National and Kapodistrian University of Athens

Department of English Language and Literature

MA Programme "Linguistics: Theory and Applications"

A Statistical Approach to Automatic Essay Scoring

Student's Name: Angeliki Nektariou

ID number: 217019

Committee

Prof. G. Mikros (supervisor)

Prof. B. Mitsikopoulou

Prof. N. Lavidas

Date of submission: 8/2/2019

Declaration

This submission is my own work. Any quotation from, or description of, the work of others is acknowledged herein by reference to the sources, whether published or unpublished.

Acknowledgements

I would like to express my gratitude to my supervisor, Professor George Mikros, for an inspiring introduction to the field of Computational Linguistics. My sincerest thanks are also due for his continued guidance and support throughout the course of this project.

Last but not least, I feel indebted to all my academic Instructors for generously offering their expertise both at the undergraduate and at the postgraduate program of the Department.

Abstract

Taking into consideration escalating need for testing writing ability and the potential of Automatic Essay Scoring (AES) to support writing instruction and evaluation, the aim of the present study is to explore the relationship between stylometric indices, widely used in AES systems, and the degree of sophistication of learner essays, captured by the score provided by expert human raters. The data analyzed were obtained from a recently organized public AES competition and comprise persuasive essays written in the context of public school in the United States. Stylometric information taken into consideration greatly focuses on measures of cohesion, as well as lexical diversity and syntactic sophistication. Results indicate a clear relationship between quantifiable features of learners' written responses and the impression which they have made on expert raters. This observation reinforces the importance of pursuing further experimentation into AES, which would yield significant educational and social benefits.

Table of Contents

1.Introduction

Essay writing has long been established as a significant instructional activity, as it incorporates both opportunities for language development and critical and analytical skills on the part of the learner. Writing, one of the core skills which second and foreign language teaching and learning aims at developing, forms a prominent part both of language syllabi and language testing rationales, where it is now the norm for the test-taker to be expected to respond, among other items, to open-ended questions in writing.

The significance of writing tasks in the language classroom and the need for widespread testing of language ability for education or immigration purposes place high demands on schools and examination boards regarding assessment of written responses. The need to score very large numbers of long-form responses manually may discourage educational contexts from engaging learners in such tasks constantly and complicate the process of ensuring the reliability of scores assigned and the provision of timely feedback. Additionally, the need for scores to be provided by multiple raters in high-stakes examinations further contributes to the significant resources necessary for evaluating candidates' essays manually.

Additional developments in the current educational landscape, such as the frequent engagement of large numbers of learners and the provision of ample educational opportunities in the context of digital learning environments, such as Massive Open Online Courses (MOOCs), further emphasizes the need to better tailor systems for scoring and for the provision of feedback to changing educational needs.

This need for assistance in the evaluation process in various instructional and examination contexts highlights the importance of investigating automatic essay scoring as an alternative means for the evaluation of constructed responses. The employment of Natural Language Processing techniques, which enable machine processing of discourse, appears promising in this direction.

Skepticism sometimes surrounds evaluations of discourse by machines, due to the lack of a capacity for machines to process discourse in the same way as human speakers. This observation sometimes leads to dismissing automatically produced scores as inherently lacking validity. However, the performance of human raters has been observed to be sensitive to factors including fatigue, halo effects, inadvertent bias or individual differences, which may affect the validity and reliability of scores provided by human raters (Shi 2001). The question then arises of whether manually assessing very large numbers of essays indeed constitutes the best possible use of time and energy on the part of the instructor. Incorporating automated systems to support educational activities appears a promising way to boost learner performance and maximize the instructor's active involvement in more meaningful pedagogical activities within the context of writing lessons which would directly involve learners with a view to contributing to the development of their writing skills.

The present study will, therefore, explore the interaction between information of a stylometric nature, central in automatic essay scoring, and features of written discourse quality attested in learner writing, as these have been evaluated through scores assigned by human experts. Through experimentally exploring this relationship, our aim is to model the predictive power of combinations of stylometric indices in

relation to the score awarded by human experts. This process will allow us to discuss the significance of specific stylometric indices, as we aim to identify through our experimentation which quantifiable stylistic features of texts appear to most strongly correlate with essay scores provided by human experts, and therefore appear to most closely reflect discourse qualities.

We will begin with a discussion of theoretical accounts of discourse organization, exploring different approaches to the notion of cohesion, which have informed our methodological approach in the present study. As will be discussed in subsequent chapters, we expect various aspects of cohesion and coherence to have received significant attention, among other factors, including lexical diversity and syntactic sophistication, in the evaluation of written work on the part of human assessors. Therefore, as will be seen in our experimental design, we have concentrated our efforts on capturing, from a stylometric perspective, information whose relationship with evaluations of discourse qualities by human readers originates in theoretical accounts of features at work in discourse organization.

After discussing cohesion and coherence, and the role of discourse organization in comprehension of the text on the part of the reader, we will provide a cursory overview of the background of automated scoring systems, with a brief indicative description of the design of a few engines central in the development of the field or in operational use today. Subsequently, we will discuss considerations of the benefits and limitations of automated evaluation systems in relation to their social and educational significance. We will then proceed to reporting two experiments conducted with a view to investigating the relationship between our stylometric analysis of a corpus of learner script and the score which the texts in this corpus were awarded by human experts. After discussing the predictive strength of our resulting models, we will analyze the

manner in which stylometric information has been found, through our experiments, to be closely related to discourse qualities by analyzing the nature of those stylometric indices which appeared to have the greatest predictive power of the score assigned by human readers, and the relationship of these indices with the discourse qualities which they are intended to represent.

2. Text and cohesion

Halliday and Hasan ([1976] / 2013: 1, 2) define the notion of *text* as a unit of meaning which is enabled through the linguistic code, so that its surface realization is the set of sentences that it comprises but its essence is its underlying semantic dimension. When presented with a stretch of discourse, a speaker of a language is able to identify, with some degree of certainty, whether or not it constitutes a coherent text on the basis of the unity of its meaning. This quality represents *texture* (Halliday & Hasan [1976] / 2013: 2), the state of being a coherent text, or in other words representing a rather concentrated unit of meaning which can be distinguished from its environment in a relatively straightforward manner.

Texture, the quality of being a text, as defined by Halliday and Hasan ([1976] / 2013: 8) is realized through cohesion, which is necessary for presupposing and presupposed elements to coexist in the text and thus ensure its unity of meaning. In other words, a cohesive device is not functional in isolation from some other element in the text, whose presence it presupposes by referring to it. For instance, more than one textual elements which refer to the same entity are said to constitute a case of *co-reference*, through which texture is achieved (Halliday & Hasan [1976] / 2013: 5). In this sense,

"[c]ohesion is a semantic relation between an element in the text and some other element that is crucial to the interpretation of it." (Halliday & Hasan [1976] / 2013: 8)

As a semantic relation, coherence is realized through the lexicogrammar, with some elements of cohesion being realized through grammar and others through vocabulary, but it is not intrinsically related to form of the text. Therefore, cohesion can be found within the sentence or across more sentences in the text. The linguistic form of the text is thus not directly related to cohesion but enables the presence of cohesive devices, which facilitate the reader in reconstructing the semantic unit which the text represents (Halliday & Hasan [1976] / 2013: 8, 9).

Reference (Halliday & Hasan [1976] / 2013: 31) is described as one quality of certain discourse elements such as pronouns, demonstratives and comparatives, which are not bearers of complete meaning in their own right, but rather refer to something else which is present in the text. In this sense, reference contributes to cohesion as it overtly captures the relationship between different concepts in the text.

Cohesion in the text can additionally be achieved through *substitution* and *ellipsis* (Halliday & Hasan [1976] / 2013: 85), which are described as processes between elements of the text resulting in one item appearing in the text instead of another or an item being omitted from the text. Substitution, entailing nominal, verbal or clausal relationships in the text, is distinct from reference in that it is understood as a process related to words, whereas reference is described as a process related to meaning. Ellipsis (Halliday & Hasan [1976] / 2013: 142) is related to information which is not explicitly expressed in the text through its encoding in lexicogrammatical form. However, ellipsis is not related to meaning which is not mentioned and therefore not understood, but rather to meaning which is communicated despite not being overtly present in the physical text. This process constitutes an aspect of cohesion, since understanding the omitted element is achieved through its co-dependence with other elements, which are overt in the encoded text.

Another aspect of cohesion is conjunction (Halliday & Hasan [1976] / 2013: 226). *Conjunction*, which is enabled through the grammatical patterns used in the text, is described in different terms in comparison to the aspects of cohesion described above, as it is not an anaphoric relation. In other words, conjunction does not constitute a process through which an element of the text refers to another one, whose presence it presupposes. Rather, conjunction encodes the relationship between two different textual elements, the presence of both of which is presupposed.

Thus, cohesion is described by Halliday and Hasan as a "tie" ([1976] / 2013: 229, 230) between textual elements, whose logical relationship is signaled in the text through the devices described above. Cohesion is seen as contributing to the overall quality of texture by capturing the interrelationship between different elements communicated through a stretch of discourse. In other words, cohesion is associated with discourse elements which contribute to coherence, which constitutes the unity of meaning of a text.

Kintsch and van Dijk (1987) discuss textual coherence in the context of investigating the underlying cognitive processes that operate in the course of text comprehension. Their analysis is similarly related to the meaning of the text as a concrete unit, but, while Halliday and Hasan ([1976] / 2013), as discussed earlier, explicate the elements present in the system of the text, emphasis is placed here on the manner in which these elements interact with processing factors with a view to accounting for the manner in which the reader achieves reconstruction of meaning from a text. A distinction is made between perception and comprehension, with the former being restricted to mere recognition of information and the latter encompassing a rather complex set of processes which entail a more active response on the part of the reader. These processes are dependent on elements of semantic organization of a text which

appear both locally, within shorter parts of the text, and globally, when considering the text as a unit. These two levels of semantic organization are both central in achieving comprehension due to processing considerations on the part of the reader, who will rely on both sources of information in reconstructing the overall meaning of the text.

More specifically, while in the process of comprehending a text, the reader is required to derive unified meaning from the distinct features present in discourse. Due to limitations in working memory and consequent constraints in the information which the reader has the capacity to retain active and process at any one time, these distinct textual elements are explicated (Kintsch & van Dijk 1987) to be revisited by the reader to variable extents, so that the reader will unify different information into a coherent whole. The process of deriving meaning from a text is therefore described by Kintsch and van Dijk (1987) as interactive, since new interpretations of textual elements might arise in the course of comprehension through this variable allocation of the reader's memory resources to different information. The reader's attention and retainment of information is also associated with a specific goal for reading that is said (Kintsch & van Dijk 1987) to delimit the comprehension process.

A prominent role in comprehension is attributed (Kintsch & van Dijk 1987) to the semantic structure of a text, on the basis of which processing and retention of textual information is achieved. The text is seen as a propositional system, with semantic relations holding among propositions. These semantic relations are either discursively encoded in an overt manner or remain inferentially available to the reader, who is understood to utilize their background knowledge, on the basis of which these inferences can be drawn. The authors discuss a distinction between the *microstructure* and *macrostructure* of discourse, associated with different levels of associations between textual elements. In particular, coherence at the *microstructure* level is

associated with unity among propositions achieved through devices such as reference, by virtue of which different propositions are structured, in the reader's interpretation of the text, as belonging to a continuous system. The *macrostructure* is described (Kintch & van Dijk 1987) as necessitating the overall propositional unity of a text to be semantically compatible with a broader *discourse topic*, in relation to which the textual propositional content can be interpreted.

Comprehension is also focused on by Kintsch (1988), through his *construction-integration model*. Comprehension is seen as the process of reconstructing the meaning of the text incrementally. In other words, comprehension is described as originating strictly at the local level within the text, for instance with word recognition, initially without considering the wider discourse context. Subsequently, it is hypothesized that, through the formation of connections between meaningful elements in the text, the reader gradually proceeds to capturing more global relations between textual features. As a result, emphasis is not placed on the reader's expectations of the propositional content of a text, but rather on the input provided through the text and its interaction with the reader's internalized representations of various notions, which are differentially activated and utilized when interpreting a text. Discourse comprehension is posited by Kintsch (1988) to be enabled through interpreting the discourse input with the aid of the background knowledge of the reader, the nature of which is described as a network of associations. Therefore, discourse comprehension is conceived of as a process of integrating the input obtained through a text into a meaningful unit by utilizing the associative capacity of the reader's internalized notional system.

Coherence is also placed at the center of comprehending a text by Graesser, McNamara and Louwerse (2002), who emphasize a need for the reader to transcend the initial process of decoding the structural units of the text and capture its overall unity

of meaning. A distinction is made between cohesion as a characteristic of the text, and coherence, which is associated with the mental representation of the text achieved by the reader (Graesser, McNamara & Louwerse 2002). The authors associate higher coherence with texts which are structurally and conceptually unified, and when links between concepts are not made overt in the text, the reader is sufficiently equipped for inferring them. In this sense, the reader's background is discussed as interacting with the structure and underlying conceptual organization of the text.

Coherence is thus associated with the degree to which the ideas attested appear to form an undisrupted unity and is achieved through discourse relations, which Graesser, McNamara and Louwerse (2004) describe as those linguistic and discourse elements present in the text which make explicit the associations among the various concepts represented in it. More specifically, the process of identifying *coherence relations* on the part of the reader is facilitated by the presence of *discourse markers*, which aid in deriving the meaning of elements of the text, in forming associations between these elements, and in integrating this acquired information into a broader system, such as the conceptual framework of the text. Such discourse markers can be pronouns, which, by referring to entities already made known to the reader, encode *text-connecting relations* and in this sense make associations between concepts in the text explicit. A similar effect is attributed by Graesser, McNamara and Louwerse (2002) to conjunctions, which can perform a signaling function of the relationship between parts of the text which have already been read and subsequent information.

The interconnectedness of concepts underlying the formation of a text is not, however, described by Graesser, McNamara and Louwerse (2002) as being restricted to lexicogrammatically encoded relationships, but it can rather be detected at the level of organization of the text as a unit. As illustrated by Graesser, McNamara and

Louwerse (2002), this is exemplified by the exact significance of a speech act being largely dependent on information available in its co-text. Such elements of coherence can be derived inferentially, as their interpretation requires the formation of associations not overtly signaled in the text.

It appears, then, that cohesion, both at the level of the formally encoded text, and regarding relations understood through inferencing, is directly related to coherence of meaning and facilitates the reader in reconstructing this conceptual representation when attempting to comprehend a text. In subsequent chapters, we will discuss the manner in which observations of such textual qualities inform our methodological approach and can be utilized in our endeavor. More specifically, the theoretical accounts discussed above will provide the basis for explicating the nature of stylometric information utilized in our experimentation and the textual information which these indices are intended to represent. Before reporting our automatic scoring experiments, we will now proceed to presenting the background of such applications, from their conception of automatic essay scoring systems until their operational use today.

3. The development of Automatic Essay Scoring

3.1 Conception of Automatic Essay Scoring

Ellis Page, having a professional background in education, draws attention to the possibility of scoring essays automatically as early as 1966. Observing limitations of multiple-choice tests in adequately reflecting the reasoning skills and critical ability of students, he emphasizes the need for essay writing to form a prominent part of educational practices and to enter standardized testing procedures more systematically. However, Page also notes a number of practical considerations which needed to be overcome in order for the education community to fully benefit from writing instruction. In particular, he expresses concern about the excessive workload into which manually grading a large number of student essays on a regular basis translates for instructors, which he considers a source of concern and complication of the instructional procedure. Page views this as necessitating the quest for alternative means to assess students' work without compromising the reliability of the process.

In addition to the need for an automated method to score student essays, Page (1966) also emphasizes the feasibility of such an endeavor. He contradicts arguments relating to a computer merely performing a task as it has been instructed to, by insisting on the capacity of an automated system to 'learn' for itself by analyzing the information that it has at its disposal and using this information to formulate a method for achieving a specified goal.

Page (1966) identifies as the biggest challenge for developing an automated essay scoring system the fact that seeking to achieve maximal imitation of human behavior would be inadequate, since the exact internal process by which human readers evaluate written texts is not understood in its entirety and is very likely to entail aspects which would not be fully replicable.

