead of having many connectors in a single sentence. Based on the frequency of each connector and their distribution over the sentences, a weight is assigned to every connector. The weights of all connectors are then added together and divided by the number of sentences in an essay. The minimum value that the index may generate is 0. There is no specified maximum limit on the index. The index values for the 723 processed essays from dataset #8 ranged between 0 and 2.35. Zero (0) denotes no connector at all, while an index close to 2 shows that there are many effective connectors in (almost) every sentence. The connectivity index is calculated for each essay.

The semantic similarity feature measures the semantic distance between each pair of consecutive sentences in an essay using LSA (Villalon & Calvo,

2013). The vocabulary of an essay is represented in a semantic space where the sentences are represented as vectors within that space. The distance between any pair of consecutive sentences is computed using cosine similarity where the distance value ranges between -1 and 1, inclusively. A value of 1 shows that the two consecutive sentences are identical or deal with the same topic (the sentence vectors are pointing exactly in the same direction based on the vector space model theory). A value of 0 means that the sentence vectors are 90° apart, thus demonstrating an important distance between the topics of the sentences. A value of -1 means that the sentence vectors are 180° apart, thus demonstrating that the two sentences are extremely unrelated. The average distance between all pairs of consecutive sentences is calculated for every essay.

The appendix lists the set of all writing features that were extracted by the writing analytics tool and that were part of the multiple regression analysis in this study. The purpose of this study is to assess the effectiveness of the following set of metrics in predicting the holistic scores of essay dataset #8: spelling and grammatical accuracy; semantic similarity and connectivity; and lexical diversity. Thus, the model's writing features can be categorized as cohesive devices (Crossley et al., 2016a), mechanics, and style.

4.3 Data Analysis

To answer the first research question, a multiple linear regression analysis was conducted. Since conducting a regression analysis is an iterative process where 1) a set of variables must first be selected, 2) a model must be formulated, and 3) the underlying assumptions of the model must be checked, several multiple starting models were generated using the R leaps package. The three best models for each size of model having between 1 and 11 independent variables were calculated. The original dataset consisted of 86 independent (also called explanatory or predictive) variables and one dependent (also known as predicted) variable, which was the essay scores (10-60 scale). These data types were described in 4.1 and 4.2. In order to assess whether the variables of the writing analytics tool's model (spelling accuracy, grammatical accuracy, semantic similarity, connectivity, and lexical diversity) form a good predictive model, the standard errors, the adjusted R^2 values, and the C_p Mallows values were calculated. The purpose of this regression analysis was not necessarily to select the most optimal model that generated, for example, the highest adjusted R^2 value. The objective was rather to select the smallest model having the highest predictive power.

To ensure that the individual predictive power of each variable in the model could be well assessed, some tests were performed to look for multicollinearity between the variables. First, the correlations between the

independent variables were evaluated. Second, the variance inflation factor (VIF) was calculated for every variable. A Shapiro-Wilk test was performed to test whether the distribution for each predictive and predicted variable came from a non-normal distribution. The QQ plots and histograms of all distributions were also examined. It was then decided that the coefficients of both Spearman (non-parametric) and Pearson (normality assumption) correlations would be included in the results. In addition, a matrix of scatterplots between the independent variables was generated to see if there was any obvious non-linear relationship.

The next step in the process of conducting a regression analysis was to formulate the model, that is, assess whether certain variables required some transformation. The scatterplots of every independent variable, along with both the residuals of the regression model and the essay score sample distribution, were analyzed. In addition, the λ constant was calculated to determine if and which one of the Box-Cox transformations should be applied. After having applied some transformations proposed by the scatterplots and the λ constant of the Box-Cox transformations, the resulting models were rejected since they did not significantly improve the agreement level among the regression model's predicted scores and the human essay scores. Moreover, departure from four regression assumptions slightly increased.

Finally, four assumptions underlying regression analyses were verified. The regression model's residuals were first standardized using studentized residuals to check whether the expected value of the residuals was equal to 0 and whether the residual variance was equal to 1. The plots of the standardized residuals against every independent variable were also examined. Second, to check for constant variance of residuals, a plot of the residuals versus the predicted values (essay scores) was examined. Since differentiating a good plot from a bad plot might be subjective, the Breusch-Pagan statistic (which assumes normality) was employed to test quantitatively the research hypothesis that there were heterogeneous variances among the residuals against the null hypothesis, which assumed homogeneous variances. Third, the QQ plot and histogram of the model's studentized residuals were inspected to see if there was any significant departure from normality among the residual distribution. Fourth, the Durbin-Watson test statistic was calculated to check whether there was an autocorrelation between residuals, that is, if the residuals were statistically independent. No major violations of these assumptions were observed even though there was a weak positive autocorrelation among the residuals of the regression model.

To answer the second research question, both Pearson and Spearman correlations were calculated to report the correlation between the resolved essay scores and predicted essay scores; the percentage of perfect matches between the resolved and predicted scores was calculated; the percentage of adjacent marks

between the computer and human raters was computed; the quadratic weighted kappa was derived to measure the strength of agreement among the AES system and human markers; and the standard deviation and mean of the predicted scores were generated.

5.0 Results

5.1 RQ1: What Is the Predictive Power of the Regression Model?

Table 6 lists the results of a regression subset analysis demonstrating the best model for each number of independent variables (up to 11 variables for a total of 11 models), the number of independent variables and the list of variables selected, the R^2 value and the adjusted R^2 value (i.e., the percentage of essay score variability explained by the model), the C_p Mallows value (the value the nearest to the number of variables in the model is the best), and the standard error (s_ε) for every model (the smallest value is the best).

