The Effect of Specific Language Features on the Complexity of Systems for Automated Essay Scoring

Yoav Cohen, Anat Ben-Simon & Myra Hovav
National Institute for Testing & Evaluation

Paper Presented at the IAEA 29th Annual Conference
Manchester, UK, October, 2003

The Effect of Specific Language Features on the Complexity of Systems for Automated Essay Scoring[1]

Yoav Cohen, Anat Ben-Simon & Myra Hovav[2]
National Institute for Testing & Evaluation, Israel

Automated essay scoring (AES) can be a reliable and efficient assessment procedure. AES is currently performed using three types of methods: those based on analysis of surface features of the text, those based on analysis of semantic space, and those based on natural language processing (NLP).

Each type of method is sensitive, to a certain extent, to specific language features that tend to vary widely across languages. The current paper examines the effect that such differences may have on the complexity of AES systems developed to grade essays in a specific language. The above analysis is performed with respect to the various writing scales customarily used to assess writing products.

Automated scoring of essay items (AES) offers numerous advantages in the fields of assessment and instruction, among them objectivity, standardization and efficiency (quick, on-line, and inexpensive scoring). Each of these advantages contributes to different aspects of the assessment process.

Computer-generated scores can be used in several ways in a given assessment process: (1) as a sole measure of writing ability; (2) as a joint measure along with human ratings; namely weighted with human ratings in order to produce an average score; and (3) as a monitoring or quality control procedure over the human rating process. Though there are psychometric advantages and disadvantages to the different approaches in all cases, the use of computer-generated scores contributes significantly to the standardization and accuracy of the test scores. This is always a highly desirable feature, particularly in high-stakes testing.

In addition to the psychometric advantages, computerized scoring can be considerably faster than human scoring and the operational costs markedly lower. These aspects are particularly important in large scale assessments, which often involve up to a hundred thousand examinees at a given administration (e.g., NAEP). This particular feature of AES programs has the potential to enable the delivery of essay tasks efficiently and routinely in large scale

[1] The authors extend their thanks to Janine Woolfson and Randy Bennett for their helpful comments on earlier versions of the paper.
[2] Paper presented at the 29th Annual Conference, Manchester, UK, October, 2003.

testing operations, and thus encourages the inclusion of such tasks in large scale assessments (Bennett, 1999).

Some AES programs offer immediate (on-line) scoring, which may be of major importance in situations where immediate decisions are needed. Other programs offer detailed on-line feedback which highlights the particular strengths and weaknesses of the writer as they are reflected in a given writing product (i.e., My Access/Vantage Learning, Criterion/ETS). Occasionally the programs also provide suggestions for correction. This last feature is highly appreciated in instructional settings and thus often incorporated in instructional writing programs. It should, however, be noted that with few exceptions this application of AES is restricted to the use of pre-determined and pre-tested writing prompts and cannot be applied to any writing prompt.

The salient advantages of AES notwithstanding, certain issues are often raised with regard to its application and development. The first issue relates to the validity of the particular scoring procedure or, more specifically, to the consequential validity or washback effect that it might have on the instruction of writing skills and the practice of writing. Educators argue that essays foster communication between people, whereas essays written for computer grading will elicit a different type of writing. Moreover, once the scoring rules are made public, students will manipulate their texts to fit these rules while neglecting other, perhaps more important, principles of good writing. Critics also claim that the process of grading essays helps teachers connect with their students and get to know them better. This by-product of human grading will be lost in automated scoring. Furthermore, critics express concern that automated grading, because it is based on prototypical essays, will discriminate against students who have unique or idiosyncratic writing styles.

Though automated scoring of essays is far cheaper than human scoring, it is not cost free. The development of new software entails extensive investment of both time and money. Most of the programs currently available took as long as ten years to develop and were highly costly endeavors. The application of available programs often involves a substantial initial investment because the programs need to be trained to score every new writing task (prompt). Such training may involve the calibration of scoring scales or identification of an optimal set of predictors.   In other cases, where the assessment is administered in the P&P mode, additional effort and cost is involved in converting the writing samples into electronic form in order to facilitate computerized scoring.

The third issue relates to the feasibility of using AES across languages, or more particularly to the generalizability and applicability of the various AES methods and algorithms available so far, to languages other than English.

The current paper focuses primarily on the last two issues. It examines the relationship between different aspects of the linguistic structure of a given language and the complexity of the computer program, whether existing or prospective, that is to be used for the scoring of essays in that language.

The first part of the paper discusses common scales used to assess writing products, then briefly describes various methods of AES and reviews several AES programs currently in use. It also presents empirical results attesting to the reliability and validity of these programs, principally with regard to essays written in English.

The second part of the paper presents various linguistic features that may vary extensively across languages and examines the ramifications of these features on the complexity of the AES operational system. This analysis is presented chiefly with respect to Hebrew and English, which are used to illustrate the differences that may exist between languages.

Scoring dimensions for writing

Most AES procedures were originally developed for the purpose of assessing writing skills. In light of this objective they try to mimic human readers as closely as possible. The modeling of human readers is achieved in two ways: (1) Text features believed to be close estimates of various writing characteristics are defined and used to estimate the quality of writing ; and (2) human reader ratings are used (almost exclusively) to validate the performance of computer-generated scores.

Though few AES procedures were developed with a keen focus on specific writing characteristics, most AES procedures attempt to produce scores that at least correspond to common writing dimensions.

Table 1 presents five scoring dimensions commonly used in the assessment of writing tasks. It also presents the writing features associated with each dimension. Each of the writing features can be further broken down into more specific features that may later be translated into quantifiable measures. For example, "register" can be assessed by the average frequency of the essay's content words in the language; the lower the frequency the higher the register. Syntactical complexity can be indicated by average sentence length or by the average number of clauses in the essay's sentences.