Therefore, building an automated system to undertake such a task, he argues, would necessitate a transformation of the problem. Rather than attempting to imitate human behavior per se, the goal, when designing an automated system, should be (Page 1966) to identify possible correlates of the underlying factors which may be influencing human behavior in this complex task. To this end, Page (1966) introduces the concept of *proxes* and *trins*, the latter being intrinsic variables representing a quality of interest, and the former possible approximations of these latent variables. For instance, when considering an essay submitted for evaluation, the degree of elaboration of ideas that it displays would be a variable intrinsically related to the scoring process on the part of the human assessor, whereas the length of the essay expressed in its total number of words may be a quantifiable approximation of this quality.

Based on the concept of *proxes* and *trins*, Page (1966) designed the first essay scoring model, Project Essay Grade (PEG), using regression analysis of features identified in high school essays to predict scores assigned by human raters. The variables most heavily weighted were the length of the essay and the standard deviation of word length. Importantly, when experimentally attempting to identify the scores assigned by the automated model amongst a pool of scores assigned by human raters, the system was found to be virtually indistinguishable from humans as its predictions correlated with human raters almost as strongly as they correlated with each other.

3.2 Operational use of Automatic Essay Scoring and learner corpora

The next significant step in automatic essay scoring, henceforth AES, was taken in the 1990s (Burstein, Braden-Harder, Chodorow, Hua, Kaplan, Kukich, Lu, Nolan, Rock & Wolff 1998) with the development of e-rater system by Educational Testing Service, an updated version of which is currently used, in combination with human raters, in a number of examinations. Given its early initial design and wide use, we will now discuss features of this system as an example of many related applications available today. In this context, we will subsequently present and discuss related issues which concern the examinee's profile, as well as research activities related to these issues.

In similar conceptual orientation to the first AES engine, Project Essay Grade, the e-rater system was developed to approximate human scores by quantifying a set of features in the essays and using regression analysis to determine their respective significance in predicting the essay score. This system is the first AES application introducing an element of hybridity in its scoring method (Shermis, Burstein & Bursky 2013: 10), since factors related to grammatical accuracy and organization were also included in the model in addition to word and essay length. As a result, the weight assigned through statistical analysis to the length of the composition when predicting the score was reduced, which enhanced the security of the system against essays written in bad faith potentially achieving a high score.

The e-rater scoring engine (Attali & Burstein 2006) is designed to build both generic, rubric-independent and rubric-specific AES models which would differ for essays under each topic. Models are pre-operationally trained with a large number of

essays marked by at least two human raters. Predictive features are selected and weighted using regression analysis on these training essays and then these features are applied to scoring new essays. With a view to more closely approximating human scores, the features available for selection exclude essay length and are instead informed by rating criteria used for human scoring, such as lexical choices, grammatical accuracy and discourse organization. Both conventional and unconventional patterns are identified through positively and negatively weighted features which can vary between scoring situations. For instance, certain types of unconventional structural patterns which are more frequently attested in second language writing are identified through specialized error features focusing on usage of prepositions, articles or collocations (Chodorow, Gamon & Tetreault 2010). As the researchers emphasize, detection and categorization of errors less characteristic of native speaker production, but typical of second language writing, greatly contribute to essay evaluation and to the provision of feedback through automated systems.

The manner in which the AES engine e-rater scored non-native speaker script was therefore analyzed (Burstein & Chodorow 1999) for fear of the system, which was originally designed to assess native speakers' performance, responding poorly to texts written by candidates of a different language background. Some degree of variation appeared between the performance of test-takers of a different first language, which was considered to require further inquiry into the extent to which the test-taker's language background may interact with different essay rubrics. However, the models produced in this experiment selected the same factors to be taken into consideration in scoring for each prompt, regardless of whether the majority of essays was written by native or non-native speakers, indicating adequate generalizability of the method for essays produced by speakers of other languages. Similarly, the system was found

(Burstein & Chodorow 1999) to successfully evaluate essays which represented dialectal variations in terms of grammatical and stylistic choices. Corpora representing variations in terms of dialect have been analyzed (Breland, Bonner & Kubota 1995, Bridgeman & McHale 1996) and corpora comprising non-native speaker written production have been compiled to further investigate the manner in which essay rubrics or scoring features interact with these variations and continue to make allowances for them in scoring models.

Learner corpora used for machine learning research include, for instance, TOEFL11 (Blanchard, Tetreault, Higgins, Cahill & Chodorow 2013), which was compiled from essays written in response to the Test of English as a Foreign Language college admission examination in the years 2006 and 2007. In this corpus, essays are organized on the basis of the first language of the author and the score range that it achieved, which allows investigations of score distribution in relation to the examinee's profile.

Another learner corpus compiled through examinations contains script of candidates in the Cambridge First Certificate in English test (Yannakoudakis, Briscoe & Medlock 2011), containing essays extracted from the Cambridge Learner Corpus (CLC), which has been collaboratively developed by Cambridge University Press and Cambridge Assessment. The First Certificate in English Corpus is considerably smaller but richer in information about each candidate, as it contains two essays per author on which grammatical errors have been manually annotated. Besides the scores provided in the testing situation, metadata available for the essays include the test-takers' first language and age.

Analyzing learner corpora can aid significantly in making observations regarding the linguistic behavior of speakers of English as a second language in relation to their linguistic background and inform error detection and correction and essay scoring applications.

As discussed above, an updated version of the AES engine e-rater, e-rater V.2 (Attali & Burstein 2006), is operationally used by Educational Testing Service, an examination board offering a number of tests which include language components. These examinations aim at certifying language proficiency in English mainly for college admission purposes. In some cases, two scores are provided for the constructed-response component, one of which is provided by a human assessor and the other by the automated engine e-rater, and in other settings an automated score is provided for validation purposes. In the cases of operational use of automatically predicted scores, essays are directed to an additional human rater in cases of significant disagreement between the two scores or in essays displaying very infrequent characteristics, which the system is designed to detect.

This most recent version of the e-rater AES engine, e-rater V.2, has been designed to allow for judgmental selection of the features assigned the greatest weight in the scoring process, introducing further hybridity in its design (Attali & Burstein 2006). The system also incorporates provision for detecting essays which do not respond to the topic, and efforts are made to generalize the robustness of this classification to predict responses uncharacteristic of the topic at hand without topic-specific training data (Burstein & Higgins 2005).

As described by Attali and Burstein (2006), features of a scoring model and their degree of contribution to the score may change considerably among rubrics and scoring

situations to best capture the characteristics of the scoring situation and the training data. However, this variability in models is said (Attali & Burstein 2006) to obscure the scoring process and to complicate the interpretation of the score achieved for users of the system. In this light, there is a shift of focus in e-rater V.2 towards more constant across topics and thus more readily interpretable scoring criteria. This is achieved through the relatively small feature set of e-rater V.2, which constitutes a unified basis for the design of all rubric-specific models.

3.3 Automatic evaluation of content: Latent Semantic Analysis

Apart from utilizing stylistic information, AES has also focused on analyzing the content of the texts under evaluation. This is achieved through Latent Semantic Analysis (LSA), a technique comparing quantified representations of the content of units of discourse of varying length with a view to identifying their degree of conceptual similarity (Landauer & Dumais 1997, Furnas, Deerwester et al 1988, Landauer, Foltz & Laham 1998, Foltz, Kintsch & Landauer 1998).

Latent Semantic Analysis offers an inferential solution to the learning problem of automatically detecting semantic similarity. This inferential modelling is developed in dialogue with broader accounts of human acquisition and representation of knowledge. In particular, the question is raised of the extent to which the inductive process of interpreting possible relationships between certain observations may underlie human learning, which is achieved through insufficient input information. In

relation to semantic qualities, it is hypothesized (Landauer & Dumais, 1997) that words which co-occur in discourse will bear some degree of semantic association. The strength of this association is determined by the frequency of co-occurrence, which is not only quantified on a surface level but is also utilized to induct contextual information regarding usage of lexical items in discourse (Landauer, Foltz & Laham 1998).

By identifying lexical relationships, LSA is closely related to elements of textual coherence explicated by Halliday and Hasan ([1976] / 2013, as discussed in chapter 1). In particular, Latent Semantic Analysis relies heavily on co-reference relations achieved through lexical reiteration, hypernymy and hyponymy and combines this information with observations of lexical items tending to co-occur in particular contexts (Foltz, Kintsch & Landauer 1998).

The performance of LSA systems in identification of conceptual relatedness between lexical items was tested (Landauer & Dumais 1997) and found to be comparable to human performance on questions of the Test of English as a Foreign Language (TOEFL) examination requiring the test-taker to choose the word which, among four options, is closest in meaning to a given word. In another empirical study (Landauer, Foltz & Laham 1998), a Latent Semantic Analysis model was trained on a psychology textbook and then tested through a final multiple-choice exam. Despite receiving a lower score than students taking the same exam, the model achieved a passing score, and overall committed errors similar to human test-takers.

Essay scores provided by the Intelligent Essay Assessor system, an automated essay scoring engine which uses the Latent Semantic Analysis machine learning technique, were experimentally found (Laham 1997, Landauer, Laham, Rehder &

Schreiner 1997) to correlate equally or almost equally with the correlation between scores provided by two human raters when providing a holistic score of the degree of conceptual similarity between essays. Despite not capturing information available to human readers through morphology or syntax, this alternative representation of information conveyed through a text successfully approximated human judgements of semantic relatedness (Landauer, Laham, Rehder & Schreiner 1997).

LSA can be utilized when scoring essays by identifying the degree of similarity between their content and the content of a large number of texts of acknowledged quality, such as textbook extracts, which handle the same topic (Foltz, Laham & Landauer 1999). Alternatively, as this technique also allows comparison of content between essays, it can be used for comparing essays to a model response or to all other essays in a sample. Importantly, comparisons between essays or between student essays and instructional material are possible for the full text of an essay or for fractions of it. Foltz, Laham and Landauer (1999) highlight that these affordances permit automatically detecting suspicious similarity between two essays or between essays and textbooks, thus aiding significantly in instructors' efforts to identify instances of plagiarism. Such submissions, together with highly original script and essays which fail to plausibly respond to the topic can be automatically identified through such analysis and "flagged" as special cases requiring human judgement, a provision which contributes to the reliability of the system.

The Intelligent Essay Assessor, as described by Foltz, Laham and Landauer (1999), is another established AES engine which uses Latent Semantic Analysis. Its scoring criteria do not focus on considerations of accuracy of mechanical aspects of writing, but are instead intended to capture the degree of accurate and complete awareness of the subject matter that the student displays, the extent to which their

argumentation is well-formed and the logical flow of ideas, contributing to ease of comprehension.

3.4 Automatic essay scoring from a discourse perspective

A number of applications of Natural Language Processing techniques have automatically identified discourse relations in natural language corpora. For instance, Wolf and Gibson (1985) annotated a corpus of newswire texts on the basis of discourse relations as defined by Hobbs (1985, cited in Wolf and Gibson) and used it for predictive purposes. In particular, discourse relations identified in the corpus focused on cause and effect, similarity, elaboration, evaluation, generalization, attribution and temporal sequence relations

In response to such applications, efforts have been made to incorporate discourse based information into AES systems. Yannakoudakis and Briscoe (2012), for instance, have designed a model for evaluating cohesion in learner script, which is by definition characterized by a great degree of variation. The researchers employ both more shallow information related to the distribution of different parts of speech and word length, and features more intuitively related to cohesion, such as the presence of connectives, which, as discussed earlier, make explicit the logical associations between different concepts in the text, the degree of co-reference of pronouns, as well as the degree of semantic similarity between words which appear in close proximity in the text.

Persing and Ng (2013) take into consideration in their scoring model the degree of clarity in thesis statements, using as training data essays evaluated for this trait by human readers. As a considerable extension, a method has been introduced (Stab & Gurevych 2014) for identifying elements of discourse structure in argumentative essays. In particular, the internal structure of arguments is modeled through a classification task, resulting in identification of components of arguments and the subsequent recognition of these components as reinforcing or opposing the thesis in question.

A discourse based approach to AES is employed by Persing and Ng (2015), whose scoring model is based upon strength of argumentation in the essays under evaluation. The researchers compiled a corpus of essays annotated by human readers in relation to their evaluations of the persuasive potential of the manner in which arguments present in the essays are structured and employed these annotations as training data. Features utilized in the scoring model include part of speech combinations, the degree of relevance to the essay prompt, the degree of co-reference present in the essay, as well as semantic information.

Having indicatively discussed examples of existing AES engines, we will now see possible threats to these scoring methods, originating from similar technology. In addition, we will discuss the provisions which can be made in AES systems with a view to reinforcing their reliability in detecting such threats.

3.5 Securing Automatic Essay Scoring systems against adversarial technologies

Unfortunately, as the fields of Natural Language Processing and Machine Learning advance, so do tools and applications which may enable dishonorable use of the techniques developed. This causes concern regarding possible vulnerability of automated scoring systems to strategies intended to manipulate their scoring methodology.

The possibility of automated scoring engines to be manipulated through bad-faith response strategies has provided new foci for research into intelligent scoring systems, related to the developing field of adversarial machine learning, which, as Huang, Joseph, Blaine, Rubinstein and Tygar describe (2011, cited in Cahill, Chodorow & Flor 2018) aims at refining a machine learning algorithm by securing it against software designed to exploit its areas of observed vulnerability.

In particular, Cahill, Chodorow and Flor (2018) experimentally explored the interaction of e-rater, an operationally used scoring engine and Babel, a text generation application designed to construct an essay from a small number of key words provided by the user. Observing that essays produced by the Babel application replicate certain aspects of text grammar and contain grammatically legal sentences, as well as a high proportion of low-frequency words, traits which may be evaluated positively, the researchers sought to identify the areas in which such essays significantly differ from naturally produced discourse in an empirically observable manner.

Lexical-semantic cohesion, viewed as the degree of semantic similarity of all the content words of the text, was identified as the area where automatically constructed responses most strikingly failed to emulate naturally produced discourse. In similar vein to existing provisions in the AES system experimentally used by Cahill, Chodorow and

Flor (2018) for detecting, for instance, essays not responding to the topic, integration in the scoring engine of a classification algorithm resulted in successfully identifying essays generated through the Babel tool without erroneously extending this observation to good-faith responses. The area which presented the greatest challenge was found to be reliably discriminating between very incoherent, but humanly produced essays from machine-generated responses.

It appears then that classifying machine-produced essays by detecting their systematic differences from human writing is a feasible machine learning task. The researchers additionally highlight the importance of further investigating semantic violations in machine-generated responses, including semantic violations at the sentence level, so that anomalous fractions of text can also be detected in real examination situations.

The fact that the rapidly developing field of Artificial Intelligence allows a proportion of human decisions to be replicated only underscores the value of the remaining human behavior, which would include moral considerations of whether or not one *should* perform an action that they have the capacity to perform. It may be argued that enhanced technological assistance in fulfilling one's goals necessitates a more acute sense of responsibility as to the nature of the goals prioritized and pursued. Consequently, cultivating values and attitudes that promote moral integrity and social responsibility, already a distinct part of pedagogical objectives in many educational contexts, appears an increasingly essential curricular goal.

Simultaneously, however, exercising rigor in analyzing the products of 'gaming strategies' described above and continuing to reinforce AES systems against them remains of the essence for ensuring validity of AES applications.

In the following chapter, we will approach the promising aspects, as well as limitations to AES from a social and education perspective, and we will discuss possible and observed pedagogical effects of AES.

4. Discussion of the relevance of Automatic Essay Scoring

4.1 Implementation of AES systems: discussion of benefits and limitations

Automated essay evaluation has often been approached with skepticism related to the fact that scoring engines do not process the text to be scored in the same way as humans do. Page and Petersen (1995) broadly categorize objections to AES into humanistic, defensive and construct criticisms.

Humanistic objections are related (Page & Petersen 1995) to a lack of a human audience for the written product, which constitutes a remarkable difference from humanly evaluated essays, since AES engines do not appreciate the potential communicative effect of the essay they are processing. Defensive criticisms, as discussed by Page and Petersen (1995) are related to an increased risk for AES systems, when compared to human raters, to be 'deceived' by bad-faith strategies carefully tailored onto their scoring methods, the detection of which would at times require human judgement.