Table 6

Best Subset Regression Models (1-11 Independent Variables) from a Set of 86 Independent Variables (Appendix)

Model ($\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_j x_j$)	# of var.	Adj. R^2	R^2	C_p	S_ε
$x_5, x_{29}, x_{31}, x_{35}, x_{42}, x_{43}, x_{61}, x_{62}, x_{73}, x_{75}, x_{83}$	11	0.613	0.619	15.08	3.78
$x_5, x_{29}, x_{31}, x_{35}, x_{42}, x_{43}, x_{61}, x_{62}, x_{73}, x_{82}$	10	0.610	0.616	19.03	3.80
$x_5, x_{29}, x_{31}, x_{35}, x_{42}, x_{43}, x_{61}, x_{62}, x_{73}$	9	0.608	0.613	22.65	3.82
$x_3, x_{29}, x_{31}, x_{35}, x_{61}, x_{62}, x_{74}, x_{81}$	8	0.601	0.605	34.71	3.85
$x_3, x_{29}, x_{31}, x_{35}, x_{61}, x_{62}, x_{81}$	7	0.597	0.601	40.66	3.87
$x_3, x_{31}, x_{35}, x_{61}, x_{62}, x_{81}$	6	0.588	0.591	56.34	3.92
$x_{30}, x_{42}, x_{61}, x_{62}, x_{86}$	5	0.573	0.576	83.49	3.99
$x_{31}, x_{42}, x_{61}, x_{86}$	4	0.560	0.563	106.64	4.05
x_{30}, x_{61}, x_{86}	3	0.536	0.538	150.76	4.17
x_{61}, x_{86}	2	0.503	0.505	211.30	4.32
x_{61}	1	0.456	0.456	299.52	4.52

The purpose of this regression analysis is to determine if the writing analytics tool's five-variable model (spelling accuracy, grammatical accuracy, semantic similarity, connectivity, and lexical diversity) forms a good predictive model and if the model's variables are uncorrelated enough to identify accurately poor-quality essays. In other words, is the writing analytics tool's model a proper

trade-off between the most optimal models generating, for example, the highest adjusted R^2 values and the models having the smallest numbers of independent variables? It can be seen from Table 6 that the adjusted R^2 value for a five-variable model ranges from 0.560 to 0.573. Hence, it is reasonable to conclude that the writing analytics tool's model is a proper trade-off between predictive power and number of variables in the model. From the list of the three best models derived from the regression subset analysis for each number of independent variables (1-11) (33 models in all), the second best five-variable model (which does not figure in Table 6) contained only variables that were related to the proposed model of the writing tool (spelling error ratio, grammatical error ratio, number of connectors, number of unique words, and semantic similarity):

$$y = \beta_0 + \beta_1 x_{29} + \beta_2 x_{31} + \beta_3 x_{42} + \beta_4 x_{61} + \beta_5 x_{86} + \varepsilon$$

Table 7 shows the adjusted R^2 and the standard error (s_ε) of the writing analytics tool's model. Approximately 57% of the essay score variability is accounted for by these five variables. Compared to an 11-variable model, which accounts for only 61% of the essay score variability, this model looks reasonably fitted to represent the predictive power of the writing analytics tool, while keeping a small number of independent variables in the regression model.

Table 7

Results of the Multiple Linear Regression Analysis on the Following Five Selected Writing Features

Variable	β_1	β_2
Intercept	24.9884	-1.152e-15
Ratio of misspelled words	- 38.8725	-0.1108
Ratio of grammatical errors	- 59.6438	-0.1707
Number of connectors	- 0.0365	-0.2780
Number of unique words	0.0700	0.7381
Semantic similarity	4.3892	0.1625

Note: $R^2 = 0.5725$; adj. $R^2 = 0.5695$; $s_\varepsilon = 3.775$; $F_{5,717} = 192$; $p < 0.001$. β_1's represent unstandardized coefficients, while β_2's represent standardized coefficients.

The next step was to analyze the model for significant evidence of multicollinearity; that is, it was important to determine if there was any significant correlation among the independent variables. To test for multicollinearity, the correlations among the independent variables were analyzed and the variance

inflation factor (VIF) was calculated for each of them. By analyzing the QQ plots and histograms of the five independent variables plus the dependent variable and by performing Shapiro-Wilk normality tests on each variable, it was seen that the distributions of the variables *number of connectors*, *number of unique words*, and *essay scores* did not depart significantly from normality. Thus, both the Pearson and Spearman coefficients were calculated and are presented in Table 8. Table 8 demonstrates that all correlations have significant results ($p < 0.05$) and that all correlations among independent variables are weak or moderate except for the correlation between the *number of connectors* and *number of unique words* variables, which have a strong relationship (0.76). Elliot et al. (2016) recommend interpreting correlation strength as follows: 0.00-0.29 as weak, 0.30-0.69 as moderate, and 0.70-1.00 as strong. Table 8 also shows the variance inflation factor (VIF) for each explanatory variable. All of them were between 1 and 3. Kleinbaum et al. (2013) suggest that there is strong evidence of multicollinearity in the explanatory variables when a value of VIF exceeds 10. Thus, there seemed to be enough evidence to conclude that multicollinearity was not a major problem with the variables of this regression model.

Table 7 reveals in addition that the variables *ratio of misspelled words*, *ratio of grammatical errors*, and *number of connectors* penalize the predicted essay scores, while the variables *number of unique words* and *semantic similarity* contribute to obtaining a better predicted score. By analyzing the standardized beta coefficients from Table 7, it is observed that the variable *number of unique words* contributes by far the most when determining the predicted score of an essay.

Table 8 shows two interesting correlations among the independent variables of the model. A moderate positive correlation (Spearman) of 0.40 exists between the *semantic similarity* and *number of unique words* variables and a strong correlation of 0.76 between the *number of connectors* and *number of unique words* variables. As for the correlations between the predictive variables and essay scores, the variables *ratio of misspelled words* and *ratio of grammatical errors* have weak-moderate correlations of -0.27 and -0.33, respectively, with the essay scores meaning that as the numbers of spelling and grammatical errors decrease, the essay score will tend to increase slightly. The variables *semantic similarity* and *number of connectors* have moderate correlations of 0.39 and 0.40, respectively. Finally, the *number of unique words* variable has a moderate-strong correlation with the essay scores of 0.68, which supports the same conclusion as the standardized beta coefficients of the regression analysis shown in Table 7.