Methodical analysis of the quality of a given AES procedure and its applicability to a given writing assessment should involve comprehensive and systematic analysis of the particular text features that are extracted from an essay and used to assess its quality, and their relevance and suitability to a given writing task and its scoring dimensions. Also, in addition to the reliability evaluations performed by comparing computer-reader agreement with inter-rater agreement, a construct validity analysis should be performed in order to assess the extent to which the interrelationship between the scoring dimensions obtained for human-rater scores is reconstructed for computer-generated scores.

These practices are particularly important in light of the fact that most AES procedures are based on brute empirical approaches and run the risk of using partially irrelevant text features to assess specific writing skills. Such a scenario is highly probable due to the moderate to high correlation between the different writing scales often observed in human ratings.

A deep understanding of the specific text features used by a given AES procedure may also prove to be highly relevant to its feasibility for application or adaptation to languages other than English, which may differ from the English language in a wide range of linguistic features.

Table 1: Scoring scales for writing

Content	Rhetorical Structure / Organization	Style	Vocabulary	Syntax & Grammar/ Mechanics	Creativity
Relevance	Organization	Clarity	Richness	Complexity	
Richness of ideas	Coherence	Fluency	Register	Syntactical accuracy	
Originality	Cohesion	Accuracy	Accuracy	Grammatical accuracy	
	Paragraphing			Spelling	
	Focus				

Methods of automated essay scoring

Chung & O'Neil (1997) classify the approaches to text analysis in two categories: methods designed to perform classification of documents, and methods that attempt to understand the meaning of a text. This classification can easily be applied to AES procedures.

The first category includes methods that are based predominantly on analysis of superficial features of a text (surface variables), such as average sentence length, number of paragraphs and the average and SD of word length. In the extreme case, these methods are completely language-blind, meaning they have no "knowledge" whatsoever of the particular language in which the text is written and can thus be applied to almost any given language. In other cases, some features of the language are incorporated into the scoring procedure (e.g., a list of prepositions, a table of word frequency or a list of specific key words). These features can be predetermined and applied to all writing samples in a given language, or else extracted from a sample of texts initially matched to the writing topic.

The second category is comprised of methods that are capable, to some degree, of interpreting a text. These methods are based predominantly on Natural Language Processing (NLP) techniques, which can extract meaning from a given text with varying degrees of success. In other words, they can perform semantic, morphological and syntactical analysis of a text. These techniques require metalinguistic knowledge of the language in which the text is written, including knowledge of the semantic and morpho-syntactical rules such as inflection and derivation, identification of speech parts and sentence structure.

The various AES methods differ principally in the type of text features that they extract from the text and use for scoring, and the statistical procedures they use to determine the weight of these features and combine them into one or more scores. Once the text features are extracted from a text, classical or modern statistical approaches such as factor analysis, discriminant analysis, linear and non-linear regression or neural network analysis are applied to identify the best predictors among these features and determine their optimal weight.

All the methods, with a few exceptions, require training samples to calibrate the system. Such training needs to be carried out separately for each writing task (prompt) and for each designated population. Once the training is completed, scoring development is carried out in a fairly similar way. In the training stage human readers, preferably more than one, grade a sample of 200-300 essays. An AES program applied to the calibration sample extracts 30-100 text features, which are pre-determined by each particular program, and assigns a proximity value (e.g., regression weights) to each feature. The scoring procedure is then applied to a second sample in order to produce cross-validation estimates. Once training is satisfactorily completed, new essays can be automatically scored by extracting the relevant text features and applying the appropriate weights to generate a final score or scores.

The quality of the training process for any given AES program depends on a variety of factors, including the number of sample essays used, the number of points on the scoring scale, the variability of scores across the sample essays and the number of human-rater scores upon which the criteria are based. The quality of the automated scoring process can be also affected by variables such as the average length of the essays, the genre of the texts, and the sincerity of the writers.

As many as six different programs are currently known to be in use for automated essay scoring. A brief review of each follows, including the scoring procedure used by each program, the scoring dimensions used to report scores, and their applicability to languages other than English.

Project Essay Grader (PEG)

PEG was the first computer program developed for essay scoring. The first version of the program was developed by Ellis Page in 1966 (Page, 1966). The first version of Page's program used a regression model in which the independent variables were mostly surface features of the text. The independent variables, termed "proxy" variables by Page, served as estimates of some intrinsic quality of writing style. The first version of PEG used 28 proxy variables including features such as: title, average sentence length, number of paragraphs, and number of various text characters, such as parentheses, commas, and periods. Most of the proxy variables used in this version of PEG are language-blind yet some features – such as the number of prepositions and connectives, spelling errors and common words – are language dependent to the extent that they need to be fed into the program for every new language to which the program is applied. The statistical procedure used by PEG to produce weights for the proxy variable is simple multiple regression.

A revised and perhaps more sophisticated version of the program was released in the 90s and is known to make use of some NLP tools such as grammar parsers and speech part taggers

(Page & Peterson, 1995; Page, 1995). The current version of PEG was released as a commercial product and as a result hardly any information is available regarding the proxy variables and the exact procedures used to generate essay scores.

PEG was successfully applied to three data sets: NAEP essays (Page, 1994; Page, Poggio & Keith, 1997), PRAXIS/ETS essays and GRE/ETS essays (Page & Petersen, 1995; Petersen, 1997). In all three studies, the correlation between PEG scores and reader scores was equivalent to the correlation between reader scores. In some cases PEG demonstrated an even higher rate of correspondence. In all three studies PEG produced a holistic score.