Finally, Page and Petersen (1995) explicate construct objections as related to the difference between the scoring criteria employed by human raters and automated engines. This difference is captured in the terms *trin* and *proxy*, from the terms intrinsic and approximation, which, as described above (paragraph 3.1), represent the latent criteria hypothesized to affect human scoring decisions and the quantifiable features which are expected to most closely approximate these criteria, which are thus taken into consideration in the design of AES empirical studies. The construct validity of AES

systems would, as a result, be expressed in the degree to which the scores predicted by automated systems successfully imitate human scoring decisions.

Doubts have, however, been expressed (Bennett & Bejar 1998) regarding the degree to which comparisons to human scoring behavior and efforts to approximate it as closely as possible would constitute the best possible criterion for evaluating the validity of automatically produced scores. The researchers argue that if similarity to scores provided by human raters is employed as the only validation criterion, then biases and limitations inherent in human scoring will also be inbuilt in automatic scoring engines. In response to this concern, the researchers propose analyzing the features intended to be used for scoring and their interaction with test-takers' performance on each task as an additional measure which would provide substantial information regarding the performance of scoring models.

Automated evaluation of essays has also received criticism from an interpersonal perspective, related to the lack of involvement of a human audience for the essays being constructed. Objections have been raised by writing professionals of the Conference on College Composition and Communication organization (CCCC), (2004, as cited in Deane 2013) regarding the risk of the essentially social character of writing being absent from instruction or examination situations where AES engines are used. This aspect was considered hazardous for the writing process on the part of learners, as it was suggested that employment of AES systems might affect learners' conceptualization of their intended reader in unintended ways, and this might result in a lack of appreciation of writing as a social process.

Concern has additionally been expressed (Rothermel 2006: 200) about the application of AES systems introducing an impersonal atmosphere in learning contexts

and imposing restrictions on the teaching and learning environment by over-emphasizing a particular set of skills at the cost of more diverse learning experiences which would equip learners better for coping with the demands of the modern world. Rothermel (2006: 208) highlights the need to ensure clarity in evaluations of AES implementation by exploring learners' and teachers' reactions to AES systems, which she found to report an undesirable degree of uniformity in written production being associated with AES systems (Rothermel 2006: 209). Studies gauging the effects of introducing AES systems in educational contexts both through measures of participants' performance, and through exploration of their attitudes and views have, however, not always indicated a need for concern. Combined with some degree of mistrust towards evaluations provided by computer programs, particularly when these evaluations were negative, interaction with AES systems has frequently been positively received by participants in learning contexts.

Herrington and Stanley (2012) discuss potential limitations of AES systems in terms of possibly failing to recognize dialectal variations in the script evaluated, thus implicitly introducing a differential status for the standard dialect. This aspect of sociolinguistic variation has not, however, escaped the attention of AES research, as efforts are made to identify the manner in which essay writing and scoring interacts with dialectal features. To this end, stylistic preferences of Asian American, African American and Hispanic/Latino ethnic groups in essay corpora have been analyzed, also taking gender into consideration (Breland, Bonner & Kubota 1995). Factors of ethnic group and gender have also been investigated in relation to performance on various items of college admission tests, including essays where a combination of automated and human scoring is involved (Bridgeman & Mc Hale 1996, Breland, Kubota,

Nickerson, Trapani & Walker 2014). These analyses are taken into consideration to promote validity of test results.

Strong opposition of AES is voiced by Perelman (2012, 2013), who expresses serious doubts regarding the possibility for automated systems to capture the complexities of the writing process. In particular, the author criticizes the implementation of scoring engines in which essay length can have a significant impact on the score provided (Perelman 2012: 123,124). Perelman's objections to AES systems focus on differences in the definition of the writing construct when considering human and machine evaluations. Although it is true that an automated system by no means reads and appreciates an essay in the same way as a human, the distinction between factors affecting human and machine decisions remains the only feasible manner to construct an automated model to imitate human behavior. Rigorous statistical testing of the performance of AES models is meant to evaluate the degree of approximation achieved between these two distinct procedures.

Perelman (2012: 126) further dismisses the manner in which stylistic features are measured as excessively abstract and remains skeptical as to the informative value of employing numeric indices in evaluating the quality of written work. It can be argued, however, that quantified expressions of qualities of discourse need not be dismissed as irrelevant to the structure of discourse phenomena. The informative value of word length, measured in letters, for instance, is doubted by Perelman (2012: 126). Although not immediately recognizable as a feature of language usage, this index is in fact based on Zipf's law, which is expressed in the inverse relationship between the length of words and their frequency in large natural language corpora, capturing the least effort principle in language usage (as described, for instance, in Wyllys, 1981).

Word frequency, another numeric factor heavily relied upon in AES systems, as is the case, for instance, in the e-rater system, has also been empirically found to correlate with the degree of difficulty that this word appears to present for speakers. Breland (1996), for instance, notes that, with a noteworthy degree of consistency, word frequencies as obtained through different corpora were found to strongly correlate with test-takers' performance in items testing less frequent words. It appears, then, that quantified observations with regard to the language used in written production can represent very significant, linguistically relevant qualities.

Perelman (2014) expresses further reservations in relation to the validity of the agreement rates with human scores obtained through AES studies. He argues that essay length is systematically overemphasized in AES systems, leading to inflated similarities between scores and complains about developers of AES engines not always disclosing details about their system. However, this lack of code disclosure appears understandable given the proprietary nature of the systems developed.

Another complaint expressed by Perelman (2014) is related to essay length being crucial in human judgements of essays in impromptu writing tasks which are carried out within a time limit, on the premise that, when students are constructing a response to a topic they have not prepared beforehand under pressure of time, the length of their response may be indicative, in combination with other factors, of ease of written production. In general, since the conception of AES systems by Ellis Page in the 1960s, there have been systematic efforts on the part of developers of AES systems, as described above (chapter 2), for reduction of the weight of essay length in their predictive models. The fact remains that the behavior of stylistic indices is to some extent influenced by essay length, but this by no means renders essay length the only source of information taken into consideration in automated scoring systems. The

challenge of fully disentangling calculations of stylometric indices from essay length only indicates the complex nature of discourse, which, in the case of evaluating impromptu timed written tasks, is also evident in human behavior.

Despite some degree of reservation, automated scoring systems have frequently been received more positively, emphasizing their potential to support the learning process. McAllister and White (2006 :27), for instance, encourage awareness of the potential of computer-assisted writing evaluation to aid the learning or evaluation process and contribute to the improvement of practice in certain instructional contexts. The author cautions against the possibility of AES technology, if shunned by educators, to be used in manners deprived of their pedagogical expertise, and thus to be rendered counter-productive. Brent and Townsend (2006: 197,198), observing the implementation of automated evaluation in a Sociology college course, also draw attention to the need for judicious use of AES systems so that they can best support learning through a combination of their affordances with instructors' wider-encompassing role.

Deane (2013) advocates abandoning extreme positions related both to uncritically implementing AES in all writing situations and to categorically rejecting potentially fruitful applications of computer-assisted essay evaluation. The author emphasizes the importance of occupying middle ground as to the manner and degree of reliance on AES, with a view to judgmentally reap its benefits while incorporating it constructively in instructional contexts by taking into consideration the varying user or assessment situation needs.

Similarly, Brent and Townsend (2006: 179, 180) have noted a need to adapt scoring methods depending on the evaluation setting. They suggest that essay rubrics

testing language ability which have been designed to be very general in scope and aid test-takers from a wide variety of backgrounds to respond to them may be adequately scored through models capturing more mechanical aspects of writing, such as accurate language usage and punctuation. By contrast, the researchers argue that essay rubrics more closely related to specific academic disciplines, which would appear in the context of higher education, would require considerations of semantic aspects of the written work.

Deane (2013) additionally highlights that objections to AES very closely reflect objections to standardized testing in general, rather than being hesitations about the nature of the technology developed per se. The author proposes that automated evaluation lends itself and can be configured to support a variety of uses, so that its applications can be tailored to various instructional needs. Additionally, Deane (2013) argues that continued research efforts into AES systems can yield models which would capture more factors at work in the writing process, such as contextual information as this becomes evident in discourse, or provide scores based on distinct scoring traits as an alternative to holistic scores.

The social aspect of AES is discussed by Shermis (2014) in response to changes in curricular objectives in the United States. Emphasis is intended for essay writing with a view to better preparing the student population for college entry (Tucker 2009, as cited in Shermis 2014). As a result, Shermis (2014) anticipates a need for increasing numbers of long-form constructed responses in standardized examinations within the high school context in response to these education policy changes. This need underscores the potential of AES for supporting the education system and for increasing writing practice opportunities for a large number of students.

Shermis (2014) describes two events organized by the William and Flora Hewlett Foundation which demonstrated the potential of AES applications. In a national event in the United States, AES vendors provided scoring models whose performance was compared with scores provided by human raters and successfully replicated the mean scores for all essay sets provided. The second, international, event was the public competition hosted by the Kaggle platform, from which data were obtained for the present study. In this competition, 159 research teams submitted models for evaluation, which is indicative of the interest that the competition generated in the Natural Language Processing Community. The performance of the three winning teams successfully replicated human scores, further reinforcing confidence in the potential of AES systems to aid education contexts.

As Davidson (2018) observes, many current educational demands could be met through implementation of AES. In particular, the author associates increasing interest in AES with emphasis on comparable standards in educational contexts, as well as with the increasing popularity of digital learning environments, such as Massive Open Online Courses (MOOCs), which require assessment of thousands of written responses. Davidson argues that AES may reinforce the development of learners' writing skills due to the immediate provision of feedback and the possibility for more individualized practice, alleviated from the risk of embarrassment in case of poor performance. Simultaneously, the author cautions against the possibility for more creative aspects of writing lessons to be neglected and for writing tasks to be approached in bad faith when AES is used. Highlighting the significance of the teacher's role in orchestrating writing development, Davidson draws attention to versatility in AES applications, which he evaluates as promising for a variety of purposes, including practice and diagnostic testing.

Massive Open Online Courses, where thousands of participants may simultaneously be enrolled in a course, can be confidently said to directly reflect the needs of the current educational landscape. In light of the increasing need of today's student population for lifelong learning and the not uncommon need for re-training, educational opportunities which allow access to specialized knowledge globally appear to have undeniable social significance. Therefore, seeking the optimal method for assessment of participants' work in digital contexts is necessary if maximal positive impact of these affordances is to be achieved. Given the often staggering student instructor ratio in MOOCs, another important social dimension of AES emerges.

In efforts to identify the benefits of different scoring methods in the context of MOOCs, Balfour (2013) compared automatically generated feedback to peer evaluation of responses in the same digital environment. Different benefits were associated with these assessment methods, with AES appearing more appropriate for tasks requiring literal use of language, where its benefits of immediacy and consistency could be best observed. Peer assessment, on the other hand, was felt to be a more complex process, although better suited to tasks which may involve figurative use of language. Feedback obtained from multiple peers based on instructor guidelines appeared to be less consistent, which seemed to be compounded by individual differences related to efficacy for reviewing and readiness to support work produced by peers. The contribution of peer assessment, however, to the development of participants' critical ability was not overlooked. Overall, it was concluded that these different scoring methods can each yield significant results depending on the task type and the stage od development of an activity.

In a similar direction, Suen (2004) compares possible ways in which the challenge of evaluating very large numbers of student responses can be met in MOOCs.

Reporting a need to inform and plan peer assessment in the best possible way to ensure its effectiveness, he also discusses AES as an important resource in tasks whose focus is language-oriented, associating this more closely with the nature of the analyses undertaken to inform AES systems. However, in light of additional developments in automated engines, which, as discussed earlier, now often include content-based techniques, evaluation of course work with a greater emphasis on the content of responses now appears more realistic.

Reilly, Stafford, Williams (2014), investigating the effectiveness of AES systems in two digital courses report automatically provided evaluations not yet fully reflecting scores provided by instructors. Taking into consideration the social benefits of MOOCs, continued efforts into exploring the potential of AES systems in such contexts remains essential.

In addition to the reservations regarding AES discussed above, as well as its promise and potential applications, we will now discuss the possibility of applying AES technologies in broader areas of instructional activity, in addition to examination situations.

4.2 Automatic Essay Evaluation: providing feedback for written work

More process-oriented applications utilizing AES tools have become available, which emphasize the pedagogical potential of Natural Language Processing technologies in the context of writing instruction. Macdonald et al (1982) observe a lack of progress in

the written work produced in the context of writing lessons despite availability of feedback from their instructor. They attribute this lack of responsiveness to feedback to challenges faced by learners in remembering the feedback which they have received, as well as in making active use of it in relation to their own work. As a result, the authors emphasize the need to support learners in advancing the quality of their writing and propose that providing feedback automatically could contribute to overcoming these obstacles.

Describing the Writer's Workbench applications for providing feedback on essays, they emphasize their capacity to provide suggestions for improvement in writing in areas other than the most widely encountered spelling corrections. In particular, the editing applications described include recommendations to writers by recognizing specific types of grammatical violations and providing more accurate versions. Additionally, the systems described detect poor stylistic choices reflected in word combinations which have been annotated as problematic in training data and recommend more appropriate alternatives, which the individual user has the opportunity to tailor to their personal writing needs by delimiting the lexicon upon which recommendations are made.

The Criterion Online Essay Evaluation Service, as described by Burtsein, Chodorow and Leacock (2004), represents another significant development in the field of automatically provided feedback, designed to support and extend the practice undertaken in the classroom. The system incorporates an AES engine, e-rater V.2, as well as the Critique suite, which generates feedback on a variety of areas including language usage, text organization and style.

In particular, grammatical violations are recognized by virtue of statistical information gained through a large training corpus, based on which observed frequencies of various combinations in learner script are compared to their expected frequencies, so that divergent patterns become apparent. Provision is also made for detecting errors with homophones, in response to frequent errors committed with usage of such words. Stylistic inadequacies detected include excessive repetition of words or overly long or very short sentences and potentially problematic overuse of the passive voice.

Discourse structure elements are also brought to the attention of the user, which is made possible through a large corpus annotated by human experts with categories such as thesis statements or supporting statements. Recommendations are made to learners in view of the highly structured nature, form a discourse perspective, of persuasive essays (described, for instance, in Hyland, 1990) It appears that the descriptive feedback provided through such systems can aid both learners and teachers greatly by reinforcing, through simplifying the process of individualized suggestions, concepts related to quality of writing which have been focused on in class and enhance their realization in learner writing.

As Ware (2011) describes, in such Automatic Essay Evaluation applications, the focus is on providing feedback on learners' work, rather than scoring their essay as an end product, which renders related technology suitable not only for summative, but also for formative assessment. As Ware describes, applications geared towards formative essay assessment provide learners and teachers with a wealth of information regarding the written work produced, including assessments expressed numerically or represented graphically and specific recommendations in relation to features of the submitted text.

Automatically saving this information is observed by Ware (2011) to facilitate conducting analyses of learner or class performance as the means for monitoring progress become readily available both for the teacher and learners. It appears that this aspect of essay evaluation tools can contribute to learner autonomy, as the user is afforded the opportunity to repeatedly edit their work in response to feedback received and submit it again for evaluation. Simultaneously, the teacher can use systems as an aid in supporting the development of learners' writing skills, but also judgmentally intervene with additional feedback, monitor learner behavior through measures such as time spent on each task and communicate with learners in an asynchronous manner.

Informed by teachers' and learners' reactions to implementation of essay evaluation tools in education contexts, Ware (2011) highlights the importance of teachers utilizing the enhanced flexibility of such systems in a discretionary manner. Critically exploring the potential of different features of these systems in relation to different learning situations and aims is encouraged (Ware 2011) with a view to selecting or adapting the most relevant tools and combining them effectively with other classroom practices.

One example of the diverse information that can be provided through automated tools is discussed by Ehsan and Faili (2012), who describe a proofreading tool based on a machine learning method used for automatic translation. Suggestions provided through this tool include spelling, grammar, as well as lexical choices. Tools of this kind can aid learners in editing their own work and can help them become more aware of areas that require their attention through interactively reviewing errors and incorporating suggestions during the course of writing.

Classroom applications of AES are also discussed by Burstein, Marcu and Knight (2003). The online essay writing practice tool Criterion is discussed, which uses machine learning algorithms in order to model teacher behavior in response to writing tasks. In particular, the authors describe a discourse analysis software embedded in Criterion, which has been designed to inform learners of any violations of discourse structure which appear in their essays in relation to the text structure which is expected of persuasive texts. This is achieved through labeling sentences in learner script under specific discourse elements, for which process the software has been trained through data annotated by human readers.