Table 8

Pearson (Top-Right Half)/Spearman (Bottom-Left Half) Correlations & Variance Inflation Factors Among Study Variables

Variable	1	2	3	4	5	6
1. Ratio of misspelled words (VIF = 1.23)	-	0.22**	-0.25**	-0.32**	-0.40**	-0.38**
2. Ratio of grammatical errors (VIF = 1.13)	0.17**	-	-0.19**	-0.20**	-0.31**	-0.34**
3. Number of connectors (VIF = 2.39)	-0.13**	-0.15**	-	0.76**	0.35**	0.40**
4. Number of unique words (VIF = 2.76)	-0.21**	-0.17**	0.75**	-	0.49**	0.68**
5. Semantic similarity (VIF = 1.50)	-0.14**	-0.12**	0.23**	0.40**	-	0.52**
6. Essay score	-0.27**	-0.33**	0.35**	0.63**	0.39**	-

* $p < 0.05$; ** $p < 0.01$

5.2 RQ2: What Is the Agreement Level Between the Computer and Human Raters?

The agreement level of the writing analytics tool's model with human readers was assessed as per the following dimensions: correlation between the resolved and predicted essay scores, percentage of perfect matches, percentage of adjacent marks, the quadratic weighted kappa, and the distributional characteristics standard deviation and mean. After examining the histogram and QQ plot of the distribution of predicted essay scores and the result of the Shapiro-Wilk normality test, it was concluded that the distribution of predicted essay scores for the regression model departed significantly from normality. Hence, the Spearman correlation was selected and its coefficient between the distributions of predicted essay scores and resolved essay scores was calculated to be 0.72 ($p <$ 0.001). It was found that out of the 723 essays, only 70 essays had perfect matches between the regression model and human raters (10% of perfect agreement); 353 essays received adjacent marks (49% of adjacent agreement); and the quadratic weighted kappa value was found to be 0.73.

The range of essay scores for ASAP dataset #8 was 10-60. Given that the scale was very large, the agreement levels were also calculated on smaller integer scales (0-50, 0-10, 0-5). To convert to the 0-50 scale, the dataset of resolved scores was transformed by subtracting 10 from each resolved score. The regression analysis was re-run and generated the same results as for the 10-60

scale except for the quadratic weighted kappa, which decreased from 0.73 to 0.69. Both scales had the same number of score categories (51). The 0-10 integer scale (11 score categories) was derived by performing integer division (5) on every resolved score on the 0-50 scale. For example, if an essay score was 48 (out of 50), then the new score was 48 / 5 = 9 (out of 10). Similarly, the 0-5 integer scale (6 score categories) was derived by performing integer division (9) on every resolved score on the 0-50 scale. For instance, if the essay score was 37 (out of 50), then the new score was 37 / 9 = 4 (out of 5). Table 9 lists the agreement levels of the writing analytics tool for these four different scales.

Table 9

Agreement Levels of the Regression Model Proposed in this Study for Different Scales

Essay score scale	Exact agree	Adj. agree (±1)	Adj. agree (±2)	Quadratic weighted κ	Pearson r	Spearman r
10-60	0.10	0.31	0.49	0.73**	0.76**	0.72**
0-50	0.10	0.31	0.49	0.69**	0.76**	0.72**
0-10	0.45	0.94	-	0.69**	0.71**	0.68**
0-5	0.68	0.996	-	0.59**	0.60**	0.58**

* $p < 0.05$; ** $p < 0.01$

Out of the 723 essays, 416 essays actually received a mark below or equal to 65% and were considered as poor-quality. Among those 416 essays, the regression model detected 383 poor-quality essays (92%; true positives); 33 actual poor-quality essays were not detected (8%; false negatives). Three hundred seven essays actually received a grade greater than 65%. Out of these 307 essays, 154 essays were correctly not identified as being of poor quality (50%; true negatives), while 153 essays that were not of poor quality were considered as poor-quality (50%; false positives). The precision is, therefore, evaluated at 71% and the recall at 92%.

6.0 Discussion

The predictive model of the writing analytics tool analyzed in this study accounts for approximately 57% of the variability in essay scores. The study performed by McNamara et al. (2013) reported a regression model containing eight variables that explained 46% of the essay quality variability with agreement levels for perfect matches and adjacent matches of 44% and 94%, respectively. As for this study's regression model, Table 9 indicates that with a large scale (10-60)

the proposed model has a 10% exact agreement and a 49% adjacent agreement, while with a smaller scale (0-10), it has 45% exact agreement and 94% adjacent agreement. Although McNamara et al. did not report the scale used in their study, it could be inferred that the range of essay quality scores used in their study was rather small. Table 9 clearly shows that the range of the selected scale has a significant impact on the agreement levels between resolved and predicted scores. Furthermore, a greater adjusted R^2 value as suggested by this study does not seem to imply better exact and adjacent agreement levels when compared to McNamara et al.'s study.

It was found in the literature (Attali & Burstein, 2006; McNamara et al., 2015; Rudner et al., 2006) that the correlation between the predicted essay scores and human scores usually lies between 0.60 and 0.85. The correlation computed (0.72) in this study confirms these results. It is also reported in the literature that the ranges of perfect and adjacent agreement are 30-60% and 85-100%, respectively. Although the exact and adjacent agreement levels obtained for the 10-60 scale are much lower than what is reported in the literature (10% and 49%), when the range of the essay score scale is reduced to 0-10, the exact and adjacent agreement rates (45% and 94%) fit exactly in the ranges reported in the literature. It is, therefore, crucial that every study assessing the accuracy of an AES system reports its prediction scale. It is important to find the proper trade-off between the reliability of the predictions and their accuracy. A small scale increases the reliability of the results, but it caps its potential accuracy.