In a more recent study (Shermis, Koch , Page, Keith & Harrington, 2002) PEG provided five trait scores – content, organization, style, mechanics and creativity – as well as a holistic score. This innovation was introduced in order to provide more detailed information regarding the quality of the writing for purposes of formative feedback about areas of strengths and weaknesses. In this study PEG was applied to 386 web-based essays serving as a placement test in a Midwestern university. 807 sample essays were used for training. Results indicate that the correlations between PEG and human raters were significantly higher than inter-rater correlations for all five traits as well as for the holistic score.

IntelliMetric

IntelliMetric Engineer was developed by Vantage Technologies in 1997 for the purpose of scoring essays and open-ended responses. It is claimed to be based on an artificial intelligence approach. IntelliMetric was developed primarily as a commercial product and as a result hardly any information is available with regard to its scoring technique, apart from the following description:

> IntelliMetric™ is an intelligent scoring system that emulates the process carried out by human scorers. ….Relying on Vantage Learning’s proprietary CogniSearch™ and Quantum Reasoning™ Technologies, the IntelliMetric™ system internalizes the characteristics of the responses associated with each score point and applies this intelligence in subsequent scoring. The approach is consistent with that underlying holistic scoring.

Though the program is known to be widely used in a variety of settings – schools, higher education and other enterprises – evaluation results are rarely published.

The training procedure employed by IntelliMetric appears similar to the one employed by PEG. Human readers score a sample of calibration essays after which Intellimetric extracts some 100 content and structural features, identifies the optimal set of predictors and estimates their weights. To score a new essay, IntelliMetric applies these weights to those text features extracted from the new essay that match the predictors identified in the training samples.

Five scoring scales are routinely used to report results, yet IntelliMetric can generate scores for any set of scoring dimensions, providing that human-rater scores are available for the calibration sample on these dimensions. The scoring scales reported by Intellimetric are:

- *Focus and unity*: indicating cohesiveness and consistency in purpose and main idea.
- *Development and elaboration*: indicating breadth of content and support for concepts advanced.
- *Organization and structure*: indicating logic of discourse, including transitional fluidity and relationship among parts of the response.
- *Sentence structure*: indicating sentence complexity and variety.
- *Mechanics and conventions*: indicating conformance to English language rules.

Since it seems that the IntelliMetric's scoring procedure is based on a brute-empirical approach, it is not clear whether all text features are attempted for all scoring scales or else a specific subset of text features is applied to each scale.

Only a few validity studies of Intellimetric scores are available (Vantage Learning, 2001; 2002; Elliot, 2001). Results reported in these studies yield fairly satisfactory results. Of the above three studies, the one conducted by Elliot (2001) is the most comprehensive. In this study IntelliMetric was applied to 612 persuasive essays taken from national k-12 NRT data sets. In the external validity study, IntelliMetric was applied to 300 creative writing essays written by students aged 7, 11 and 14, and two external criteria were used to validate IntelliMetric scores, an MC writing test and teacher judgments of writing skills. In both studies the correlation between IntelliMetric and the human raters was as high as the inter-rater correlation. Correlations between IntelliMetric and external criteria were similar to those obtained between human raters and the same criteria.

Intelligent Essay Assessor (IEA)

IEA was developed by the University of Colorado in 1997, on the initiative of Thomas Landauer, Peter Foltz and Darrell Laham. Unlike PEG and Intellimetric, IEA adopts a completely different approach to essay scoring, which is deeply grounded in the Latent

Semantic Analysis method (Foltz, 1996; Landauer, Laham, Rehder & Schreiner, 1997; Landauer, Foltz & Laham, 1998). In order to understand how IEA operates, one needs to understand the principles of Latent Semantic Analysis.

The Latent Semantic Analysis (LSA) method was originally developed for information retrieval purposes and has been used for a variety of objectives other than essay scoring, such as cross-language information retrieval, information filtering and text analysis (Chung & O'Neil, 1997). The underlying assumption of the technique is that a latent semantic structure (semantic space) of a given set of documents or texts can be captured by a representative matrix that denotes the core meaning or content of these texts. The method is based on a factor-analytic model of word co-occurrences in which information generated from a variety of content-relevant texts is condensed and represented in a matrix that defines a "semantic space" and explicitly relates words and documents. In this procedure a primary matrix is developed in which each word considered for analysis is represented by a row and each text unit (e.g., sentence or paragraph) is represented by a column. The matrix cells contain the frequency of the appearance of a word in a given text unit. Next, using a mathematical scaling procedure – Singular Value Decomposition (SVD) – a subset of three smaller matrices is created, including a diagonal matrix which is later used to reconstruct the original matrix. The condensation of the primary matrix is achieved by setting notably small "singular values" in the diagonal matrix to zero and then multiplying all three matrices to reconstruct a word-document association matrix with fewer dimensions. The word-document association in this matrix is represented by a numerical value (weight) which is conceptually similar to variable loadings on a set of factors in a factor analysis procedure. The reconstruction process may result in slight changes in the associations between words and text units, including the creation of associations between documents that do not share even a single word (Landauer, Foltz & Laham, 1998; Miller, 2003).

With its unique scoring technique, LSA-based procedures can be applied to a broad variety of open-ended language assessments such as vocabulary tests and open-ended responses to reading comprehension tests.

In the particular context of essay scoring, the specific content of an essay is only important to the extent that it matches that of other essays of varying quality. As is true for all other programs, IEA uses a training sample for calibration. Texts relevant to the essay topic are sampled and used to create the word-by-document decomposed matrix. The sampling texts can be chosen from a variety of sources that best reflect the topic of writing, such as textbooks or an essay written by an expert in the case of criterion reference tests, or sample essays scored by human raters if scores are to be calibrated to a given reference group (Landauer, Laham & Foltz, 2000).