Developers of this system have used a linear, as opposed to hierarchical, representation of the text, the consecutive parts of which are associated with a particular communicative goal, and bear labels relevant to the discourse structure of essays, such as “thesis statement”, “introductory idea”, “main idea”, “supporting idea” or “concluding statement”. The authors emphasize pedagogical uses of AES as the feedback provided is immediate and personalized, so that learners can be continuously supported in their writing development. Given the capacity of automated systems available for practice purposes to capture not only mechanical aspects of writing, but also more global features of its discourse organization, it appears that their potential for providing assistance to learners and teachers cannot be overlooked.

4.3 Observed effects of AES on learner performance

In parallel with researchers' intense efforts at optimizing AES systems, attention has also been drawn to the manner in which their implementation impacts learner performance and learner and teacher attitudes within a number of instructional contexts.

Burstein, Chodorow and Leacock (2004) report improvement in class performance, after being exposed to the Criterion interactive system for AES and automatic feedback generation. The researchers received positive feedback both from learners and teachers who used the tool in parallel with their classroom-based instructional activities. It appears that the capacity of the system for developing and electronic portfolio, while compatible with the requirements of ongoing assessment, can simultaneously facilitate the editing process of learner script and thus contribute to learner motivation to pursue writing tasks.

Lipnevich and Smith (2009) focused on feedback provided through Criterion, an automated tool developed by Educational Testing Service and ESL Assistant, developed by Microsoft Research. These applications use machine learning techniques to identify patterns of conventional article and preposition usage, so that they can detect divergent instances in second language writing and provide more accurate alternatives to the writer. In their study, Lipnevich and Smith (2009) observed the writing performance of a group of undergraduate students in the United States, native and non-native speakers, as a function of automatically provided feedback in the areas of article and preposition usage. By contrast to native speakers, whose performance when revising their essays was not found to be affected by automatically receiving feedback, non-native speakers showed greater improvement when receiving such feedback.

Importantly, when automatically provided with suggestions for improvement in article and preposition usage, non-native speaker users were found to favor more valid recommendations, indicating a tendency to use the system constructively in terms of language learning.

Chodorow, Gamon and Tetreault (2010) discuss the importance of using automated error correction systems on the written production of non-native speakers of English. The authors emphasize the need for supporting writing performance in a second language from a social perspective, observing the sheer number of non-native speakers of English residing or following studies in the United States, as well as the ever-increasing population of learners of English as an additional language and of non-native textual production in English. Valuable information which can be constructively used in this direction arises through analyzing learner corpora (discussed in paragraph 3.2), through which patterns of language usage which present challenges for learners of English as an additional language in a different manner from native speakers can be observed.

Xiaoyu (2018) examined the effects of experimentally exposing a senior high school English as a Foreign Langauge Writing class to usage of an AES system. Alteration of learners' metacognitive behaviour was explored through self-reported employment, on the part of the learners, of various metacognitive strategies related to writing, including, for instance, self-monitoring of lexicogrammatical choices, reviewing the written work produced and setting goals for improvement. Results indicated a significant increase in the usage of such metacognitive strategies. Additionally, the researcher examined whether these metacognitive effects were accompanied by changes in learners' performance in written work. Analyses of writing test results, as well as evaluations provided by the AES system used indicated progress in writing achievement. It was

concluded that usage of the AES system contributed to enhanced attention on the part of learners to the process of producing written work and supported their progress through provision of immediate feedback in areas such as spelling and grammar. It was also highlighted, however, that the teachers' role remained essential, especially in areas such as reviewing the content or coherence of essays and in providing positive feedback in order to reinforce particularly successful aspects of learner writing.

Similar benefits were observed through implementation of Intelligent Essay Assessor system in a Psycholinguistics course in New Mexico State University. As reported by Foltz, Laham and Landauer (1999), students were encouraged to submit to the system essays written in the context of the course and then freely revise and re-submit them while monitoring changes in their performance. Designed to analyze conceptual similarity between texts, in this case essays written by students and textbooks used in the course, the Intelligent Essay Assessor system alerted students to aspects of content which would have been expected under the rubric given but were overlooked in their responses. Students' scores improved considerably after this interaction with the system and a subsequent survey found a striking majority to be positively disposed to hypothetically using such a tool in other courses.

Having initiated our discussion with theoretical accounts of text organization and coherence, which we expect to arise in evaluations of learner script by expert human raters, we then proceeded to compare different viewpoints regarding the potential of AES for education purposes, as well as observations of the effects of current applications of automated systems in instructional contexts. We will now turn our attention to our own attempts at building an automated model predictive of essay scores, but most importantly at exploring, through experimentation with constructing such a model, the interrelationship of quantifiable features of texts, in this case learner script,

and theoretical notions explicating textual qualities and thus underlying human judgements of the level of writing achievement on the part of learners. In the following chapters, our experimental design will be detailed in relation to the theoretical accounts which inform our analysis. Information relating to the source and nature of our data will be presented and our incentives for opting for analyzing the corpus in question will be discussed. Subsequently, we will present and discuss the results of our experiments and attempt to interpret our results in relation to the theoretical accounts of textual qualities discussed above.

5. Experimental design

In light of theoretical accounts of text organization (discussed in chapter 2) and in an attempt to explore the relationship between the behavior of stylometric indices in learner writing and their possible relationship with variable degrees of writing achievement in learner script, we analyzed a corpus of learner essays evaluated by human experts, taking into consideration the scores provided by human raters as indicative of the degree of sophistication that learner writing displayed. In this chapter, we will provide information about the corpus analyzed, as well as the nature of stylometric indices which we took into consideration in our analysis. We will subsequently report two experiments conducted with a view to exploring the degree to which attempting to predict features indicative of discourse qualities with the aid of stylometric information will yield a significant relationship between these two types of information. Additionally, through our experiments, we will investigate which specific stylometric indices appear more closely related to human reactions to the texts analyzed.

5.1 Research questions

In our analysis, we aim to model the predictive power of a combination of different stylometric information in relation to the score assigned to essays by expert raters, which we consider indicative of discourse quality. Given the complex and multi-faceted nature of discourse, as well as variability of performance in learner script, we initially take into consideration a large set of quantifiable features, the content of which will be

detailed below, many of which represent aspects of cohesion, as have been discussed earlier.

Co-reference relations and the presence of conjunctions, for instance, constitute indications of coherence represented by a number of indices utilized in our experiment. We also expect coherence of the content of the essays under investigation to be captured by measures employing the Latent Semantic Analysis technique, through which the degree of variability of content among different segments of the text is quantified. The presence of various grammatical patterns in the texts is additionally explored, a number of which are seen as contributing to cohesion of the text and others as being informative of the degree of syntactic complexity of the sentences present in learner responses. Finally, the frequency structure of the texts is analyzed through a number of indices indicative of the degree of vocabulary richness attested in the texts, which we also expect to influence raters' reaction to texts since it would represent the degree of sophistication of the essays.

In this light, we undertake two experiments, with a view to responding to the following research questions:

1. Can essay score be significantly predicted based on a model of the stylometric indices described above?
2. Which stylometric indices most strongly predict the score assigned to essays by human experts?

We view the first research question as informative of the degree to which a significant relationship will indeed be observed between features of texts which we hypothesize to have affected scores awarded by human raters, and which constitute theoretically grounded features of discourse quality intrinsically related to

comprehension and appreciation of written texts by humans, and quantifiable features which we intend to represent machine-readable approximations of these latent variables which may be utilized by an automated system.

Through the second research question, we aim to obtain more specific information regarding which specific stylometric indices most strongly correlate with human assessments, through which we will attempt to interpret the conceptual grounding underlying points of strong convergence between our two different representations of textual qualities, those intrinsically employed by human readers and those quantified representations accessible to automated systems.

5.2 The corpus under analysis

The dataset analyzed was obtained online from Kaggle (https://www.kaggle.com/), a data science platform which hosts a number of machine learning competitions. Through this platform, the William and Flora Hewlett foundation (Hewlett) sponsored the Automated Student Assessment Prize (ASAP) competition in 2012, with a view to promoting and supporting research efforts to utilize machine learning technologies in response to the increasingly socially relevant issue of evaluating a very large number of responses to open-ended examination tasks. In particular, organizers of the competition make reference to increasing attention to essay writing in standardized testing by State Departments of Education in the United States in efforts to enrich examinations with tasks which better reflect analytic reasoning and critical ability on the part of the learner. The overall aim of the competition, therefore, in which 154

research teams participated, was for the machine learning community to concentrate their efforts on designing a viable, cost-effective solution to facilitate wider implementation of essay writing in curricular design.

Participants were provided with large essay datasets rated by human experts and asked to submit automatic scoring models whose performance would be evaluated on the basis of the strength of their predictions in comparison to additional data bearing undisclosed scores.

More specifically, essays in the Kaggle corpus are organized in eight different data sets based on the rubric to which they respond. In some data sets, essays are produced in response to the rubric only, whereas in others, students are provided with input material in addition to the rubric. Scores provided by human experts follow a holistic scoring method for seven of the data sets, and there is an additional set of essays where scores have been provided on the basis of distinct traits for evaluation. For each essay, scores are provided by two different expert raters.

These variations in the characteristics of essays and scoring methods enables a more detailed analysis of the strengths and weaknesses of each scoring model. Simultaneously, however, it introduces a degree of complexity to the machine learning task of designing a unified model to effectively predict scores across all data sets which is outside the scope of the present study. For the purposes of our attempt, we have taken into consideration in our analysis the essays of one of the data sets provided in the context of the machine learning competition.

We have analyzed a data set of student script which consists of 1,785 persuasive essays of 350 words on average. The essays have been written by students at Grade level 8 who have been asked to express their viewpoints and present relevant

argumentation in relation to the effects of the increasing use of computers. The topic and the scoring criteria used by the expert raters can be found in the appendix.

Expert raters have been provided with descriptions of six score levels which take into consideration clarity of ideas produced, organization and coherence, and communicative effect. Raters have been asked not to focus on usage or mechanics errors and it has been requested that evaluations take into consideration what has been produced irrespective of how complete the written work appears. As the essays have been produced as spontaneous responses within a time limit, these are considered elements which would have been edited in a later version of the texts. This is related to the nature of the Kaggle corpus, which contains essays produced by native speakers of English. This element thus makes the Kaggle corpus suitable for the goals of the present study, which does not focus on factors pertaining to non-native speaker production but investigates how stylometric indices behave in essays of varying quality, as encoded in the score provided by expert raters.

Following standard practices in relation to data protection and taking into consideration the students' young age, who are between grades 7 and 10, great care has been taken in the context of the competition to protect the anonymity of the students providing script. Therefore, seeking to identify the authors of essays is explicitly prohibited through the contest regulations as they appear on the Kaggle platform. As an additional measure, essays contained in the publicly available Kaggle datasets have been subjected to pre-processing for automatic anonymization purposes. Named entities as well as other information which might be informative of the authors' identity, such as numbers, have been removed and replaced with special tags indicating the type of information which has been omitted. While absolutely essential, this process introduces a slight limitation to the data under analysis, since it constitutes a slight

change in the original content of the essays. However, public access to this wealth of naturally produced script by authors of the desired profile would otherwise have been impossible. Additionally, these small omissions concern all the essays in the corpus, and their content does not appear closely related to the subject matter at hand, the benefits and threats of using computers. As a result, it does not appear that this anonymization process importantly affects our analysis.

Another limitation of our data is the fact that the texts in our corpus are not represented in paragraphs. Therefore, an analysis of textual features at the paragraph level was not possible. However, by obtaining detailed information about different aspects of the texts, which will be elaborated below, we have at our disposal sufficient information to detect the overall cohesion of the essays, as well as the homogeneity of their content.

5.3 Data analysis through the Quita tool

We used the analysis tool Quantitative Index Text Analyzer (QUITA) to obtain a number of indices useful for our analysis. QUITA is a program designed for the quantitative analysis of texts, mainly focusing on indicators which capture the frequency structure of texts through diverse measures but also including indicators of other features, as will be discussed below. The majority of these include measures of lexical diversity, which we expect to be a strong predictor of achievement in an essay task, both form a discourse and from a developmental perspective, since we expect the diversity of lexis used in the texts to be indicative of their richness of content, which

will be central in capturing and maintaining the reader's interest. Through the QUITA tool, we have analyzed our essay data with 22 indices, the content of which is as follows.

The Type/Token Ratio (*TTR*), capturing variety of words in a text, is the ratio of the number of types to the number of tokens in the text, with types representing the number of unique words that the text comprises irrespective of their frequency in the text, and tokens representing the total number of words which the text consists of. Wimmer and Altmann (1999), discussing the empirical identification of vocabulary richness in a text, draw attention to the complexity of the problem, which is compounded by a relative lack of stability in the behavior of vocabulary richness indices and by their interrelationship with other features of the text. Ultimately, the researchers associate the complexity of quantifying the phenomenon in question with the complex nature of language usage itself and highlight that seeking to probe the frequency structure of texts through a variety of indices is crucial since, if different indices were not informative, they would appear equated. Despite clarity in its conceptual foundation, the Type/Token ratio has been observed (Wimmer & Altmann 1999, Kubát & Milička 2013) to be greatly influenced by text length. To overcome this issue, Kubát and Milička (2013) introduce a method related to the Type/Token ratio for capturing vocabulary diversity while avoiding dependence on the length of the text and discuss possible interaction of their proposed method with genre and authorship variables.

The *h*-point is conceptually related to Zipf's law, (analyzed, for instance, in Wyllys 1981), which models the frequency of an observation in relation to its rank among all other observations in a population. This distribution of frequency in relation to rank displays regularity when observing natural language, as the frequency of words

appears to be inversely proportional to their rank in large natural language corpora. The *h*-point index is the point in the rank-frequency distribution of the words in a text at which the frequency value equals, or neighbors, the rank. This point separates the words in the text into *autosemantics*, which introduce the ideas of the text, and *synsemantics*, which establish and explicate the relationships between them (Popescu, Mačutek & Altmann 2009b: 24, 25). Thus, this index is directly related to the thematic organization of the text. The *R*1 is another index of vocabulary richness which is based on the *h*-point but is less sensitive to the length of the text.

Repeat Rate (*RR*) and Relative Repeat Rate of McIntosh (*RRmc*) represent lexical diversity in the essay, whereas Hapax Legomenon Percentage (*HL*) is the ratio between the words which occur only once in the text and the total number of words in the text. Through this relationship, it indicates the degree of variety that the words in the text display.

The Lambda (Λ) indicator expresses the frequency structure of a text by incorporating the diversity of words in the text as well as taking into consideration the Euclidean distance of the frequencies between words which are nearest in the frequency distribution of words in the text (Popescu, Čech & Altmann 2011: 3). In this way, the Lambda indicator expresses vocabulary richness through incorporating the additional measure of the degree to which the frequency of each word exceeds that of the word subsequent in rank. This measure captures the degree of uniformity which the frequencies of the words which appear in the text display (Popescu, Čech & Altmann 2011: 66). Gini Coefficient (*G*) is another measure of vocabulary richness, based on statistical distribution. Lexical diversity is also measured by the *R4 index***,** whose calculation is derived from the Gini Coefficient index.

Curve Length (*L*) represents the length of the curve of the frequency-rank distribution and is the sum of the Euclidean distances between all points on the curve, whereas Curve Length R Indicator (*R*) is an index derived from Curve Length. It compares the part of the frequency-rank curve which exceeds the *h*-point, described above, to the total length of the curve. Entropy (*H*) is a measure of vocabulary richness. It is inversely related to concentration of the vocabulary in a text, as greater values of Entropy represent lower predictability, in a probabilistic sense, of the words in the text based on the relative frequency of the rest of the words that the text contains.

The Adjusted Modulus (*A*) additionally indicates lexical richness based on the frequency structure of the text. Calculation of the Modulus index is based on the frequency value of the most frequent words in the text, the vocabulary size of the text, indicated by the value of the greatest rank, and the *h*-point, described above. Through adjusting the value of the Modulus index in relation to the length of the text, an effort is made to reduce its sensitivity to text length (Popescu, Mačutek, Kelih, Čech, Best & Altmann 2010: 4, 5).