Table 10 lists the performance of 10 state-of-the-art AES systems as described by Shermis (2014) and Zupanc and Bosnic (2017), along with the performance of the presented regression model in this study. In terms of quadratic weighted kappas, SAGE ranks first, while this study's writing analytics tool shares the second rank with PEG. In terms of Pearson coefficients, the proposed model has the highest value although it is unknown for SAGE. As for the exact agreement rate, the proposed model ranks 9th out of the 11 AES systems, while it holds the 8th position out of 10 AES systems (since SAGE does not provide adjacent agreement measurements) for the adjacent agreement rate. Only 49% of all essays have a score predicted within a margin of ± 2. For instance, if the resolved score between human raters is 53/60 (88%), then this could imply that 49% of predictions will fall between 51/60 (85%) and 55/60 (92%), which is rather imprecise. The distributional characteristics of the writing analytics tool's distribution of predicted scores indicate that the mean predicted essay score is the closest to the mean of human scores, while its standard deviation is the 4th most distant. Given that the distribution of resolved essay scores does not depart significantly from normality, an AES system having a wide standard deviation may be desirable (but not at the expense of exact and adjacent agreement levels)

since it may prove its ability to identify students at both extremes (those who score very high and very low), which is very important in a classroom setting. It is important to remember that in a normal distribution, 68% of the observations are within one standard deviation (σ), 95% within two σ's, and 99.7% within three σ's. The standard deviation of this model is $\sigma = 4.36$, while the standard deviation of the resolved scores is $\sigma = 5.75$. The means of both resolved and predicted essay scores are 36.95 ($\approx 62\%$). According to Table 11, it can be seen that the regression model in this study predicts 99.7% of essay scores in a range of 40% and 83%.

Table 10

Test Set Means, Standard Deviations, Exact Agreements, Adjacent Agreements, Kappas, Quadratic Weighted Kappas, and Pearson Correlation Coefficients (Shermis, 2014)

	Means	SD	Exact agree	Adj. agree	κ	Quadratic weighted κ	Pearson r
Resolved scores	36.67	5.19	-	-	-	-	-
H1	36.45	5.93	0.35	0.53	0.26	0.75	0.87
H2	36.70	5.68	0.35	0.52	0.26	0.74	0.88
1. SAGE	-	-	0.16	-	-	0.81	-
2. PEG	37.23	5.38	0.16	0.52	0.10	0.73	0.73
3. This Study	36.95	4.36	0.10	0.49	-	0.73	0.76
4. e-rater	37.24	4.52	0.17	0.52	0.08	0.70	0.71
5. AutoScore	37.32	4.11	0.12	0.52	0.06	0.69	0.71
6. IEA	37.51	4.63	0.14	0.52	0.09	0.69	0.70
7. IntelliMetric	37.79	4.21	0.10	0.53	0.04	0.68	0.72
8. CRASE	37.04	5.16	0.20	0.48	0.11	0.68	0.68
9. LightSIDE	37.43	4.44	0.26	0.51	0.13	0.65	0.66
10. Lexile	37.54	5.91	0.08	0.41	0.04	0.63	0.62
11. Bookette	37.18	3.83	0.23	0.51	0.11	0.60	0.63

Table 11

Standard Deviations of Resolved and Predicted Essay Scores; Ranges of Essay Scores in Percentages for 1, 2, and 3 Standard Deviations

$\mu_{score} = 36.95$	68% of obs.		95% of obs.		99.7% of obs.	
	$-\sigma$	σ	-2σ	2σ	-3σ	3σ
$\sigma_{resolved} = 5.75$	52%	71%	42%	81%	33%	90%
$\sigma_{predicted} = 4.36$	54%	69%	47%	76%	40%	83%

Table 12 gives the average ranking of every AES system in relation to the five accuracy metrics discussed above (standard deviation, exact agreement, adjacent agreement, quadratic weighted kappa, Pearson correlation). PEG comes first, while the regression model developed in this study is classified in the 4th position. SAGE was discarded from the ranking given that it provided only the exact agreement and the quadratic weighted kappa measurements. The authors, Zupanc and Bosnic, are encouraged to report more about the accuracy of their system.

Table 12

Average Ranking of AES Systems According to Standard Deviation, Exact Agreement, Adjacent Agreement, Quadratic Weighted Kappa, and Pearson Correlation

	AES system	SD	Exact agree	Adj. agree	Quadratic weighted κ	Pearson r	Average rank
1	PEG	2	5	3.5	1.5	2	2.8
2	e-rater	5	4	3.5	3	4.5	4
3	IEA	4	6	3.5	4.5	6	4.8
4	This Study	7	8.5	8	1.5	1	5.2
5	IntelliMetric	8	8.5	1	6.5	3	5.4
6	AutoScore	9	7	3.5	4.5	4.5	5.7
7	CRASE	3	3	9	6.5	7	5.7
8	LightSIDE	6	1	6.5	8	8	5.9
9	Bookette	10	2	6.5	10	9	7.5
10	Lexile	1	10	10	9	10	8

The results of this study confirm the objections raised by Perelman (2013, 2014) and Deane (2013) that state-of-the-art AES systems are not mature enough to replace human scoring and that they largely award scores based on word counts. This study reports that the correlation (Pearson) between the number of unique words (type of word count) and the resolved scores is moderate-strong (0.68), the correlation (Pearson) between the number of connectors (another type of word count) and resolved scores is moderate (0.40), and the correlation (Pearson) between number of unique words and number of connectors is strong (0.76). These are the strongest correlation coefficients among the five variables of the regression model derived in this study, which play an important role in the prediction of essay scores. Besides, the standardized coefficients of the multiple linear regression model confirm that the number of unique words and the number of connectors have the largest magnitudes suggesting that they contribute the most to the determination of predicted scores. The small contribution of semantic similarity corroborates previous findings that cohesive devices do not play an important role when determining essay quality (Crossley et al., 2016a, 2016b).