The scoring of a new essay can be done in several ways. Commonly, the essay is represented by k-dimension vectors (e.g., essay sentences), and each vector is then compared with all text units of the decomposed matrix. If sample essays are used for calibration, the score assigned to a new essay may be based on the average scores assigned to the N most closely matched pre-graded essays.

IEA usually provides scores for three dimensions:

- *Content* - assessed by two components derived from the decomposed vectors extracted from a given essay: (1) a 'quality' score equal to the weighted average of the scores for the k most similar calibration essays and; (2) a 'domain relevance' score based on the length of the essay’s vector.
- *Style* - assessed by two components: (1) 'coherence' - an LSA-based measure of conceptual relatedness among words, sentences, and paragraphs in an essay; and (2) 'grammaticality' - based on the resemblance between any given sentence’s grammatical structure to some standard that represents good writing.
- *Mechanics* - assessed through analysis of punctuation and spelling.

Miller (2003), in a review of AES procedures, concluded that LSA-based procedures are most effective in assessing the "content" dimension of essays. Although IEA provides scores on two other dimensions, "style", and "mechanics", given its underlying scoring method it seems safe to assume that the estimated scores for these dimensions are of a lesser quality.

IEA was successfully applied to GMAT/ETS essays (Landauer, Laham & Foltz, 2000). In this study, correlation between IEA and readers fell only slightly short of correlation between pairs of readers. Satisfactory results were also reported for the application of IEA to subject-dependent essays – history (Foltz (1996) and biology (Landauer et al, 1997).

With regard to validity checks, IEA employs various techniques to detect ‘Unusualness’, ‘Plagiarism’, English word order and content-score confidence.

One of the most intriguing features of LSA-based procedures is that in spite of the fact that they rely almost exclusively on the analysis of semantic relationships, they are in fact language-blind and can hence be applied to any language, irrespective of its alphabet. Nonetheless, their feasibility and effectiveness may depend to a large extent on the specific structure of the language, especially its morphological complexity.

E-rater

E-Rater was developed by the Educational Testing Service in 1997 under the leadership of Jill Burstein. The first version of the program was developed for the purpose of assessing essays written as part of the GMAT test and has been operational since 1999. Today, the program and its derivative application (Criterion) are applied in a wide variety of contexts that require the assessment of writing skills: schools, university selection and placement tests and tests administered by public as well as private organizations.

E-rater is predominantly based on Natural Language Processing techniques. Its prototype (Burstein, Braden-Harder, Chodorow, Hua, Kaplan, Kukich, Lu, Nolan, Rock & Wolf, 1998) was developed for the scoring of issue and argument-type essays. It uses hybrid feature methodology and incorporates both variables derived statistically via simple counting procedures and variables that were extracted using NLP techniques. A somewhat detailed description of the first version follows. Though e-rater has been extensively modified since the release of this version, far fewer details regarding the text features used by the subsequent versions are available. Familiarization with the first version can hence shed some light on the nature of the currently-used version.

The first version of the e-rater employed five sets of critical feature variables in order to build a final regression prediction model:

(a) *Surface feature variables* – such as essay length and various mathematical derivatives of this variable;

(b) *Structural features* - such as syntactic variety as reflected in the variability in clauses and verb types;

(c) *Rhetorical structure* – such as the number of occurrences of independent arguments in the final paragraph of the essay, number of occurrences of parallel words that begin an argument and number of occurrences of arguments starting with a summary phrase;

(d) *Content vector analysis* – this analysis is carried out by extracting the 100 most frequent words from a given essay and matching them to several lists of 100 words each. These lists contain the most frequent words extracted from a sample of essays assigned a specific score by human raters. The matching procedure can be further refined by restricting the frequency lists to content words only (excluding function words) and by applying morphological rules to identify various possible inflections of each word.

(e) *Content vector analysis by argument* – this analysis is similar to "content vector analysis" except for the fact that it is carried out on the sentence level rather than on the word level. The matching of sentences rather than words is believed to capture the association between words more effectively and hence be more likely to capture associations between ideas.

The current version of the e-rater extracts 50-60 essay features from each essay sample and uses only three dimensions to report essay scores:

Content – vocabulary related to the topic, such as relevant information and precise or specialized vocabulary.

Structure – syntactic variety, or the use of various structures in the arrangement of phrases, clauses, and sentences, such as the number of clauses of different types, the number of subjunctive modal verbs, and the ratio of syntactic structure types per essay and per sentence.

Organization – characteristics associated with the orderly presentation of ideas such as rhetorical features and linguistic cues (logical connections between sentences and clauses).

"Content" scores in the current version are obtained by applying a vector space model. This technique is similar in principle to the one used by LSA, except that instead of using SVD analysis, the original co-occurrence matrix is condensed. This is achieved by extracting only content words from the essays, leaving out various parts of speech (defined in a stop list) that are irrelevant to the specific topic of the essay. It should be noted that this method would not recognize similarity unless there was word overlap. The features in the other two dimensions are extracted through NLP techniques. NLP analysis is applied in order to parse the text sentences. The analysis returns a syntactically analyzed version of each sentence in a form of a vector that indicates the sentence and clause type, and the parts of speech and other morphological features (e.g., inflections) of each word. The parsed text is subjected to statistical analyses that generate specific text features indicative of various aspects of text structure and organization (Burstein & Marcu, 2000; Burstein, 2001). It should be noted that in the current version of the e-rater the role of surface features has been minimized in order to improve the validity of the scoring procedure.