Various other textual features are identified through the QUITA analysis tool. The Verb Distances (*VD*) index represents the mean number of tokens which appear between two verbs, annotated by a part of speech (POS) tagger. Activity (*Q*) and Descriptivity (*D*) are related to verbs and adjectives in the text. The Writer's View index (*a*) is based on text length, the frequency-rank curve and its *h*-point and seems to represent a 'golden section' where the author exercises control over the developing trajectory of word frequencies in the text (Tuzzi, Popescu & Altmann 2010). Average Tokens Length (*ATL*) represents the mean of the length of all words attested in the text. Thematic Concentration (*TC*) captures the degree to which the author focuses on a specific topic in the text, as indicated by the presence of autosemantic words. Secondary

Thematic Concentration (*STC*) is directly derived from the Thematic Concentration index, although it employs the *h*-point described above in a different manner in its calculation.

In efforts to analyze the texts under scrutiny more comprehensively, attempting to enhance our approximation of the multi-dimensional nature of discourse, we included in our analysis a combination of the indices described above and features of the texts that we obtained through another text analysis tool, Coh-Metrix, which focuses on a different set of measures.

5.4 Data analysis through the Coh-Metrix tool

In addition to the indices obtained through the QUITA analysis tool, described above, we analyzed the essays in our data set with Coh-Metrix, a tool primarily focusing on indices of text cohesion, which, as explained by Graesser, McNamara and Kulikowich (2011), was the original incentive for its development. In fact, the very design of this tool is characterized by sound theoretical grounding (Graesser, McNamara & Louwerse 2004) in terms of features of discourse organization contributing to cohesion, and such as co-reference (as discussed in chapter 2), with subsequent effects on the process of text comprehension.

Despite readability levels of the texts not being the focus of our study, we considered text cohesion, with its great significance for comprehension processes (Halliday & Hasan [1976] / 2013: 8,9, Kintsch & van Dijk 1987), essential in our

investigation of the interaction between text features and human evaluation of text quality, since we expect cohesion indices, with their effect on readability, to additionally reflect clarity of argumentation and sophistication of discourse structure organization.

In addition to coherence, which was the original incentive for the designers of the tool, more diverse information was added to the design of the Coh-Metrix tool very soon after its conception. Reflecting increased attention to multiple levels of linguistic analysis (Kintsch 1988, Graesser, McNamara & Kulikowich 2011), this information is captured by indices including semantic representations of information, measures of syntactic complexity and of ease of text comprehension, resulting in a total of 108 indices.

The relevance of obtaining the stylometric information offered through the Coh-metrix tool is emphasized by Crossfield, Greenfield and McNamara (2011), who explicate the consistency of several indices measured through the tool with psycholinguistic theory. Having experimentally validated the predictive power of the tool as compared to readability assessments made by human readers, the researchers draw attention to indices included in the tool which measure word frequency, syntactic similarity among sentences and measures of co-reference in the text as having solid basis in empirical findings regarding human linguistic behavior.

The first indices provided through the Coh-Metrix tool are descriptive characteristics of the text, such as the number of words and the number of sentences in the text, the mean numbers of syllables and letters in words and their standard deviation.

The second class of indices in the Coh-Metrix tool represent ease or difficulty of text comprehension as a function of its linguistic features. Grounded on multi-level theories

of text and discourse comprehension, as explicated, among others, by Kintsch and van Dijk (1987), they incorporate measures of textual features on different linguistic and discourse levels. In particular, the ease of comprehension metrics available through this tool go beyond basic decoding measures such as graphemic-phonemic correspondence or rate of decoding. Emphasis is placed instead on factors relating to comprehension of the global organization of the text, involving processes such as inferencing or awareness of discourse structure which reveal the interplay between cohesion of the text and the reader's background knowledge and general level of reading skill development (McNamara, Kintsch, Songer & Kintsch 1996).

Attention has been drawn (McNamara et al 1996, Graesser, McNamara, Louwerse & Cai 2004) to the complex relationship between cohesion and ease of comprehension of a text, positing that higher levels of cohesion would be related to ease of comprehension for the majority of readers, but in some cases, texts displaying lower cohesion may invite readers with richer background knowledge on the subject matter handled to pay closer attention to the text and infer more meaningful connections among propositions. In this light, it appears crucial to take cohesion measures, and their effect on ease or comprehension, into consideration when analyzing their relationship to evaluations of essays, since these measures appear related to the effect of a text on the reader and to the degree of sophistication of a text.

The interaction of cohesion evident in the text and the reader's ease of reconstructing the message conveyed is also described in similar terms by Ozuru, Dempsey and McNamara (2009), who experimentally found cohesion to interact with the existence of available schemata on the part of the reader and with participants' reading skill, which was found to be vital for making use of textual features making the relationships between propositions more explicit.

Taking into consideration the multi-faceted nature of cohesion when its presence in the text and its effect on the reader are both taken into consideration, the importance of assessing cohesion in a unified manner across texts, in the sense described by Graesser, McNamara and Kulikowich (2011), appears essential for our investigation, since we seek to identify the interaction between quantified features of the text and the impression that it makes on the reader.

In our case, considerations of reading skill development or existence of prior knowledge may not directly influence overt evaluations of cohesion, and underlying cohesiveness of texts, since our readers are trained assessors of written discourse and the clear asymmetry between their age and educational background and that of the authors' can be said to guarantee the existence of necessary prior knowledge for unimpeded comprehension of the texts. However, it appears of interest how empirical measures of cohesion that we have taken into consideration in our analysis relate to evaluations by language experts, since these evaluations will allow us to see, not the extent to which our raters themselves are able to comprehend the texts under evaluation, but what degree of cohesiveness, comprehensibility and overall effect which they implicitly hypothesize, through the score that they assign, the texts to have on the general reader.

Indices which measure ease of comprehension through our tool include Narrativity, which captures elements of orality and familiarity of linguistic features used. Syntactic Simplicity, which is based on the length of sentences in words and on the familiarity of syntactic structures used, is also related to comprehension ease , as well as Word Concreteness, which is related to the proportion of words which have been found to tend to evoke images in the reader as opposed to more abstract words, whose processing is more challenging.

A number of indices capture different aspects of cohesion, including Referential Cohesion, which indicates conceptual overlap between parts of the text, so that links between ideas are made explicit in the text, facilitating comprehension of these links. Global Cohesion is measured taking into consideration the presence of causal and intentional connectives in parts of the text which display logical connections. When these connections are made explicit in discourse, the reader does not have to infer them. Verb Cohesion is a measure of the overlapping verbs in the text, greater values of which are associated with simpler narrative reconstruction on the part of the reader. Connectivity refers to logical relationships between concepts in the text made explicit through the usage of transition words capturing these relations. Temporality is based on the presence of references to time, as well as the degree of consistency of these references, for instance in terms of tense or aspect, which is hypothesized to represent a more unified structure of events which is less challenging to process.

A variety of indices contribute to identifying the degree of referential cohesion that the text displays. Referential cohesion, or co-reference, represents the degree to which the discourse structure of a text displays connectedness among its elements and as a result facilitates the reader in identifying connections between its features (Halliday & Hasan [1976] / 2013: 8). Co-reference is captured in Coh-Metrix through measures of the overlap attested among a number of features, which is measured through different indices locally, between adjacent sentences, as well as globally, taking into consideration the whole text. Features whose overlap is identified include nouns, arguments and content words.

Referential cohesion is captured in the Coh-Metrix tool through indices representing stem overlap, which refers to words representing the same lemma. These indices capture co-referential relations more freely, since no constraint of word class is

taken into consideration. In similar orientation, indices capturing anaphor overlap are additionally utilized in this tool to quantify co-reference relations, including the presence or absence of an anaphoric relationship between adjacent sentences, as well as the presence of such relationships across the whole text.

In addition to the discourse perspective discussed above, highlighting the contribution of co-reference to the quality of texture (Halliday & Hasan [1976] / 2013: 5), the significance of analyzing referential cohesion and taking it under consideration in an analysis of textual qualities and their effects on the reader's behavior in response to a text is also reinforced through observations made from a psycholinguistic perspective, exploring the relationship of co-reference relations present among sentences with processing factors on the part of the reader. Haberlandt and Graesser (1985), for instance, have reported a reading experiment where the rate at which participants read a sentence was found to increase proportionately with the number of words in a sentence which referred to new information in relation to what the participant had been exposed to earlier in the experimental session. This empirical observation of co-reference being related to ease of processing may also indirectly indicate a relationship between co-reference and ease of comprehension.

Another set of features are intended to measure cohesion through the degree of conceptual uniformity of the text. This textual quality is detected through the Latent Semantic Analysis (LSA) technique (as discussed in paragraph 3.3), which aims at detecting conceptual similarity among specific units of discourse by quantifying semantic proximity between lexis present in them on the basis of the frequency of co-occurrence of different lexemes in larger samples of discourse (Landauer and Dumais, 1997) . Indices in the Coh-Metrix tool which are based on the LSA technique capture the degree of semantic similarity between different parts of the text, and also measure

the degree to which each sentence presents the reader with new or given information in relation to the sentences that have preceded it.

The next class of indices is related to lexical diversity, which, as described above, can be identified through the Type/Token ratio (*TTR*) in a text. In this tool, the *TTR* index is measured, as well as indices directly derived from it. A comparison between types and tokens in the text enables the identification of the proportion of words which are unique in the text, as opposed to words which occur repeatedly. Lexical diversity is inversely related to cohesion, since the presence of a large proportion of words with low frequency in the text indicates a more varied thematic structure. Another set of indices constitute connectives, which explicitly encode cohesive links between concepts in the text.

These include a score of the presence of all connectives per one hundred words of the text, as well as measures of all classes of connectives (Halliday & Hasan [1976] / 2013: 226). A group of indices is devoted to textual features related to situation modelling. Through these indices, the incidence of causal and intentional verbs and particles. The compound index of causal cohesion represents the ratio of causal particles to causal verbs. This relationship is closely related to cohesion, since the presence or absence of causal particles relative to causal verbs is expected to be expressive of the degree to which information is provided regarding the manner in which the actions referred to in the text are connected, thus simplifying the reconstruction of the events represented.

Indices of syntactic complexity measure different features related to the structure of the sentences in the text. Lower syntactic complexity is generally associated with ease of processing, as less syntactically complex sentences are less demanding for

the working memory. The features measured here include the number of words which occur before the verb of the main clause, which is hypothesized to be very representative of working memory load the number of modifiers per noun-phrase.

Another source of information regarding syntactic complexity across all sentences in the text is the density of syntactic patterns, word types and phrase types, represented by the frequency of their presence in the text. This type of information is expected to reveal patterns relevant to qualities of the text such as information density, which would be represented by high frequency of verbs and nouns.

Information at the word level is additionally analyzed, focusing on content words, with a view to capturing the subject matter discussed in the text. Descriptive features are calculated, such as the frequency of words representing different grammatical categories, including the frequency of various personal pronouns. Indices derived from word frequency are also calculated, taking into consideration their relationship with reading rate (Haberlandt & Graesser 1985), which would be indicative of the degree of complexity of the text.

Psychometric characteristics of words are also taken into consideration. These are informed by the MRC Psycholinguistic database, which has been developed to assist inquiry into the properties of words which may affect the way in which they are processed. In particular, as described by Coltheart (1981), this database contains a large number of words accompanied by detailed information related to their psychological characteristics, which allows researchers to control for several factors when selecting trials for a Psycholinguistic experiment. One such factor is Age of Acquisition, representing the mean age at which these words have been empirically found to first appear in the speech of typically developing children. Other psychological

characteristics of words include meaningfulness, concreteness, imageability and familiarity. Information related to these factors has been obtained through ratings by large numbers of adult speakers. As explicated by Coltheart (1981), familiarity represents the degree to which adult native speakers report being familiar with a word. The following trait, concreteness, represents ratings of the degree to which a word is understood as representing a tangible, rather than an abstract entity. Imageability is related to the reported readiness to form a mental image in response to a word and meaningfulness expresses the degree to which a word is understood as closely related to other words. This information is captured in indices of our analysis tool, which will allow us to analyze the extent to which psychological characteristics of words may affect raters' perceptions of sophistication that students' written expression displays.

Having developed our experimental design, the nature of our corpus and of the information derived through stylometric analysis of the texts that it contains, as well as our incentives for taking into consideration these particular sets of stylometric indices, we will now report the results of our experiments and discuss their significance.

6. Results and discussion

Having analyzed the texts in our corpus with the text analysis tools and on the basis of the stylometric features described above, we subsequently sought to identify the relationship of these features with scores provided by human raters in the context of the international competition hosted by the Hewlett foundation through the Kaggle platform. We undertook this investigation by means of statistical analyses in order to model the relationship of different stylometric features and of their combinations with scores awarded by human experts and identify their predictive power of these scores. In this chapter, we will discuss the methods of analysis used as well as the significance of our findings in relation to our theoretically motivated exploration of the interaction of stylometric properties of learner script and the evaluative reaction to this script by expert human raters, which is represented in our experiments by the score assigned to essays.

6.1 Experiment one

We conducted this experiment with a view to answering our first research question, which aims at investigating whether essay scores assigned by human experts can be significantly predicted through a model of the indices described above, which capture a range of discourse qualities. Through this analysis, we wish to explore whether observations of a stylometric nature appear to be related to human evaluations of texts,

which reflect the degree to which different texts display elements which are understood as desirable of writing.

To answer our first research question, we conducted a multiple linear regression analysis, with essay score as the criterion variable and the total set of stylometric indices obtained through our two analysis tools, as described above, as independent variables. The total list of variables investigated in both our experiments can be found in Appendix II. The results of this regression analysis indicated that the model explained 61.4% of the variance in scores provided by human raters and that the model was a significant predictor of the essay score [$F(118,24)=2.914$, $p < .05$]. The resulting predictive equation and the corresponding coefficients table can be found in Appendices III and IV, respectively.

It appears then, that information obtained through our stylometric analysis, which yields a predictive model of the essay score, indicates a statistically significant relationship of this stylometric information with latent criteria influencing human decisions when evaluating writing performance. Better identifying the nature of this relationship has been the quest of experimentation in automated essay scoring since its conception by Ellis Page in the 1960s, and its wider development in the 1990s, with engines such as e-rater (Burstein, Braden-Harder, Chodorow, Hua, Kaplan, Kukich, Lu, Nolan, Rock & Wolff 1998) and has since been the focus of much related work. To this end, emphasis has been placed on identifying the most relevant stylometric features

which most strongly predict human scores and are related to theoretically informed features of discourse quality.

Apart from the overall predictive power of the model constructed in the present study, we will now focus on those specific variables which were found, through our

regression analysis, to significantly contribute to the model. We consider analyzing these specific variables to be particularly informative in relation to understanding the interaction between stylometric information and the essay score, as a measure of human appreciation of the texts.

In particular, the length of the essay expressed in its total number of words *(DESWC)* was one of the features which were found to significantly contribute to predicting the essay score (p=.004). This index belongs to the earliest set of predictors of essay score identified by Page (1966). Despite its intuitive association with the degree of elaboration of ideas which an essay presents, the informative value of text length when attempting to automatically predict essay scores has been heavily contested and has invited strong opposition to AES systems (Perelman 2012, 2013) due to its lack of association with more linguistically relevant features and due to the possibility of essays written in bad faith, by intentionally intending to manipulate such a feature, achieving an unwarranted high score. These reservations are based on the premise that text length does not guarantee relevance or coherence of ideas discussed, and by no means reflects lexicogrammatical sophistication.

While it cannot be denied that relying solely, or most heavily, on text length when providing scores automatically would be perilous in many respects, our results are in agreement with the frequent observation in AES that essay length appears to be related to quality of written work which has been evaluated by human experts. Therefore, this appears to be an informative conclusion in relation to the development of learners' writing skills. Additionally, however reasonable reservations towards heavy reliance on text length may be, if its predictive power is taken into consideration, it can be argued that including this feature in a scoring system does not appear to be undesirable, on condition that its relative weight is reduced by taking into consideration other

stylometric indices. Simultaneously, introducing into AES models measures of multiple levels of discourse sophistication, as well as automated analyses of content and of the degree of uniformity or redundancy of information in texts can interact constructively with essay length, as such measures would capture undesirable writing behaviors which would not be detected through text length alone.