Although the literature reports and this study demonstrates that state-of-the-art AES systems roughly assess text production skills and do not measure the quality of discourse features such as strength of argumentation, rhetorical effectiveness, and attention to audience, can they still be of any help to the teacher in a classroom setting (e.g., automatically identifying poor-quality essays)? As a complement to this study, the authors investigate if the proposed writing analytics tool is accurate enough to separate the 723 essays of ASAP dataset #8 into two categories according to their predicted essay scores, that is, those essays below a certain threshold score and those above that threshold score. Two thresholds have been analyzed, one slightly greater than the mean of resolved scores and the other slightly smaller than the mean ($\mu_{resolved} = 36.95$). The first threshold was set at 39/60 (65%) and the second at 36/60 (60%) (scale is 10-60). For example, all essays, whose score is below 40, are separated from those essays having 40 or more. Table 13 lists the precision and recall rates of both classifiers along with the performance of their corresponding majority classifier. A majority classifier simply assigns every essay to the set having the greatest number of essays. For example, 416 out of the 723 essays have scores below 40 and 307 essays have scores equal to or greater than 40. Since the majority of essays have scores below 40, then the majority classifier will classify (predict) all essays as having less than 40. The precision and recall of this majority classifier are reported to be 58% and 100%. This means that among all essays that were predicted to have a score less than 40, 58% of those essays were actually below 40 and that all essays (100%) having actually less than 40 were correctly reported as having a score smaller than 40. More specifically, lower precision overestimates the number of poor-quality

essays, while lower recall underestimates that number. It is always important to find the proper trade-off between both metrics. In this situation, the best trade-off appears to be the 65%-threshold classifier with a precision of 71% and accuracy of 92%. While it is beyond the scope of this study to provide a comprehensive background to the reader regarding the usefulness of performing such a dichotomy and what would be the best criteria to perform such a dichotomy, it still shows that using a so-called state-of-the-art AES system to identify poor-quality essays does not prove to be an optimal solution.

Table 13

Classifiers of Poor-Quality Essays

	60%-threshold classifier	60%-majority classifier	65%-threshold classifier	65%-majority classifier
N	723	723	723	723
True positives	177	284	383	416
False positives	52	439	153	307
True negatives	387	0	154	0
False negatives	107	0	33	0
Precision	77.29%	60.72%	71.46%	57.54%
Recall	62.32%	100%	92.07%	100%

7.0 Conclusions

This study analyzed the predictive power and accuracy of a new AES system based only on five writing features (spelling accuracy, grammatical accuracy, semantic similarity, connectivity, and lexical diversity). It showed in a transparent manner how the proposed AES system is comparable to other state-of-the-art AES systems. A literature review listed the state-of-the-art automated essay scoring systems and reported their range of accuracy in terms of quadratic weighted kappa. The implementation process of AES was also described: 1) identification and extraction of writing features (three types of writing features: style, content, and semantic), 2) creation of predictive models, and 3) actual essay scoring. AES systems were then categorized in four categories of implementation methods: hybrid, LSA-based (latent semantic analysis), TCT-based (text categorization techniques), and miscellaneous. Automated essay scoring was also classified as either holistic scoring or trait scoring. This study presented, in addition, different methods to measure the accuracy of an AES system: correlation, perfect agreement, adjacent agreement, and the quadratic weighted

kappa. Finally, the flaws of AES were acknowledged (e.g., assessment of text production features instead of discourse ones, predictions based on word counting) (Deane, 2013; Perelman, 2013, 2014), as well as the limitations underlying the studies reported by Shermis (2014).

This study performed a multiple linear regression analysis on the eighth dataset (723 narrative essays written by Grade 10 students) provided in the context of the ASAP demonstration. The essays were further processed by the writing analytics tool proposed in this study, and 86 writing features were extracted in all. One goal of the analysis was to verify if the writing analytics tool's model was representative of its potential predictive power. After having executed a regression subset analysis and verified that there were no major violations in four underlying regression assumptions, it was found that the five-variable model was representative of the tool's predictive power and explained approximately 57% of essay score variability. In addition, no serious issue of multicollinearity was reported allowing to identify the *number of unique words* variable as a moderate-strong predictor of essay scores with a correlation of 0.68.

The accuracy of the writing analytics tool was evaluated along the following dimensions: Pearson/Spearman correlation, exact agreement, adjacent agreement, quadratic weighted kappa, and the distributional characteristics standard deviation and mean. The results of the study showed that there was a 0.76 correlation (Pearson) between resolved essay scores (human) and predicted essay scores, 10% of the essays were given the same score by both the computer and human raters, 49% of the essays were given adjacent marks (± 2) by the computer and human raters, and the quadratic weighted kappa computed was 0.73. The results were measured on an essay score scale of 10-60. The analysis underscores the importance of specifying the scale when measuring the accuracy/reliability of an AES system. The findings of this study are supported by the literature.

The main contribution of this study consists in shedding light on the development process of a relatively simple AES system that compares very well to state-of-the-art AES systems. Since AES systems enclose very prized intellectual property, very few details are unveiled in terms of implementation, analysis techniques, and accuracy measurements. It is hoped through this study that this will reveal what lies behind so-called state-of-the-art AES systems, that is, shallow assessment of text production skills instead of discourse features such as argumentation and rhetoric. The results of this study uphold the objections expressed by Perelman (2013, 2014) and Deane (2013). In addition, this study demonstrates how the proposed writing analytics tool struggled in the simple identification of poor-quality essays (discovering all essays scoring below or equal to a certain threshold). This study provides additional evidence that machine

scoring is not ready to replace human scoring for the moment. However, analytics-based, large-scale causal models have the potential to track and measure cognitive traits associated with writing, paving way for future AES systems to approach the quality of human grading.