Burstein and her colleagues also investigated the feasibility of using Neural Network techniques for automated essay scoring. Results, as reflected in the agreement between computer-generated scores and human ratings, were found to be promising (Burstein et al, 1998).

E-rater typically requires ~250 calibration essays per each new prompt. Human scores are regressed (step-wise) on the e-rater feature scores to produce weights for combining the scores. The weights are then cross-validated in a new sample. A different scoring model is usually created for each prompt. Only the most predictive features (8-12) are retained for any given prompt. To grade a new essay, e-rater extracts the relevant features and applies the regression weights to these features in order to compute a score.

E-rater has been successfully applied to a wide range of essay questions administered as part of tests like the GMAT (Burstein et al., 1998), GRE (Powers, Fowles & Welsh, 1999), and TWE (Burstein & Chodorow). In all of the above studies the procedure proved highly effective and accurate, yielding above 95% agreement with human raters for exact and adjacent scores. E-rater's scores were less related to external indicators of writing than were readers' scores. When combined with one human reader, validity appeared more comparable (Powers et al, 2000). (See Table 2 for detailed validity results).

A recent study (Powers, Burstein, Chodorow, Fowles & Kukich, 2001) investigated e-rater's susceptibility to manipulation. Two particular aspects were investigated:
(1) Do scores fail to capture important features relevant to good writing? and (2) Are scores unduly influenced by extraneous features? The above aspects were investigated by asking 27 participants to produce 63 essays according to specific instructions. Results indicated that e-rater could be tricked into awarding higher scores than deserved by means of strategies such as: repetition of paragraphs, overuse of key words and content-related words out of context or with faulty logic. The awarding of lower scores than were deserved was less frequent.

In light of the specific nature of e-rater, the procedure cannot be applied directly to other languages. Nevertheless, a detailed description of its underlying principles can be beneficial for the development of similar procedures in languages other than English.

Other AES procedures

Three other programs for automated essay scoring are known to exist: InQuizit, developed by InQuisit Technologies in 1998, the Text Categorization Technique (TCT), developed by Larkey (1998) and a third program developed by Rudner. Larky's program uses Bayesian independent classifiers to assign probabilities to documents, estimating their likelihood of matching a specific category of documents. The analysis relies on word-co-occurrence. Nearest neighbor technique is used to find the k closest essays. Computer-reader agreement, found for TCT, for issue essays and argumentative essays was above 95% for exact and adjacent scores. Little information is available with regard to the other two programs.

Summary of representative results from validity studies

Validation of Automated essay scoring systems can be performed in various ways. A common way to validate computer-generated scores is by comparing the correlation between computer-generated and human-rater scores to the correlation obtained between two human raters. In addition to this agreement index, correlation between computer-generated scores and external criteria are also compared with correlation between human rates and these criteria.
Table 2 summarizes representative validity results with respect to the agreement index.

Table 2: Representative results from studies of the correlation between computer and human reader scores are given in the following table:

Method	**Author/s**	**Essay type**	**Sample size**	**HH correlation**	**HC correlation**
PEG	Petersen 1997	Praxis	300	.65	.72
PEG	Petersen 1997	GRE	497	.75	.74-.75
PEG	Shermis et al. 2002	English placement test	386	.71	.83
Intellimetric	Elliot 2001	NRT	102	.84	.82
IEA	Landauer et al 1997	GMAT	188	.83	.80
IEA	Foltz et al 1999	GMAT	1363	.86-.87	.86
e-rater	Burstein et al 1998	GMAT (13 prompts)	500-1000 per prompt	.82-.89	.79-.87

These validation studies differ from each other with regard to essay type and inter-rater reliability (human). Nevertheless, in almost all of them, the computer-generated scores are correlated with the human scores at a level which is almost as high as the inter-rater reliability. (One exception is the PEG-Praxis study in which the computer scores appear to be more reliable than human scores. In this case, however, the inter-rater reliability is exceptionally low. The other exception is the PEG 2002 study, which differs from the other studies in the definition of the validation criterion (Shermis et al., 2002).

Application of AES procedures to different languages

All the AES programs described above were originally developed for essays written in English. Some of them can be easily applied to languages other than English without major modifications, while others require major adaptation. Moreover, each method type is somewhat sensitive to specific language features that tend to vary widely between different languages. Thus, any prospective application of a given AES program to other languages should involve thorough analysis of the specific analytic procedures it employs and their effectiveness when applied to languages with grammatical structures that differ significantly from those of the English language.

The second part of this paper examines the effect that such differences may have on the complexity of AES systems developed to grade essays in different languages. The case of the Hebrew language is used to illustrate potential structural and grammatical differences that may exist between languages. This examination is carried out with respect to the various text features used by different AES procedures to generate dimension scores for writing products.

Both statistical and NLP-based approaches seem to yield fairly similar results with regard to agreement between computer and human judgments, and between computer scores and external measures of writing. However, NLP based methods have a notable advantage with regard to consequential validity.

Text features used to automatically score essay items can be classified into five categories. These categories are listed here in ascending order, according to the degree of effort required to adapt them to any given languages:

1. *Surface variables* – all text features in this category, such as essay length, number of sentences or paragraphs, and average word or sentence length, are completely language-blind and can thus be applied by any program to any given language with no further effort or investment.
2. *Variables based on stop lists* – some text features can be extracted from short, easily generated "stop lists". Text features in this category include variables such as: the number of prepositions or connectives, total occurrence of arguments starting with summary words or using parallel words, and total occurrence of pronoun references (him, theirs, etc.).
3. *Lexicon-dependent variables* – certain text features, such as erroneous spelling, require the availability of a complete digital lexicon and quick, efficient, scanning or search procedures.
4. *Corpus-dependent variables* – some aspects of writing skills, particularly those associated with the vocabulary dimension (e.g., richness, register and accuracy), are

based on the average frequency in the language of the words the essay contains. These features require the existence of a corpus in the given language. Moreover, in languages which have complex morphology, an effective measure of word frequency will also require use of a morphological analyzer that can extract lexemes from inflected words.