Another index which was found to significantly contribute to the model (p=.019) was the Flesch Kincaid Grade Level *(RDFKGL)*. This is a readability measure which takes into account calculations of ease of comprehension of a text and converts them into values representing the twelve grade levels of schools in the United States. In this sense, this index approaches textual features from a developmental perspective, naturally associating texts whose comprehension would require a more effortful process with higher grade levels and texts whose content is more readily interpretable with earlier school grades. The conceptual grounding of this index immediately reflects the observation made by Ozuru, Dempsey and McNamara (2009) regarding the importance of the readers' background knowledge, which was found to equip readers for making use of textual features intended to assist comprehension. In our analysis, this measure of readability was found to significantly predict appreciations of the degree of sophistication and development of learners' writing skills on the part of human assessors. This observation is comparable with the complexity identified (McNamara et al 1996, McNamara, Louwerse and Cai 2004) in the relationship between cohesion of a text and its effect on the reader. In particular, texts displaying lower cohesion were at times expected to more successfully engage the interest of readers who have sufficient background knowledge of the topic discussed, since they would invite the reader to engage in the comprehension process more actively. In our case, a similar pattern emerges of an inverse relationship between holistic impressions of the degree

of ease of comprehension of a text and its degree of appreciation by human raters. in this sense, ease of comprehension would in this context appear to reflect simplicity, and thus an earlier stage of writing development on the part of learners.

The Rindex variable *(RINDEX)* was also identified as a significant predictor of essay score (p=.028). This factor is related to lexical repetition in the text, since it identifies the distance between reiterations of lexis. Thus, this index is related to co-reference, and resulting coherence of a text, as explicated by Halliday and Hasan ([1976] / 2013: 8,9), which represents unity of the content of a text, and facilitates the reader in identifying the logical flow of ideas, which are interconnected in an undisrupted manner and are relevant to the topic handled.

Finally, the ratio of causal particles to causal verbs *(SMCAUSr)* was found to significantly predict the essay score through our first experiment (p=.034). Reflecting a specific grammatical pattern, this index captures those semantic elements which represent this specific type of logical relationships attested in sentences of learner essays. Being a rather abstract notion, it may be inferred that causal relationships would be indicative of a certain degree of sophistication in the content of essays and of the degree of development in learners' writing ability which would be required in order to explicitly signal these relationships in the script produced. It may also be argued that the relationship between causal particles as opposed to causal verbs would also indicate a greater degree of sophistication, due to the different degrees of dependence relationships necessitated by these two categories, which would result in greater syntactic sophistication and may function as evidence of more detailed information provided in the text.

Having observed a statistically significant relationship between our stylometric information and the degree of sophistication of the essays in our corpus, captured in the score provided by human experts, we then proceeded to investigating which specific indices will display the strongest relationship with learner performance in the writing task analyzed.

6.2 *Experiment two*

This experiment was undertaken with a view to answering our second research question, which focuses on identifying those factors, among the total set of the stylometric indices analyzed, which contribute the most to predicting the essay score.

A stepwise regression analysis was conducted in order to determine which stylometric variables are most significant in predicting the essay score. The resulting stepwise predictive equation and the corresponding coefficients table can be found in Appendices V and VI, respectively. The results of the regression indicated that the model explained 64.8% of the variance in scores provided by human raters and that the model was a significant predictor of essay score [$F(9,133)=30.088$, $p=0$].

By selecting those features which most significantly predict the essay score, experimentation through this statistical method allowed us to examine more closely the relationship which is sought after in AES between quantifiable features and factors intrinsically related to human judgements of quality of learner essays. Therefore, it

appears essential to analyze the factors which were found to be statistically significant predictors of the essay score through this experiment as well.

The length of the text expressed in its total number of words *(DESWC)* was found to significantly predict essay score (p=0), as was the case through experiment one. As discussed above, text length is a classic index in AES systems, whose predictive power is verified here. However, it remains an index which requires judicious combination with other factors if its informative value is to be fully exploited.

Another statistically significant index in this experiment (p=0) was the average word frequency for content words *(WRDFRQc)*. Content words appear to have received a greater degree of attention than grammatical words, since the former would capture the content of the text from a semantic perspective, whereas the latter would attest to grammatical patterns involving the relationships between different words, to which thematic roles would be assigned. Another factor which appears to have influenced raters' behavior and is captured by this index is the frequency of lexis used in learner script, which is directly related to lexical diversity of the text, and thus to the range of linguistic resources which the learner displays through their script.

Connectivity *(PCCONNz)* is one more index which was found to significantly predict essay score through our second experiment (p=0). This factor captures the cohesive function of conjunction, explicated by Halliday and Hasan ([1976] / 2013: 226). This textual process comprises elements through which the relationship between other elements in the text is overtly signaled. By taking into consideration different types of conjunction relations, this index can be confidently inferred to be predictive of the degree to which learner script displays evidence of the writing skills development necessary to handle the linguistic code in such a manner that would explicitly encode

logical relationships between propositions, and would thus facilitate the reader in more readily recognizing and appreciating the logical relationships intended to become more salient by the author, through this form of very concise commentary on the relationship between elements of textual content, representative of reflection upon the content of the text produced and of mastery of the linguistic resources necessary to signal such reflection.

In similar orientation to the previous index, the presence of temporal connectives *(CNCTemp)* in the text is identified as a statistically significant predictor (p=.022). As with the elements discussed above, the presence of temporal connectives would also serve purposes of conjunction and thus aid the reader in comprehending the content of the text. Perhaps more importantly in our case, though, presence of such elements in learner script serve as evidence of awareness, on the part of the learner, of patterns relating to various relationships of content, in this case temporal relationships, and adequacy of linguistic resources and writing skill in signaling these relationships to the reader.

Another aspect of cohesion which appeared to be significant in predicting essay score (p=.013) was the presence of personal pronouns *(WRDPRO)* in the text. Pronouns are also highlighted by Halliday and Hasan ([1976] / 2013: 31) as reinforcing coherence of the text by contributing to reference relationships, which would include elements whose interpretation is dependent on information which has already been made available to the reader. Therefore, the relationship between personal pronouns and rating behavior which appeared through our experiment can be attributed to a preference, on the part of raters, for usage of personal pronouns rather than repetitions of names or other nouns indicating actors in narrative parts of the texts. Additionally to providing greater variety, the presence of pronouns referring to entities previously

mentioned can be expected to have contributed to the logical flow of events and ideas in the text in a manner that displays some degree of awareness of the reader through taking into consideration the information which the reader has already been exposed to.

In addition to measures of cohesion, syntactic patterns have also arisen as factors bearing a significant relationship with essay scores. In particular, an editorial distance index between sentences *(SYNMEDwrd)* is found to be a significant predictor (p=0) capturing the degree of difference among syntactic patterns attested in adjacent sentences, and thus capturing the variety of syntactic patterns attested in the text.

The presence of infinitives *(DRINF)* in learner script was also found to be significantly related to essay score (p=.014). This observation may be attributed to infinitives capturing actions or states referred to in learner essays, which would otherwise have been expressed through verbs. In this sense, the frequency of infinitives in learner essays necessitates a higher frequency of slightly more complex syntactic patterns and may indicate that the learner has refrained from repetitively using very simple patterns.

Similarly with our first experiment, information relating to readability has been found to impact rating behavior. Ease of text comprehension is in this case represented through an index of second language readability score *(RDL2),* which was found to significantly contribute to predictions (p=.009). In this case, the degree of ease of comprehension is sought when the reader of the text uses the language in question as a second language. This is an interesting observation indicating a possibly comparable behavior of certain types of stylometric information across speakers' languages.

Finally, essay scores were found to be significantly predicted (p=.014) by an index representing semantic similarity of content between fractions of texts. The Latent

Semantic Analysis sentence adjacent *(LSASS1d)* index is based on a probabilistic technique of identifying the degree of conceptual similarity between texts, or fractions of texts (Landauer & Dumais 1997, Furnas, Deerwester et al 1988, Landauer, Foltz & Laham 1998, Foltz, Kintsch & Landauer 1998). In this case, the degree of semantic similarity between subsequent sentences in the text is quantified, thus indirectly indicating the density of learner script in content, as well as the degree of coherence of concepts appearing in close proximity within the text. This index is therefore intended to capture human reactions to the degree of unity of content in learner essays as well as reflecting, to some extent, the degree of interest with which the reader proceeds from one sentence to the next.

It appears through our analyses that stylometric information pertaining to learner essays can significantly predict scores provided by human raters. In particular, information relating to textual coherence, achieved, among others, by means of conjunction or reference, sophistication of the texts captured through readability indices, as well as certain indices of syntactic complexity and lexical diversity are points of clear convergence between the two forms of representation of information relating to texts, the scores provided by human experts which we had at our disposal, and the quantified representation of stylistic information obtained through our stylometric analysis.

7. Conclusion

It is clear that machines do not have the capacity to process language in the same way that humans do. The underlying coherence of discourse organization in a text, its clarity of argumentation, the novelty of its ideas and its emotional appeal on the reader are not directly translatable into quantifiable measures in any straightforward manner.

Despite the difference in the manner in which discourse is processed by humans and by automated systems, our analysis indicates that human behavior in this task can be modeled by approximating human reactions to written texts on the basis of stylometric analysis of the texts. This shows that stylometric information can capture those qualities of texts which influence variable degrees of appreciation and different evaluative reactions to written texts on the part of human readers. Processing and evaluation of discourse qualities remains, of course, an extremely complex process influenced by a multitude of factors, not all of which are necessarily consciously accessible to human raters, but available for attempts at imitation through their quantifiable correlates.

Our results show a clear relationship between information of a stylometric nature, which identify the frequency structure of our texts and capture information which escapes conscious detection by human speakers, and quality judgements of the same corpus of texts, represented by the score which expert raters have assigned to the essays. This clear relationship between quantifiable, machine readable features of texts and their appeal to human raters enables the employment of stylometric information for

solving various problems, including the increasingly socially relevant potential of automatic essay scoring.

Taking into consideration the escalating need for fast and reliable assessment of large numbers of long-form constructed responses in many educational contexts, we have explored the possibility for automated essay scoring being a tool in the hands of learners and teachers. Given their benefits as an assessment instrument, since it introduces elements of critical thinking and reasoning skills absent from other standardized assessment tasks, such as multiple-choice questions, open-ended tasks now form an essential part of effective pedagogical evaluation. This need, in combination with the affordances of the World Wide Web for education purposes, as powerfully realized through Massive Open Online Courses, highlight the vital contribution that AES can make to promoting wide accessibility of learning opportunities while simultaneously preventing a potentially overwhelming workload for instructors.

In addition to the social factors which make research into automatic essay scoring highly relevant, we have discussed its limitations and ethical considerations regarding responsible use of its features. We have provided a cursory overview of existing AES systems since their inception by Ellis Page in the 1960s and we have discussed some of their features and methodologies.

For our own attempt, we have analyzed essay data originally provided in the context of a recent essay scoring competition. We have utilized stylometric indices in our effort to build a predictive model that can successfully predict essay scores provided by human raters. Our results show a clear relationship between quantifiable stylistic

information and features of discourse which appear to affect human judgements of the quality of a written assignment in the context of a language examination.

These findings highlight the usefulness and relevance of automated essay scoring systems, which makes it important to extend investigations of the interplay between stylometric features of texts and their effect on the reader, taking into consideration a variety of stylistic indices as well as varying textual and characteristics.

References

Attali, Y., & Burstein, J. (2006). "Automated essay scoring with e-rater V. 2". *The Journal of Technology, Learning and Assessment, 4*(3).

Balfour, S. P. (2013). "Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review™". *Research & Practice in Assessment, 8*: 40-48.

Bennett, R. E., & Bejar, I. I. (1998). "Validity and automad scoring: It's not only the scoring". *Educational Measurement: Issues and Practice*, 17(4), 9-17.

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., & Chodorow, M. (2013). "TOEFL11: A corpus of non-native English". *ETS Research Report Series, 2013*(2), i-15.

Breland, H. M. (1996). "Word frequency and word difficulty: A comparison of counts in four corpora". *Psychological Science, 7*(2), 96-99.

Breland, H. M., Bonner, M. W., & Kubota, M. Y. (1995). "Factors in performance on brief, impromptu essay examinations". *ETS Research Report Series, 1995*(2), i-35.

Breland, H., Kubota, M., Nickerson, K., Trapani, C., & Walker, M. (2004). "New SAT® Writing Prompt Study: Analyses of Group Impact and Reliability". *ETS Research Report Series, 2004*(1), i-18.

Brent, E. & Townsend, M. (2006). "Automated essay grading in the Sociology classroom: finding common ground". In P.F. Ericsson & R.H. Haswell (eds.), *Machine Scoring of Student Essays: Truth and Consequences*. Logan: Utah State University Press, 177-198.

Bridgeman, B., & McHale, F. (1996). "Gender and ethnic group differences on the GMAT Analytical Writing Assessment". *ETS Research Report Series, 1996*(1), i-14.

Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., & Wolff, S. (1998). "Computer analysis of essay content for automated score prediction: A prototype automated scoring system for GMAT analytical writing assessment essays". *ETS Research Report Series, 1998*(1), i-67.

Burstein, J., & Chodorow, M. (1999). "Automated essay scoring for nonnative English speakers". In *Proceedings of a Symposium on Computer Mediated Language Assessment and Evaluation in Natural Language Processing*. Association for Computational Linguistics, 68-75.

Burstein, J., Chodorow, M., & Leacock, C. (2004). "Automated essay evaluation: The Criterion online writing service". *AI Magazine, 25*(3), 27.

Burstein, J., & Higgins, D. (2005). “Advanced Capabilities for Evaluating Student Writing: Detecting Off-Topic Essays Without Topic-Specific Training”. In *Artificial Intelligence in Education*. Amsterdam: IOS Press.

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (2001). “Enriching automated essay scoring using discourse marking”.

Burstein, J., Marcu, D., & Knight, K. (2003). “Finding the WRITE stuff: automatic identification of discourse structure in student essays”. *IEEE Intelligent Systems, 18(1), 32–39.*

Cahill, A., Chodorow, M. & Flor, M. (2018). “Developing an e-rater advisory to detect Babel-generated essays”. *The Journal of Writing Analytics* 2: 203-224.

Chodorow, M., Gamon, M., & Tetreault, J. (2010). “The utility of article and preposition error correction systems for English language learners: Feedback and assessment”. *Language Testing*, *27*(3), 419-436.

Clauser, B. E. (2000). “Recurrent issues and recent advances in scoring performance assessments”. *Applied Psychological Measurement*, *24*(4), 310-324.

Coltheart, M. (1981). “The MRC psycholinguistic database”. *The Quarterly Journal of Experimental Psychology*, *33*(4), 497-505.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). “Assessing text readability using cognitively based indices”. *Tesol Quarterly*, *42*(3), 475-493.

Davidson, P. (2018). “Robo-grader: Can computers really rate students’ essays?”. In T, Aksit, H.I. Mengu & R. Turner (eds.), *Bridging Teaching, Learning and Assessment in the English Language Classroom*. Newcastle: Cambridge Scholars Publishing, 44-51.

Deane, P. (2013). “On the relation between automated essay scoring and modern views of the writing construct”. *Assessing Writing*, *18*(1), 7-24.

Ehsan, N. & Faili, H. (2013) Grammatical and context-sensitive error correction using a statistical machine translation framework. *Software: Practice and Experience* **43**:2, 187-206.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). “The measurement of textual coherence with latent semantic analysis”. *Discourse Processes*, *25*(2-3), 285-307.

Foltz, P.W., Laham, D. & Landauer, T.K. (1999). “Automated Essay Scoring: Applications to Educational Technology”. In B. Collis & R. Oliver (Eds.), *Proceedings of ED-MEDIA 1999--World Conference on Educational Multimedia, Hypermedia & Telecommunications*. Seattle, WA USA: Association for the Advancement of Computing in Education (AACE), 939-944.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). "The intelligent essay assessor: Applications to educational technology". *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1*(2), 939-944.

Furnas, G. W., Deerwester, S., Dumais, S. T., Landauer, T. K., Harshman, R. A., Streeter, L. A., & Lochbaum, K. E. (1988). "Information retrieval using a singular value decomposition model of latent semantic structure". In *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 465-480.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). "Constructing inferences during narrative text comprehension". *Psychological Review, 101*(3), 371.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). "Coh-Metrix: Providing multilevel analyses of text characteristics". *Educational Researcher*, *40*(5), 223-234.

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). "What do readers need to learn in order to process coherence relations in narrative and expository text". *Rethinking Reading Comprehension*, 82-98.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). "Coh-Metrix: Analysis of text on cohesion and language". *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193-202.