8.0 Directions for Further Research

To help the future development of AES, the collective understanding of the relationship between the explanation of essay score variability and the inter-rater agreement level should be further investigated to better represent the changes in terms of level of agreement when a new variable is added to a regression model instead of deciding on more arbitrary statistical values like standard errors and adjusted R^2 values.

The impact of anonymizing personally identifiable information should also be investigated further to assess whether it significantly impacts the accuracy of AES. For instance, the presence of anonymized tags (e.g., @PERSON) in a sentence could diminish the capacity of the natural language processing software to extract correctly the sentence structure. Hypothetically, it could generate more grammatical errors. Besides, this study should be replicated at a larger scale and extended if possible to various writing settings (institution, grade, English as native/second language, type of writing, demographic characteristics).

Researchers in the writing analytics community are encouraged to contribute to making the process of developing AES systems more transparent so that a clearer understanding of the predictive power of every major writing feature in their respective writing context can be communicated to the community at large. Researchers are invited to share further details on their implementation processes given that it is not sufficient to know the correlation between the output of a writing feature calculated through an imperfect machine learning algorithm and essay quality scores. The actual correlation between a writing feature as assessed by human raters and human essay scores should be first reported followed by the accuracy measurement of the analytic technique used to automate the extraction of the writing feature. Reporting the predictive power of a writing feature based solely on the results of an in-progress algorithm can prove misleading. It will also be important to focus on finding new metrics that will be highly uncorrelated to the bank of existing writing features and highly correlated to essay scores. It is expected that this will help partly explain the unexplained sources of variability in essay scores. Finally, predictive models should be developed at a finer level of granularity in order to improve holistic scoring as well as the generation of more precise feedback. Again, high percentages of the variability in the rubric scores should be elicited and explained by the most

relevant writing features. Researchers could consider conducting causal analyses instead of "correlational" regression analyses to better identify how these variables interact (Clemens, 2017). Identification of potential suppressor variables may be important when it comes to assessing writing. Their impact among independent variables may improve predictive accuracy of a regression model, even when there is no correlation between the independent and dependent variables (Huck, 2009).

Measuring the connectivity of a text using TF-IDF (where individual sentences are treated as "documents") as proposed in this study needs further testing to ensure it accurately assesses effective usage of connectors. Researchers could exploit the normal distribution of the number of connectors to improve their correlation with essay scores (the same is true for any writing feature). Connectors are key for the cohesion of an essay. Even though it has already been shown that cohesive devices (Crossley et al., 2016a, 2016b) play a limited role in prediction of essay scores, new algorithms could challenge the current status quo and shed light on unknown sources of predictability. Finally, the area of discourse analysis remains wide open for new advances by the research community.

Author Biographies

Dr. Vivekanandan Kumar is a Professor in the School of Computing and Information Systems at Athabasca University, Canada. He holds the Natural Sciences and Engineering Research Council of Canada's (NSERC) Discovery Grant on Anthropomorphic Pedagogical Agents, funded by the Government of Canada. His research focuses on developing anthropomorphic agents, which mimic and perfect human-like traits to better assist learners in their regulatory tasks. His research includes investigating technology-enhanced erudition methods that employ big data learning analytics, self-regulated learning, co-regulated learning, causal modeling, and machine learning to facilitate deep learning and open research. For more information, visit http://vivek.athabascau.ca.

Dr. Shawn N. Fraser is an Associate Professor at Athabasca University and an Adjunct Assistant Professor in Physical Education and Recreation at the University of Alberta. His research interests include understanding how stress can impact upon rehabilitation success for heart patients. He teaches research methods courses in the Faculty of Health Disciplines and is interested in interdisciplinary approaches to studying and teaching research methods and data analysis.

David Boulanger is a student and data scientist involved in the learning analytics research group at Athabasca University. His primary research focus is on

observational study designs and the application of computational tools and machine learning algorithms in learning analytics including writing analytics.

References

Aluthman, E. S. (2016). The effect of using automated essay evaluation on ESL undergraduate students' writing skill. *International Journal of English Linguistics, 6*(5), 54–67. http://doi.org/10.5539/ijel.v6n5p54

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment, 4*(3), 1–31. Retrieved from https://www.ets.org/research/policy_research_reports/publications/article/2006/hsjv

Bennett, R. E. (2011). CBAL: Results from piloting innovative K–12 assessments. *ETS Research Report Series, 2011*(1), 1–39. http://doi.org/10.1002/j.2333-8504.2011.tb02259.x

Brenner, H., & Kliebsch, U. (1996). Dependence of weighted kappa coefficients on the number of categories. *Epidemiology, 7*(2), 199–202. Retrieved from http://www.jstor.org/stable/3703036

Chen, H., He, B., Luo, T., & Li, B. (2012). A ranked-based learning approach to automated essay scoring. In *2012 Second International Conference on Cloud and Green Computing* (pp. 448–455). http://doi.org/10.1109/CGC.2012.41

Clemens, C. (2017). *A causal model of writing competence* (Master's thesis). Retrieved from https://dt.athabascau.ca/jspui/handle/10791/233

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

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

Crossley, S. A., & McNamara, D. S. (2011). Text coherence and judgments of

essay quality: Models of quality and coherence. In *Proceedings of the Cognitive Science Society* (Vol. 33, No. 33). Retrieved from https://www.semanticscholar.org/paper/Text-Coherence-and-Judgments-of-Essay-Quality-Mode-Crossley-McNamara/89c191a8053412356eb8a68144ca59d8b5eb6a63