5. *Variables requiring natural language processing* – many text features, particularly those used to assess structure and organization dimensions, require the use of NLP procedures that can parse sentences and tag speech parts. Procedures such as these require a considerable amount of effort to develop.

The following is a list of language characteristics that tend to vary to a large extent across languages. These characteristics may have a considerable effect on the complexity of AES procedures developed for automated scoring of essay items in a given language.

- *Lexicon size* – number of lexemes (primary entries in a dictionary).
- *Prevalence of inflexions* (e.g., person, gender, number, possessive) – high prevalence of inflected words (e.g., nouns, adjectives and verbs) require a use of morphological analysis to parse all words into their grammatical components.
- *Prevalence of prefixes* (e.g., connectives/conjunctives, pronouns and articles) and suffixes (e.g., gerund, accusative case). High prevalence of prefixes and suffixes in a language may pose difficulties in extracting lexemes from a given letter string.
- *Homograph's rate* – homographs are words which have identical orthography yet different meaning, for example the word *orange* can denote a fruit (noun) or a color (adjective). A high rate of homographs in a language presents a challenge to both LSA-based procedures and NLP procedures, which use the identification of parts of speech to perform syntactical analyses.
- *Flexibility of syntactic structures* – some languages are very flexible with regard to sentence structure, offering many valid ways to construct a sentence that denotes an identical meaning. This feature contributes significantly to the complexity of the syntactical analysis (sentence parsing) of the text.

Language Specific Features – The Case of Hebrew

Although simple surface variables, such as word length, essay length, and the relative frequency with which one character or another appears in the text, are extremely predictive of essay score, they can clearly not serve as the sole basis for an automatic essay scoring system. The consequences of using a surface-based AES alone are that preparing students to the test will become a matter of teaching them to write longer texts containing longer words, with no regard for rhetorical structure, the logic of argumentation, and so forth. Surface variables can only be used alongside more substantive characteristics of the text, such as content and syntactic and rhetorical structure, not to mention more elusive concepts such as literary or aesthetic value.

In order to evaluate content and rhetorical structure we need to perform a syntactic analysis of the text. This is where the characteristic features of the language in question become relevant. Hebrew, along with other Semitic languages (Akkadian, Arabic, Aramaic, Ethiopic, etc) has orthographic, morphological and syntactic features that might affect the complexity of an automated system for scoring essays written in Hebrew.

Hebrew orthography

The letters in a written Hebrew word are usually consonants. Very few of the vowels are represented orthographically. Hence, a string of Hebrew letters can usually be pronounced in several ways. This is known as orthographic ambiguity. Moreover, some letters can represent different consonants. The letter ש (“Shin”), for instance, can represent both “sh” and “s” sounds; the letter פ (“peh”) can be pronounced both as “f” and as “p”, and the letter ב (“bet”) can be pronounced “b” or as “v”. The letter ו (“vav”) in some instances represents a vowel (“o” or “u”), and in others a consonant (“v”). The various pronunciations generally constitute different interpretations of the word. At the end of the first millennium CE the orthographic ambiguity of Hebrew gave rise to the development of a method of diacritical marking known as “vocalization” (the Tiberian vocalization method) which removes all ambiguity from Hebrew orthography. The method involves adding diacritic marks either above, inside, or beneath the letters. The diacritic marks indicate the correct pronunciation of the word either by representing vowels or by specifying the correct sound of the consonants. Although children learn to read and write vocalized Hebrew, contemporary Hebrew texts intended for adult native speakers are not vocalized. (Poetry is an exception – it is usually written in vocalized Hebrew.) Thus, most of the printed words in unvocalized Hebrew texts are ambiguous in the sense that they can be read in different ways. To make matters even more complicated, Hebrew writing has undergone additional changes in recent decades. To make the reading of unvocalized texts easier, it is now permissible to write various words with extra letters instead of diacritical marks (this is known as “full orthography”). Today, Hebrew texts can be written in three ways: vocalized, unvocalized, or using full orthography, but any text nay include a mixture of orthography systems. It is considered good practice to follow the rules of full orthography, but these rules are not common knowledge of writers, even experienced ones, so that it is acceptable to use a free mixture of orthography styles.

The following table demonstrates the various ways in which the word ספר (“SFR”) can be written and pronounced, and the meaning associated with each variant.

Table 3: Seven vocalized forms of the word ספר “SFR”.

Vocalized	**Unvocalized**	**Full orthog.**	**Pronunciation**	**Meaning**
סֵפֶר	ספר	ספר	SEFER	book
סְפָר	"	"	SFAR	border
סַפָּר	"	"	SAPPAR	barber
סָפַר	"	"	SAFAR	counted
סַפֵּר	"	"	SAPPER	tell / cut hair (imperative)
סִפֵּר	"	סיפר	SIPPER	told/ cut (past) hair
סֻפַּר	"	סופר	SUPPAR	was told/ his hair was cut

In this case, a single unvocalized form represent seven different words. Such morphological ambiguity is by no means exceptional in Hebrew. Itai and Segal (2002) estimate that more than half the words in Modern Hebrew are morphologically ambiguous. More than 10% of the words have four or more different vocalized forms and approximately 1% of the words have 7 or more different vocalized forms.