Haberlandt, K. F., & Graesser, A. C. (1985). "Component processes in text comprehension and some of their interactions". *Journal of Experimental Psychology: General*, *114*(3), 357.

Halliday, M.A.K. & Hasan, R. (2013). *Cohesion in English*, 2nd ed. [1st ed.: 1976]. New York: Routledge.

Herrington, A., & Stanley, S. (2012). "Criterion SM: Promoting the standard". *Race and Writing Assessment*, 47-61.

Hyland, K. (1990). "A genre description of the argumentative essay". *RELC Journal*, *21*(1), 66-78.

Kintsch, W. (1988). "The role of knowledge in discourse comprehension: A construction-integration model". *Psychological Review*, *95*(2), 163.

Kintsch, W., & van Dijk, T. A. (1978). "Toward a model of text comprehension and production". *Psychological review*, *85*(5), 363.

Kubát, M., & Milička, J. (2013). "Vocabulary richness measure in genres". *Journal of Quantitative Linguistics*, *20*(4), 339-349.

Laham, D. (1997). “Latent semantic analysis approaches to categorization”. In *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, 979.

Landauer, T. K., & Dumais, S. T. (1997). “A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge”. *Psychological Review*, 104(2), 211.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). “An introduction to latent semantic analysis”. *Discourse Processes*, *25*(2-3), 259-284.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). “How well can passage meaning be derived without using word order?” A comparison of Latent Semantic Analysis and humans. In *Proceedings of the 19th Annual Meeting of the Cognitive Science Society*, 412-417.

Li, X., Chen, M., Nie, J., Liu, Z., Feng, Z., & Cai, Y. (2018). “Coherence-Based Automated Essay Scoring Using Self-attention”. In M. Sun, T. Liu, X. Wang, Z. Liu & Y. Liu (eds.), *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*: *Proceedings of the 17th China National Conference, CCL 2018, and 6th International Symposium, NLP-NABD 2018*, Changsha, China, October 19–21, 2018, 386-397.

Lipnevich, A. A., & Smith, J. K. (2009). “Effects of differential feedback on students’ examination performance”. *Journal of Experimental Psychology: Applied*, *15*(4), 319.

Macdonald, N. H., Frase, L. T., Gingrich, P. S., & Keenan, S. A. (1982). “The Writer's Workbench: Computer aids for text analysis”. *Educational Psychologist*, *17*(3), 172-179.

McAllister, K.S. & White, E. (2006). “Interested complicities: The Dialectic of Computer-Assisted Writing Assessment”. In P.F. Ericsson & R.H. Haswell (eds.), *Machine Scoring of Student Essays: Truth and Consequences.* Logan: Utah State University Press, 8-27.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). “Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text”. *Cognition and Instruction*, *14*(1), 1-43.

Ozuru, Y., Dempsey, K., & McNamara, D. S. (2009). “Prior knowledge, reading skill, and text cohesion in the comprehension of science texts”. *Learning and Instruction*, *19*(3), 228-242.

Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, *47*(5), 238-243.

Page, E. B., & Petersen, N. S. (1995). “The computer moves into essay grading: Updating the ancient test”. *Phi Delta Kappan*, *76*(7), 561.

Perelman, L. (2012). “Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring (AES)”. In C. Bazerman, C. Dean, J. Early, K. Lunsford, S. Null, P. Rogers, & A. Stansell (eds.), *International Advances in Writing Research: Cultures, Places, Measures.* Colorado: Parlor Press, 121-131.

Perelman, L. C. (2013). “Critique of Mark D. Shermis & Ben Hamner,'Contrasting State-of-the-Art Automated Scoring of Essays: Analysis.”. *Journal of Writing Assessment*, *6* (1).

Perelman, L. (2014). “When ‘the state of the art’ is counting words”. *Assessing Writing*, *21*: 104-111.

Persing, I., & Ng, V. (2015). “Modeling argument strength in student essays”. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Volume 1: Long Papers.* Vol. 1, 543-552.

Popescu, I. I., Mačutek, J., Altmann, G. (2009b). *Aspects of Word Frequencies*. Lüdenscheid: Ram-Verlag.

Popescu, I. I., Mačutek, J., Kelih, E., Čech, R., Best, K. H., & Altmann, G. (2010). *Vectors and Codes of Text*. Lüdenscheid: RAM-Verlag.

Popescu, I. I., Čech, R., & Altmann, G. (2011). *The Lambda-Structure of Texts*. Lüdenscheid: Ram-Verlag.

Reilly, E. D., Stafford, R. E., Williams, K. M., & Corliss, S. B. (2014). “Evaluating the validity and applicability of automated essay scoring in two massive open online courses”. *The International Review of Research in Open and Distributed Learning*, *15*(5).

Rothermel, B.A (2006). “Automated Writing Instruction: Computer-assisted or computer-driven pedagogies?”. In P.F. Ericsson & R.H. Haswell (eds.), *Machine Scoring of Student Essays: Truth and Consequences*. Logan: Utah State University Press, 199-210.

Shi, L. (2001). “Native- and nonnative-speaking EFL teachers’ evaluation of Chinese students’ English writing”. *Language Testing*, 18, 303–325

Shermis, M. D. (2014). “State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration”. *Assessing Writing*, *20*: 53-76.

Shermis, M. D., Burtsein, J. & Bursky, S.A. (2013). Introduction to automated essay evaluation”. In M.D. Shermis & J. Burstein (eds.) *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. New York: Routledge.

Stab, C., & Gurevych, I. (2014). “Identifying argumentative discourse structures in persuasive essays”. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing EMNLP*, 46-56.

Suen, H. K. (2014). “Peer assessment for massive open online courses (MOOCs)”. *The International Review of Research in Open and Distributed Learning*, *15*(3).

Tuzzi, A., Popescu, I. I., & Altmann, G. (2010). “The golden section in texts”. *ETC–Empirical Text and Culture Research*, *4*: 30-41.

Wang, X. (2018). “The effect of AES system on writing metacognitive strategy in senior high school EFL writing: Taking Nanchong First Middle School as an example”. In *Proceedings of the 5th International Conference on Education, Language, Art and Inter-cultural Communication (ICELAIC 2018)*. Atlantis Press, 104-110.

Ware, P. (2011). “Computer-generated feedback on student writing”. *TESOL Quarterly* 45(4): 769-774.

Wimmer, G., & Altmann, G. (1999). “Review Article: On Vocabulary Richness”. *Journal of Quantitative Linguistics, 6(1), 1–9.*

Wolf, F., & Gibson, E. (2005). “Representing discourse coherence: A corpus-based study”. *Computational Linguistics*, *31*(2), 249-287.

Wyllys, R. E. (1981). Empirical and theoretical bases of Zipf's law.

Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011). “A new dataset and method for automatically grading ESOL texts”. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 180-189.

Yannakoudakis, H., & Briscoe, T. (2012). “Modeling coherence in ESOL learner texts”. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics, 33-43.

APPENDIX I

The following information, obtained through the Kaggle platform (available at https://www.kaggle.com/competitions), provides a description of the features of the essay set analyzed in the present study, as well as the scoring guidelines upon which scores provided by human experts were based.

Essay Set #1

Type of essay:	Persuasive/ Narrative/Expository
Grade level:	8
Training set size:	1,785 essays
Final evaluation set size:	592 essays
Average length of essays:	350 words
Scoring:	Score1, Score2, Resolved Score
Rubric range:	1-6
Resolved score range:	2-12

Prompt

More and more people use computers, but not everyone agrees that this benefits society. Those who support advances in technology believe that computers have a positive effect on people. They teach hand-eye coordination, give people the ability to learn about faraway places and people, and even allow people to talk online with other people. Others have different ideas. Some experts are concerned that people are spending too much time on their computers and less time exercising, enjoying nature, and interacting with family and friends.

Write a letter to your local newspaper in which you state your opinion on the effects computers have on people. Persuade the readers to agree with you.

Rubric Guidelines

Score Point 1: An undeveloped response that may take a position but offers no more than very minimal support. Typical elements:

- Contains few or vague details.
- Is awkward and fragmented.
- May be difficult to read and understand.
- May show no awareness of audience.

Score Point 2: An under-developed response that may or may not take a position. Typical elements:

- Contains only general reasons with unelaborated and/or list-like details.
- Shows little or no evidence of organization.
- May be awkward and confused or simplistic.
- May show little awareness of audience.

Score Point 3: A minimally-developed response that may take a position, but with inadequate support and details. Typical elements:

- Has reasons with minimal elaboration and more general than specific details.
- Shows some organization.
- May be awkward in parts with few transitions.
- Shows some awareness of audience.

Score Point 4: A somewhat-developed response that takes a position and provides adequate support. Typical elements:

- Has adequately elaborated reasons with a mix of general and specific details.
- Shows satisfactory organization.
- May be somewhat fluent with some transitional language.
- Shows adequate awareness of audience.

Score Point 5: A developed response that takes a clear position and provides reasonably persuasive support. Typical elements:

- Has moderately well elaborated reasons with mostly specific details.
- Exhibits generally strong organization.
- May be moderately fluent with transitional language throughout.
- May show a consistent awareness of audience.

Score Point 6: A well-developed response that takes a clear and thoughtful position and provides persuasive support. Typical elements:

- Has fully elaborated reasons with specific details.
- Exhibits strong organization.
- Is fluent and uses sophisticated transitional language.
- May show a heightened awareness of audience.

Adjudication Rules

- If the two scores are adjacent, the final score for an item is the sum of the two scores.
- If the two scores are non-adjacent, the final score is determined by an expert scorer.

APPENDIX II

In the experiments described above, the following stylometric features of texts were utilized, which are represented in the resulting predictive equations by the variables listed below.

Number of paragraphs (DESPC), Number of sentences (DESSC), Number of words (DESWC), Mean length of paragraphs (DESPL), Standard deviation of the mean length of paragraphs (DESPLd), Mean number of words (length) of sentences in (DESSL), Standard deviation of the mean length of sentences (DESSLd), Mean number of syllables (length) in words (DESWLsy), Standard deviation of the mean number of syllables in words (DESWLsyd), Mean number of letters (length) in words (DESWLlt), Standard deviation of the mean number of letter in words (DESWLltd), Narrativity (PCNARz), (PCNARp), Syntactic Simplicity (PCSYNz), (PCSYNp), Word Concreteness (PCCNCz), (PCCNCp), Referential Cohesion (PCREFz), (PCREFp), Deep Cohesion (PCDCz), (PCDCp), Verb Cohesion (PCVERBz), (PCVERBp), Connectivity (PCCONNz), (PCCONNp), Temporality (PCTEMPz), (PCTEMPp), Noun overlap (CRFNO1), (CRFNOa), Argument overlap (CRFAO1), (CRFAOa), Stem overlap (CRFSO1), (CRFSOa), Content word overlap (CRFCWO1), (CRFCWO1d), (CRFCWOa), standard deviation of LSA cosines for adjacent units (CRFCWOad), Anaphor overlap (CRFANP1), (CRFANPa), LSA sentence adjacent (LSASS1), (LSASS1d), mean LSA cosine (LSASSp), standard deviation of LSA cosine (LSASSpd), mean of the LSA cosines between adjacent paragraphs (LSAPP1), standard deviation

of LSA cosines between adjacent paragraphs (LSAPP1d), givenness of sentence (LSAGN), standard deviation of givenness of sentence (LSAGNd), Type-token ratio (LDTTRc), Type token ratio for all words (LDTTRa), the LDMTLDa lexical diversity index, the LDVOCDa lexical diversity index, the incidence of connectives (CNCAll), Causal Connectives (CNCCaus), logic connectives (CNCLogic), adversative/contrastive connectives (CNCADC), temporal connectives (CNCTemp), extended temporal connectives (CNCTempx), additive connectives (CNCAdd), positive connectives (CNCPos), negative connectives (CNCNeg), causal verbs (SMCAUSv), causal verbs and causal particles (SMCAUSvp), intentional actions, events, and particles (SMINTEp), the ratio of causal particles to causal verbs (SMCAUSr), the ratio of intentional particles to intentional actions/events (SMINTEr), the LSA overlap between verbs (SMCAUSlsa), the WordNet overlap between verbs (SMCAUSwn), Temporal cohesion (SMTEMP), Words before main verb (SYNLE), Modifiers per noun phrase (SYNNP), the mean minimum distance between adjacent sentences (SYNMEDpos), the SYNMEDwrd index, the SYNMEDlem index, syntactic structure similarity of adjacent sentences, (SYNSTRUTa), Syntactic structure similarity of all sentences (SYNSTRUTt), noun phrases (DRNP), verb phrases (DRVP), adverbial phrases (DRAP), preposition phrases (DRPP), agentless passive voice forms (DRPVAL), negation expressions (DRNEG), gerunds (DRGERUND), infinitives (DRINF), nouns (WRDNOUN), verbs (WRDVERB), adjectives (WRDADJ), adverbs (WRDADV), personal pronouns (WRDPRO), first person, single form pronouns (WRDPRP1s), first peron, plural form pronouns (WRDPRP1p), second person pronouns (WRDPRP2), third person, single form pronouns (WRDPRP3s), third person, plural form pronouns (WRDPRP3p), average word frequency for content words (WRDFRQc), average word frequency for all words (WRDFRQa), average minimum

word frequency in sentences (WRDFRQmc), Age of acquisition (WRDAOAc), Familiarity (WRDFAMc), Concreteness (WRDCNCc), Imagability (WRDIMGc), Meaningfulness (WRDMEAc), Polysemy (WRDPOLc), Hypernymy (WRDHYPn), (WRDHYPv), (WRDHYPnv), Flesch Reading Ease: RDFRE, Flesch_Kincaid Grade Level: RDFKGL, second language readability score (RDL2), h-point (h), entropy (ENTROPY), normalized entropy (NORMENTROPY), average token length (AVTOKENLEN), standard deviation of token length (TOKENLENSD), hapax legomenon percentage (HAPAXPERCENTAGE), h-point (H), adjusted modulus (ADJUSTEDMODULUS), Gini coefficient (GINISCOEF), the R4 index (GINISCOEFR4), the L index (L), Lambda (LAMBDA), the P/A ratio index (PARATIO), the R1 index (R1), the (RINDEX), Repeat Rate (RR), Relative Repeat Rate of McIntosh, (TLFS) index, the Writes's View index (WRITTERSVIEW) and the Yule's K index (YULEK).

APPENDIX III

Below is the first predictive model of the essay score constructed in the present study through our first experiment, in which the relationship of the stylometric features listed above with scores awarded to the essays by human raters was explored through a linear regression analysis. As discussed in chapter 6, results indicated that the essay score can be predicted through the following model in a statistically significant manner and the *RINDEX*, *DESWC*, *SMCAUSr* and *RDFKGL* indices were found to significantly contribute to the model. Coefficients used in this predictive equation are presented in the following appendix.