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

El Ebyary, K., & Windeatt, S. (2010). The impact of computer-based feedback on students' written work. *International Journal of English Studies*, *10*(2), 121–142. Retrieved from http://revistas.um.es/ijes/article/view/119231

Elliot, N., Rudniy, A., Deess, P., Klobucar, A., Collins, R., & Sava, S. (2016). ePortfolios: Foundational measurement issues. *Journal of Writing Assessment*, *9*(2). Retrieved from http://journalofwritingassessment.org/article.php?article=110

Fazal, A., Hussain, F. K., & Dillon, T. S. (2013). An innovative approach for automatically grading spelling in essays using rubric-based scoring. *Journal of Computer and System Sciences*, *79*(7), 1040–1056. https://doi.org/10.1016/j.jcss.2013.01.021

Foltz, P. W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods*, *28*(2), 197–202. http://doi.org/10.3758/BF03204765

Gebril, A., & Plakans, L. (2016). Source-based tasks in academic writing assessment: Lexical diversity, textual borrowing and proficiency. *Journal of English for Academic Purposes*, *24*, 78–88. https://doi.org/10.1016/j.jeap.2016.10.001

Gregori-Signes, C., & Clavel-Arroitia, B. (2015). Analysing lexical density and lexical diversity in university students' written discourse. *Procedia - Social and Behavioral Sciences*, *198*, 546–556. https://doi.org/10.1016/j.sbspro.2015.07.477

Huck, S. (2009). *Statistics misconceptions*. New York, NY: Taylor & Francis.

Kleinbaum, D., Kupper, L., Nizam, A., & Rosenberg, E. (2013). *Applied*

regression analysis and other multivariable methods. Nelson Education.

Larkey, L. S. (1998). Automatic essay grading using text categorization techniques. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 90–95). New York, NY, USA: ACM. http://doi.org/10.1145/290941.290965

Latifi, S., Gierl, M. J., Boulais, A.-P., & De Champlain, A. F. (2016). Using automated scoring to evaluate written responses in English and French on a high-stakes clinical competency examination. *Evaluation & the Health Professions*, *39*(1), 100–113. http://doi.org/10.1177/0163278715605358

Manning, C. D., Raghavan, P., Schütze, H., & others. (2008). *Introduction to information retrieval* (Vol. 1). Cambridge: Cambridge University Press.

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, *45*(2), 499–515. http://doi.org/10.3758/s13428-012-0258-1

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, *23*, 35–59. http://doi.org/10.1016/j.asw.2014.09.002

Miłkowski, M. (2010). Developing an open-source, rule-based proofreading tool. *Software: Practice and Experience*, *40*(7), 543–566. http://doi.org/10.1002/spe.971

Naber, D. (2003). A rule-based style and grammar checker. Retrieved from https://www.researchgate.net/publication/239556866_A_Rule-Based_Style_and_Grammar_Checker

Perelman, L. (2013). Critique of Mark D. Shermis & Ben Hammer, Contrasting state-of-the-art automated scoring of essays: Analysis. *Journal of Writing Assessment*, *6*(1). Retrieved from http://journalofwritingassessment.org/article.php?article=69

Perelman, L. (2014). When “the state of the art” is counting words. *Assessing Writing*, *21*, 104–111. http://doi.org/10.1016/j.asw.2014.05.001

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of IntelliMetric™ essay scoring system. *The Journal of Technology, Learning and Assessment*, *4*(4), 1–22. Retrieved from https://ejournals.bc.edu/ojs/index.php/jtla/article/download/1651/1493

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, *20*(1), 53–76. http://doi.org/10.1016/j.asw.2013.04.001

Slomp, D. H. (2012). Challenges in assessing the development of writing ability: Theories, constructs and methods. *Assessing Writing*, *17*(2), 81–91. https://doi.org/10.1016/j.asw.2012.02.001

Villalon, J., & Calvo, R. A. (2013). A decoupled architecture for scalability in text mining applications. *Journal of Universal Computer Science*, *19*(3), 406–427. Retrieved from https://pdfs.semanticscholar.org/ffcc/204e98f16fd0fc47af8c2ff312f5f50df81d.pdf

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, *10*(2), 157–180. https://doi.org/10.1191/1362168806lr190oa

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Wilson, J., Olinghouse, N. G., McCoach, D. B., Santangelo, T., & Andrada, G. N. (2016). Comparing the accuracy of different scoring methods for identifying sixth graders at risk of failing a state writing assessment. *Assessing Writing*, *27*, 11–23. http://doi.org/10.1016/j.asw.2015.06.003

Zupanc, K., & Bosnić, Z. (2017). Automated essay evaluation with semantic analysis. *Knowledge-Based Systems*, *120*, 118–132. https://doi.org/10.1016/j.knosys.2017.01.006

Appendix

The 86 Writing Features Extracted by the Proposed Writing Analytics Tool

1. Number of characters (x_1)
2. Number of words (x_2)
3. Number of sentences (x_3)
4. Average number of characters per word ($x_4 = x_1/x_2$)
5. Average number of words per sentence ($x_5 = x_2/x_3$)
6. Number of anonymized words (x_6)
7. Number of anonymized words over total number of words (ratio) ($x_7 = x_6/x_2$)
8. Number of words expressing tone (x_8)
9. Number of words expressing tone over total number of words (ratio) ($x_9 = x_8/x_2$)
10. Number of words expressing a positive sentiment (x_{10})
11. Number of words expressing a positive sentiment over total number of words (ratio) ($x_{11} = x_{10}/x_2$)
12. Number of words expressing a negative sentiment (x_{12})
13. Number of words expressing a negative sentiment over total number of words ($x_{13} = x_{12}/x_2$)
14. Number of positive-sentiment words expressing happiness, gladness, and joy (x_{14})
15. Number of positive-sentiment words expressing happiness, gladness, and joy over total number of words (ratio) ($x_{15} = x_{14}/x_2$)
16. Number of positive-sentiment words expressing amazement and beauty (x_{16})
17. Number of positive-sentiment words expressing amazement and beauty over total number of words (ratio) ($x_{17} = x_{16}/x_2$)
18. Number of positive-sentiment words expressing fame and popularity (x_{18})
19. Number of positive-sentiment words expressing fame and popularity over total number of words (ratio) ($x_{19} = x_{18}/x_2$)
20. Number of negative-sentiment words expressing fright and annoyment (x_{20})
21. Number of negative-sentiment words expressing fright and annoyment over total number of words (ratio) ($x_{21} = x_{20}/x_2$)