Hebrew Morphology

The verb system

Hebrew verbs, like those of other Semitic languages, are root based. The root is usually comprised of three letters and can be conjugated according to one of several templates (known as "Binyanim"). As in romance languages, a conjugated verb form conveys the tense, the person, the number, the voicing and the gender in a single word. Thus, the word "A-HAV-TI" which means "I loved" is based on the three-letter root A-H-V and is understood to be in the past tense and first-person singular. Some (but not all) conjugations of the root A-H-V are presented in the following table.

Table 4**:** Some conjugations of the verb "AHV" (to love).

Hebrew	**Transliteration**	**Meaning**	**Tense**	**Person**	**Number**	**Gender**
אהב	AHAV	he loved	past	III	singular	masculine
אהבה	AHAVA	she loved	past	III	singular	feminine
נאהב	NOHAV	we will love	future	I	plural	m/f
אוהבות	O'HAVOT	they love	present	I II III	plural	feminine
תאהבו	TO'HAVU	you will love	future	II	plural	m/f
נאהבת	NE'EHEVET	she is being loved	present	III	singular	feminine
תתאהב	TIT'AHEV	You/she will fall in love	future	II or III	singular	masculine (II) feminine (III)

Typically, each verb can be conjugated according to three or four templates. Within each template, each root can be conjugated in four tenses, for first, second or third person, in masculine or feminine, and singular or plural form. On average, each verb can be conjugated in about 25 different forms within a single template thus, the average number of conjugated forms per root is about 100.

In addition to the rich inflectional and derivational system of verbs, several grammatical markers can be attached to the verb as prefixes (e.g., “ha” for a verb in question form; “shĕ” for ‘that’ and “vĕ” for ‘and’). Thus, the word “SHE’AHAVTI” denotes ‘whom I loved’. In classical Hebrew (less so in modern spoken Hebrew) conjugated transitive verbs can appear with suffixes used as accusative markers. Thus, for instance, “AHAVTI” means ‘I loved’, but with the suffix “V” it means ‘I loved him’ (“AHAVTIV”). These suffixes convey the gender, person and number of the verb compliment.

Hebrew has about 3,000 verb roots, so the number of conjugated verbs can reach 300,000, and with various prefixes and suffixes, the number of different (vocalized) verb-related words can exceed 3,000,000. Only a fraction of these forms are used in modern Hebrew, but any AES system would need to have the capacity to recognize and analyze all of them.

The noun system

Nouns can also be inflected in various ways, although the system is not as rich as it is for verbs. Nouns are either masculine or feminine, with typical morphological features that differentiate between them. All nouns have different forms depending on number: singular, dual or plural form (SEFER = ‘book’, SFARIM = ‘books’; YAD = ‘hand’, “YADAYIM” = ‘pair of hands’). Nouns can appear in their free form or in a “bound” genitive form used in conjunction with possessive suffixes. Thus SEFER = ‘book’, but SIFRI is ‘my book’ and SIFREIHEM = ‘their books’.

Nouns can take various grammatical markers that act like prepositions, pronouns, or the definite article, as prefixes. These include: “ha” for ‘the’, “bĕ” for ‘in’ or ‘at’, “lĕ” for ‘to’, “shĕ” = ‘that’, “vĕ” (or “û”) = ‘and’, “kshĕ” or “lixshĕ” for ‘when’ “mee” = ‘from’, “kĕ” = ‘as’ or ‘similar to’; and combinations of these markers: “vĕ-ha” = ‘and the’, “û-kshĕ” = ‘and when’, etc. Proper nouns can also take these markers (excluding the marker for the definite article) as prefixes.

There are between 10,000 and 20,000 common nouns in Hebrew. However, including the possibilities for genitive inflections and prepositional prefixes, the number of different words that refer to nouns most probably exceeds two million.

The size of the lexicon and the number of words

Hebrew is not a rich language; there are a little over 20,000 lexemes – nouns and verbs in their uninflected and non-prefixed form. The relative paucity of Hebrew is exemplified by the size of a typical dictionary, editorial considerations notwithstanding. The commonly used dictionary in Hebrew, which is also the largest, is the Even-Shushan dictionary. The Even-Shushan can be regarded as the Hebrew equivalent of Webster's New World Collegial Dictionary. According to its editors, the Even-Shushan dictionary contains about 40,000 primary entries (and 70,000 primary and secondary entries). The Webster's dictionary is twice the size: containing about 80,000 primary entries and 140,000 primary and secondary entries. The English lexicon, as indicated by the size of a typical dictionary is much larger than the Hebrew lexicon. However, in terms of different words – that is, the number of different strings of characters that can appear in a text – English is probably much smaller than Hebrew. The number of different words (so called word *types*) which can appear in a text is the subject of studies conducted by J. B. Carroll (1967, 1971). Carroll (1967) found 50,406 word types in the Brown University Corpus, which at that time consisted of about one million words. Of course, as the corpus gets bigger, an increasing number of rare words appear in it. According to Carroll’s statistical analysis, the number of word-types that a 10 million word corpus would be expected to yield would be about 120,000. A 100 million word corpus would consist of 200,000 word types and in an infinitely large corpus, the number of word types would reach 340,000. Included in this number are proper nouns and strings of numerical digits and symbols. In a later work, Carroll (1971) almost doubled this estimate, claiming that there are 600,000 word types in English texts compared with our estimate of 5,000,000 word types in Hebrew. Hence, although English is much richer than Hebrew in terms of the number of lexemes, the number of word types that must be recognized by Hebrew language analysts is about an order of magnitude larger.