(1) Essay score = 98.467 – 1.543 *(TOKENS)* + .791 *(ENTROPY)* + .961 *(NORMENTROPY)* + .081 *(AVTOKENLEN)* - .032 *(TOKENLENSD)* + .942 *(HAPAXPERCENTAGE)* - .963 *(H)* + .282 *(ADJUSTEDMODULUS)* + .329 *(GINISCOEFR4)* + 1.685 *(L)* – 1.722 *(LAMBDA)* + .836 *(PARATIO)* - .336 *(R1)* - .822 *(RINDEX)* + .144 *(RR)* - .614 *(RRMC)* - .061 *(TLFS)* + .026 *(WRITTERSVIEW)* - 1.333 *(YULESK)* - .594 *(DESSC)* + 1.382 *(DESWC)* - .088 *(DESSL)* + .024 *(DESSLd)* + .329 *(DESWLsy)* - .104 *(DESWLsyd)* + .329 *(DESWLsy)* - .104 *(DESWLsyd)* - .019 *(DESWLlt)* - .028 *(DESWLltd)* + .249 *(PCNARz)* - .023 *(PCNARp)* - .681 *(PCSYNz)* - .094 *(PCSYNp)* - .229 *(PCCNCz)* - .253 *(PCCNCp)* + .340 *(PCREFz)* - .203 *(PCREFp)* + 6.000 *(PCDCz)* + .050 *(PCDCp)* - .205 *(PCVERBz)* + .006 *(PCVERBp)* - .044 *(PCCONNz)* + .111 *(PCCONNp)* + .064 *(PCTEMPz)* - .345 *(PCTEMPp)* - .047 *(CRFNO1)* - .067 *(CRFAO1)* + .018 *(CRFSO1)* + .004 *(CRFNOa)* + .163 *(CRFAOa)* - .234 *(CRFSOa)* + .201 *(CRFCWO1)* + .067 *(CRFCWO1d)* - .171 *(CRFCWOa)* - .151

(CRFCWOad) + .260 *(LSASS1)* - .399 *(LSASS1d)* - .151 *(LSASSp)* - .039 *(LSAsspd)* - .312 *(LSAGN)* + .410 *(LSAGNd)* + .124 *(LDTTRc)* + .195 *(LDTTRa)* + .020 *(LDMTLD)* - .068 *(LDVOCD)* - .156 *(CNCAII)* - .278 *(CNCCaus)* - .430 *(CNCLogic)* + .109 *(CNCADC)* - .073 *(CNCTemp)* + .187 *(CNCTempx)* - .016 *(CNCAdd)* + .379 *(SMCAUSv)* - .009 *(SMCAUSvp)* - .333 *(SMINTEp)* + .248 *(SMCAUSr)* - .333 *(SMINTEr)* - .031 *(SMCAUSIsa)* + .330 *(SMCAUSwn)* + .051 *(SMTEMP)* + .286 *(SYNLE)* - .191 *(SYNNP)* - .106 *(SYNMEDpos)* – 1.414 *(SYNMEDwrd)* + .609 *(SYNMEDlem)* + .434 *(SYNSTRUTa)* + .132 *(SYNSTRUTt)* - .277 *(DRNP)* - .199 *(DRVP)* - .011 *(DRAP)* + .020 *(DRPP)* + .108 *(DRPVAL)* - .162 *(DRNEG)* + .179 *(DRGERUND)* - .235 *(DRINF)* + .367 *(WRDNOUN)* + .152 *(WRDVERB)* + .019 *(WRDADJ)* + .047 *(WRDADV)* - .089 *(WRDPRO)* + .065 *(WRDPRP1s)* + .106 *(WRDPRP1p)* + .094 *(WRDPRP2)* - .174 *(WRDPRP3s)* + .131 *(WRDPRP3s)* - .084 *(WRDFRQc)* + .022 *(WRDFRQa)* + .226 *(WRDFRQmc)* – 266 *(WRDAOAc)* + .464 *(WRDFAMc)* + .307 *(WRDCNCc)* + .066 *(WRDPOLc)* - .084 *(WRDHYPn)* + .063 *(WRDHYPv)* - .021 *(WRDHYPnv)* - .331 *(RDFRE)* – 1.763 *(RDFKGL)* - .478 *(RDL2)*

APPENDIX IV

Below is the coefficients table pertaining to experiment one, where the dependent variable is the score assigned to learner script and the independent variables are stylometric features of texts, whose analysis is detailed in our experimental section.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	98.467	185.096		.532	.600
	TOKENS	-.018	.019	-1.543	-.946	.354
	ENTROPY	2.432	7.347	.791	.331	.743
	NORMENTROPY	85.037	299.342	.961	.284	.779
	AVGTOKENLEN	.437	1.112	.081	.393	.698
	TOKENLENSD	-.230	1.271	-.032	-.181	.858
	HAPAXPERCENTAGE	15.770	28.876	.942	.546	.590
	H	-.836	.885	-.963	-.945	.354
	ADJUSTEDMODULUS	-.279	.814	-.282	-.343	.735
	GINISCOEFR4	7.517	86.908	.329	.086	.932
	L	.045	.058	1.685	.784	.441
	LAMBDA	-21.687	21.586	-1.722	-1.005	.325
	PARATIO	24.252	36.940	.836	.657	.518
	R1	-15.190	24.127	-.336	-.630	.535
	RINDEX	-50.231	21.473	-.822	-2.339	.028
	RR	42.734	530.024	.144	.081	.936
	RRMC	-100.477	376.914	-.614	-.267	.792
	TLFS	-.077	.202	-.061	-.381	.707
	WRITTERSVIEW	.176	2.246	.026	.078	.938
	YULESK	-.077	.080	-1.333	-.964	.345
	DESSC	-.086	.068	-.594	-1.268	.217
	DESWC	.016	.005	1.382	3.162	.004
	DESSL	.000	.000	-.088	-.506	.618
	DESSLd	4.469E-6	.000	.024	.090	.929
	DESWLsy	.003	.002	.329	1.576	.128
	DESWLsyd	-.001	.004	-.104	-.295	.770
	DESWLlt	.000	.001	-.019	-.059	.953
	DESWLltd	.000	.002	-.028	-.074	.942
	PCNARz	.001	.001	.249	.688	.498
	PCNARp	.000	.000	-.023	-.132	.896
	PCSYNz	-.001	.001	-.681	-.689	.497
	PCSYNp	.000	.000	-.094	-.181	.858
	PCCNCz	.000	.001	-.229	-.595	.558
	PCCNCp	.000	.000	-.253	-.753	.459
	PCREFz	.000	.001	.340	.347	.732
	PCREFp	.000	.000	-.203	-.925	.364
	PCDCz	.001	.001	.600	.921	.366
	PCDCp	2.251E-6	.000	.050	.297	.769
	PCVERBz	.000	.001	-.205	-.415	.682
	PCVERBp	3.317E-7	.000	.006	.030	.977
	PCCONNz	.000	.001	-.044	-.058	.954
	PCCONNp	1.120E-5	.000	.111	.370	.714
	PCTEMPz	2.820E-5	.000	.064	.099	.922
	PCTEMPp	.000	.000	-.345	-1.216	.236
	CRFNO1	.000	.001	-.047	-.285	.778
	CRFAO1	.000	.001	-.067	-.379	.708
	CRFSO1	9.491E-5	.001	.018	.086	.932
	CRFNOa	2.892E-5	.002	.004	.016	.987
	CRFAOa	.001	.001	.163	.751	.460

CRFSOa	-.001	.001	-.234	-1.184	.248
CRFCWO1	.005	.009	.201	.526	.604
CRFCWO1d	.003	.011	.067	.248	.806
CRFCWOa	-.004	.014	-.171	-.316	.754
CRFCWOad	-.007	.013	-.151	-.542	.593
LSASS1	.004	.007	.260	.608	.549
LSASS1d	-.011	.008	-.399	-1.404	.173
LSASSp	-.003	.008	-.151	-.326	.748
LSASSpd	-.001	.012	-.039	-.120	.905
LSAGN	-.006	.007	-.312	-.953	.350
LSAGNd	.017	.012	.410	1.452	.160
LDTTRc	.001	.003	.124	.535	.597
LDTTRa	.002	.004	.195	.656	.518
LDMTLD	1.089E-6	.000	.020	.149	.883
LDVOCD	.000	.000	-.068	-.211	.835
CNCAll	.000	.000	-.156	-.309	.760
CNCCaus	.000	.000	-.278	-.618	.543
CNCLogic	.000	.000	-.430	-.975	.340
CNCADC	1.437E-5	.000	.109	.392	.698
CNCTemp	.000	.000	-.073	-.310	.759
CNCTempx	2.834E-5	.000	.187	.967	.343
CNCAdd	.000	.000	-.016	-.030	.976
SMCAUSv	3.987E-5	.000	.379	.848	.405
SMCAUSvp	.000	.000	-.009	-.025	.981
SMINTEp	.000	.000	-.333	-1.281	.212
SMCAUSr	.001	.000	.248	2.242	.034
SMINTEr	-.001	.000	-.333	-1.733	.096
SMCAUSIsa	-.002	.013	-.031	-.121	.905
SMCAUSwn	.004	.002	.330	1.730	.097
SMTEMP	.000	.001	.051	.316	.755
SYNLE	.000	.000	.286	1.475	.153
SYNNP	-.002	.002	-.191	-1.101	.282
SYNMEDpos	-.002	.005	-.106	-.320	.752
SYNMEDwrd	-.019	.022	-1.414	-.883	.386
SYNMEDlem	.008	.019	.609	.436	.667
SYNSTRUTa	.015	.020	.434	.763	.453
SYNSTRUTt	.005	.020	.132	.266	.792
DRNP	.000	.000	-.277	-1.136	.267
DRVP	.000	.000	-.199	-.911	.371
DRAP	.000	.000	-.011	-.044	.966
DRPP	1.014E-6	.000	.020	.114	.911
DRPVAL	5.300E-5	.000	.108	.800	.432
DRNEG	.000	.000	-.162	-.935	.359
DRGERUND	1.939E-5	.000	.179	.989	.332
DRINF	.000	.000	-.235	-1.588	.125
WRDNOUN	8.597E-6	.000	.367	1.661	.110
WRDVERB	6.775E-6	.000	.152	.548	.589
WRDADJ	1.119E-6	.000	.019	.112	.912
WRDADV	2.520E-6	.000	.047	.172	.865
WRDPRO	.000	.000	-.089	-.188	.853
WRDPRP1s	7.256E-6	.000	.065	.355	.726
WRDPRP1p	1.503E-5	.000	.106	.406	.688
WRDPRP2	4.800E-6	.000	.094	.240	.812
WRDPRP3s	.000	.000	-.174	-1.012	.322
WRDPRP3p	1.131E-5	.000	.131	.536	.597
WRDFRQc	-.001	.006	-.084	-.195	.847
WRDFRQa	9.580E-5	.001	.022	.177	.861
WRDFRQmc	.001	.001	.226	.900	.377
WRDAOAc	.000	.000	-.266	-1.694	.103
WRDFAMc	6.489E-6	.000	.464	1.773	.089
WRDCNCc	8.322E-6	.000	.307	1.043	.308
WRDIMGc	1.649E-6	.000	.066	.560	.581
WRDMEAc	.000	.000	-.041	-.167	.869
WRDPOLc	.000	.000	.066	.345	.733
WRDHYPn	.000	.000	-.084	-.398	.694
WRDHYPv	.000	.002	.063	.305	.763
WRDHYPnv	.000	.001	-.021	-.129	.899
RDFRE	.000	.000	-.331	-1.421	.168
RDFKGL	.000	.000	-1.763	-2.510	.019
RDL2	.000	.000	-.478	-.631	.534

a. Dependent Variable: Score

APPENDIX V

The stepwise predictive model, designed through our second experiment, appears below, which models the relationship of essay score with those stylometric features which were found to most strongly predict the score assigned to learner script by human raters. Coefficients pertaining to this experiment are presented in the following appendix.

(2) Essay score = 18.670 + .628 *(DESWC)* - .435 *(WRDFRQc)*

+ .255 *(PCCONNz)* - .138 *(WRDPRO)* + .245 *(SYNMEDwrd)* - .127 *(CNCTemp)*

- .138 *(DRINF)* + .187 *(RDL2)* - .137 *(LSASS1d)*

APPENDIX VI

In the following table, coefficients pertaining to the second experiment are presented, based on which essay score is predicted on the basis of its relationship with the set of stylometric indices utilized in the equation presented in the previous appendix.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6.191	.287		21.591	.000
	DESWC	.006	.001	.527	7.368	.000
2	(Constant)	1.872	.675		2.775	.006
	DESWC	.010	.001	.858	10.935	.000
	LDTTRa	.007	.001	.541	6.889	.000
3	(Constant)	12.719	2.463		5.164	.000
	DESWC	.009	.001	.808	10.874	.000
	LDTTRa	.005	.001	.403	5.076	.000
	WRDFRQc	-.004	.001	-.288	-4.556	.000
4	(Constant)	14.370	2.413		5.955	.000
	DESWC	.008	.001	.709	9.232	.000
	LDTTRa	.003	.001	.271	3.184	.002
	WRDFRQc	-.004	.001	-.287	-4.719	.000
	PCCONNz	.000	.000	.224	3.529	.001
5	(Constant)	13.111	2.382		5.503	.000
	DESWC	.009	.001	.736	9.794	.000
	LDTTRa	.003	.001	.270	3.261	.001
	WRDFRQc	-.003	.001	-.231	-3.735	.000
	PCCONNz	.000	.000	.236	3.814	.000
	WRDPRO	.000	.000	-.176	-3.015	.003
6	(Constant)	13.964	2.335		5.982	.000
	DESWC	.007	.001	.604	7.073	.000
	LDTTRa	.001	.001	.114	1.185	.238
	WRDFRQc	-.004	.001	-.285	-4.533	.000
	PCCONNz	.000	.000	.232	3.863	.000
	WRDPRO	.000	.000	-.175	-3.071	.003
	SYNMEDwrd	.003	.001	.200	2.978	.003
7	(Constant)	15.513	1.937		8.009	.000
	DESWC	.006	.001	.527	9.539	.000
	WRDFRQc	-.004	.001	-.316	-5.513	.000
	PCCONNz	.000	.000	.258	4.585	.000
	WRDPRO	.000	.000	-.174	-3.065	.003
	SYNMEDwrd	.003	.001	.243	4.318	.000
8	(Constant)	15.642	1.906		8.208	.000
	DESWC	.007	.001	.568	9.967	.000
	WRDFRQc	-.004	.001	-.324	-5.739	.000
	PCCONNz	.000	.000	.237	4.234	.000
	WRDPRO	.000	.000	-.166	-2.962	.004
	SYNMEDwrd	.004	.001	.257	4.621	.000
	CNCTemp	.000	.000	-.133	-2.381	.019
9	(Constant)	15.585	1.881		8.284	.000
	DESWC	.007	.001	.602	10.299	.000
	WRDFRQc	-.004	.001	-.314	-5.626	.000
	PCCONNz	.000	.000	.257	4.583	.000
	WRDPRO	.000	.000	-.142	-2.524	.013
	SYNMEDwrd	.003	.001	.247	4.471	.000
	CNCTemp	.000	.000	-.125	-2.268	.025
	DRINF	.000	.000	-.122	-2.140	.034

10	(Constant)	15.406	1.858		8.292	.000
	DESWC	.007	.001	.631	10.656	.000
	WRDFRQc	-.004	.001	-.309	-5.597	.000
	PCCONNz	.000	.000	.256	4.625	.000
	WRDPRO	.000	.000	-.137	-2.465	.015
	SYNMEDwrd	.003	.001	.253	4.637	.000
	CNCTemp	.000	.000	-.123	-2.266	.025
	DRINF	.000	.000	-.123	-2.194	.030
	CRFNO1	-.001	.000	-.115	-2.162	.032
11	(Constant)	17.548	2.133		8.229	.000
	DESWC	.007	.001	.647	10.940	.000
	WRDFRQc	-.005	.001	-.397	-5.634	.000
	PCCONNz	.000	.000	.248	4.519	.000
	WRDPRO	.000	.000	-.136	-2.471	.015
	SYNMEDwrd	.003	.001	.227	4.096	.000
	CNCTemp	.000	.000	-.141	-2.581	.011
	DRINF	.000	.000	-.132	-2.358	.020
	CRFNO1	-.001	.000	-.111	-2.115	.036
	RDL2	4.219E-5	.000	.137	1.981	.050
12	(Constant)	18.342	2.142		8.564	.000
	DESWC	.007	.001	.648	11.095	.000
	WRDFRQc	-.006	.001	-.424	-5.983	.000
	PCCONNz	.000	.000	.253	4.670	.000
	WRDPRO	.000	.000	-.135	-2.475	.015
	SYNMEDwrd	.003	.001	.247	4.437	.000
	CNCTemp	.000	.000	-.127	-2.345	.021
	DRINF	.000	.000	-.138	-2.497	.014
	CRFNO1	.000	.000	-.085	-1.592	.114
	RDL2	5.446E-5	.000	.176	2.490	.014
	LSASS1d	-.003	.001	-.115	-2.062	.041
13	(Constant)	18.670	2.144		8.707	.000
	DESWC	.007	.001	.628	10.946	.000
	WRDFRQc	-.006	.001	-.435	-6.135	.000
	PCCONNz	.000	.000	.255	4.673	.000
	WRDPRO	.000	.000	-.138	-2.522	.013
	SYNMEDwrd	.003	.001	.245	4.388	.000
	CNCTemp	.000	.000	-.127	-2.318	.022
	DRINF	.000	.000	-.138	-2.490	.014
	RDL2	5.783E-5	.000	.187	2.642	.009
	LSASS1d	-.004	.001	-.137	-2.496	.014

a. Dependent Variable: Score