22. Number of negative-sentiment words expressing evil and meanness (x_{22})
23. Number of negative-sentiment words expressing evil and meanness over total number of words (ratio) ($x_{23} = x_{22}/x_2$)
24. Number of negative-sentiment words expressing mischievousness and craziness (x_{24})
25. Number of negative-sentiment words expressing mischievousness and craziness over total number of words (ratio) ($x_{25} = x_{24}/x_2$)
26. Number of negative-sentiment words expressing sadness and terror (x_{26})
27. Number of negative-sentiment words expressing sadness and terror over total number of words (ratio) ($x_{27} = x_{26}/x_2$)
28. Number of spelling errors (x_{28})
29. Number of spelling errors over total number of words (ratio) ($x_{29} = x_{28}/x_2$)
30. Number of grammatical errors (x_{30})
31. Number of grammatical errors over total number of words (ratio) ($x_{31} = x_{30}/x_2$)
32. Number of unknown words (x_{32})
33. Number of unknown words over total number of words (ratio) ($x_{33} = x_{32}/x_2$)
34. Number of words with 5 characters and more (x_{34})
35. Number of words with 5 characters and more over total number of words (ratio) ($x_{35} = x_{34}/x_2$)
36. Number of words with 6 characters and more (x_{36})
37. Number of words with 6 characters and more over total number of words (ratio) ($x_{37} = x_{36}/x_2$)
38. Number of words with 7 characters and more (x_{38})
39. Number of words with 7 characters and more over total number of words (ratio) ($x_{39} = x_{38}/x_2$)
40. Number of words with 8 characters and more (x_{40})
41. Number of words with 8 characters and more over total number of words (ratio) ($x_{41} = x_{40}/x_2$)
42. Number of connectors (x_{42})
43. Average number of connectors per sentence (ratio) ($x_{43} = x_{42}/x_3$)
44. Number of connectors expressing effect, result, and consequence (x_{44})

45. Number of connectors expressing effect, result, and consequence over total number of connectors (ratio) ($x_{45} = x_{44}/x_{42}$)
46. Number of connectors expressing opposition, limitation, and contradiction (x_{46})
47. Number of connectors expressing opposition, limitation, and contradiction over total number of connectors (ratio) ($x_{47} = x_{46}/x_{42}$)
48. Number of connectors expressing cause, condition, and purpose (x_{48})
49. Number of connectors expressing cause, condition, and purpose over total number of connectors (ratio) ($x_{49} = x_{48}/x_{42}$)
50. Number of connectors expressing space, location, and place (x_{50})
51. Number of connectors expressing space, location, and place over total number of connectors (ratio) ($x_{51} = x_{50}/x_{42}$)
52. Number of connectors expressing time, chronology, and sequence (x_{52})
53. Number of connectors expressing time, chronology, and sequence over total number of connectors (ratio) ($x_{53} = x_{52}/x_{42}$)
54. Number of connectors expressing conclusion, summary, and restatement (x_{54})
55. Number of connectors expressing conclusion, summary, and restatement over total number of connectors (ratio) ($x_{55} = x_{54}/x_{42}$)
56. Number of connectors expressing agreement, addition, and similarity (x_{56})
57. Number of connectors expressing agreement, addition, and similarity over total number of connectors (ratio) ($x_{57} = x_{56}/x_{42}$)
58. Number of connectors expressing examples, support, and emphasis (x_{58})
59. Number of connectors expressing examples, support, and emphasis over total number of connectors (ratio) ($x_{59} = x_{58}/x_{42}$)

60. Index of connectivity (x_{60})
61. Number of unique words (x_{61})
62. Lexical diversity (ratio) ($x_{62} = x_{61}/x_2$)
63. Number of adjectives (x_{63})
64. Number of adjectives over total number of words (ratio) ($x_{64} = x_{63}/x_2$)
65. Number of adverbs (x_{65})
66. Number of adverbs over total number of words (ratio) ($x_{66} = x_{65}/x_2$)
67. Number of articles (x_{67})
68. Number of articles over total number of words (ratio) ($x_{68} = x_{67}/x_2$)
69. Number of conjunctions (x_{69})
70. Number of conjunctions over total number of words (ratio) ($x_{70} = x_{69}/x_2$)
71. Number of determiners (x_{71})
72. Number of determiners over total number of words (ratio) ($x_{72} = x_{71}/x_2$)
73. Number of nouns (x_{73})
74. Number of nouns over total number of words (ratio) ($x_{74} = x_{73}/x_2$)
75. Number of pronouns (x_{75})
76. Number of pronouns over total number of words (ratio) ($x_{76} = x_{75}/x_2$)
77. Number of prepositions (x_{77})
78. Number of prepositions over total number of words ($x_{78} = x_{77}/x_2$)
79. Number of verbs (x_{79})
80. Number of verbs over total number of words (ratio) ($x_{80} = x_{79}/x_2$)
81. Number of content words (x_{81})
82. Number of content words over total number of words (ratio) ($x_{82} = x_{81}/x_2$)
83. Number of function words (x_{83})
84. Number of function words over total number of words (ratio) ($x_{84} = x_{83}/x_2$)
85. Reading difficulty (Flesch-Kincaid) (x_{85})
86. Semantic similarity (x_{86})