Hebrew Syntax

Hebrew word order is less restrictive than English word order. The simple SVO structure is prevalent, but almost any other order is considered grammatically correct. Different ordering of the subject, verb, and object convey different nuances of meaning. Thus, for instance, the simple sentence: “I-shall-give-him-the-book”, which in Hebrew comprises three words, can also be phrased as “The-book I-shall-give him” and as “[to] him I-shall-give the-book” or “The-book [to] him I-shall-give”. It is understood by the reader or the listener that the first or second element of the sentence is the intended focus.

Some experimental results concerning AES in Hebrew

To the best of our knowledge we, in cooperation with Vantage Learning (2001), have performed the first trial of automatic scoring of Hebrew essays. We used Intellimetric software to score essays written in Hebrew by examinees being tested in Hebrew as a foreign language. The sentences are short; their average length is about 12 words. The essays are scored on four scales: Content, Rhetorical Structure, Vocabulary and Language mechanics. Two readers, working independently, grade each essay; the grading scales range from 1 to 7 and a total score is calculated as the simple (non-weighted) sum of the four scales. The Intellimetric software system was first applied to 50 essays and then tested on a group of 194 different essays. The results are presented in the following table. First, we present the correlation coefficients between the two human readers (H1-H2), then we present the correlation between the automated scoring and one of the readers (C-H1/C-H2) and lastly, the correlation between the computer score and the score calculated on the basis of the two human readers (C-(H1+H2)). In each case, five correlation coefficients are presented: one for each of the four scales and one for the total score.

Table 5: Hebrew essays – automated and human scoring.

	Content	**Structure**	**Vocabulary**	**Language**	**Total score**
H1-H2	.89	.88	.91	.90	.94
C-H1/C-H2	.73-.76	.77-.80	.79	.74-.76	.82-.84
C-(H1+H2)	.77	.81	.81	.77	.84

It is evident that the grading of the essays by human readers was extremely reliable. The correlation between the total scores given by two readers is .94. The correlation with the automatic score is significantly lower (.82-. 84) but is within the range of agreement coefficients obtained for essays written in English (.72-. 87). The fact that a computer program developed in order to score essays in English could score Hebrew essays so successfully prompted us to investigate how successful a simple regression-based scoring system would be.

We set out to determine the zero-order correlations between the surface variables of the texts and the scores given by human raters. This study, like the one reported above, was carried out using essays that were written in the course of a test of Hebrew as a foreign language. We analyzed 100 essays that had already been scored by two human readers.

The results were unequivocal; as was the case in English, the most predictive surface features of the texts are associated with their length. Thus, for example, the number of characters per essay correlates as high as 0.80 with the human score, the number of words per text has a correlation of 0.75 and the number of sentences per text has a correlation of 0.41 with the score given by human readers. These three variables are of course highly inter-correlated, and some may hence be redundant in a multiple regression equation. There is, however, another variable which is also predictive of the human score (r=0.61), although less correlated with the other variables, and this is the variability (standard deviation) of sentence length. The average word length in the text is also correlated (r=0.51) with the scores, probably because language proficiency manifests itself in a larger vocabulary that includes words that are less frequent and thus usually longer (a manifestation of the Zipf law). In addition to these variables we also discovered that the relative proportion of specific characters in the texts is correlated with the scores of the essays. For example, the relative occurrence of the letter "vav", which, you may recall is used in Hebrew to signify conjunction ("and"), has a correlation of 0.52 with the essay score.

Following the zero-order correlation analysis, we proceeded to find linear combinations of predictors. In light of the fact that the number of characters is correlated 0.80 with the scores, it is not surprising that a combination of this variable with the sentence length variability yields a correlation of 0.83 with the scores. A three-variable regression equation, in which the relative occurrence of "vav" is added, yields a correlation of 0.84, and a five variable equation yields a correlation of 0.89.

Discussion and conclusions

Although, as described above, Hebrew differs from English in key grammatical features, we discovered that an AES system originally developed for English performs quite well on Hebrew texts. (The authors of the system claim that it was tried out on texts in other languages and performed quite well, see Vantage Learning 2002). Even extremely simple equations which weigh straightforward surface variables perform quite well in Hebrew. How then, if at all, do the special features of Hebrew influence the process?

We classified text features into five categories (cf. section 4). The first category included surface variables of the texts. Indeed, extracting the surface features of the texts seems to be possible irrespective of the language. It is likely to transpire that the same group of surface features are related to essay scores in many languages; at least those which are not pictographic but are written using an alphabet.

The second class of text features was based on word-lists. A component of a scoring system that is based on these features is of course language dependent, but is easily adapted to other languages. This is true as long as these features (connectives, pronoun references, etc.) retain their status as distinct words in the target language and are not represented by prefixes etc. If that is the case, then the adaptation to the target language may not be straightforward.

The third class is that of lexicon based features. Here, the component of the AES system is strongly coupled with the language. Naturally, lexicon based features assume the availability of a digital lexicon and a collection of word types, which as we have shown may be an order of magnitude larger than the lexicon.

The fourth class includes features that can be only extracted when a well structured corpus is available. In order to estimate the relative frequency of words or lexemes, a sizable collection of samples of the written language is needed. As Carroll (1968, 1971) has shown, in order to represent a sizable fraction of the vocabulary, the collection needs to comprise at least 10 million words. This, of course, is a weighty undertaking in any language. Furthermore, it is somewhat less complicated in English than in Semitic, Romance or other languages, which are rich in inflections.

The final category is comprised of features that depend on syntactic parsing. Although the principles of parsing may be similar in different languages, actual parsing is highly dependent on the lexicon and requires part-of-speech (POS) tagging of the texts. POS tagging is the assignment of the correct part of speech for each word in the sentence. Thus, it is equivalent to finding the most probable sequence of POSes out of all the possible ones. In languages, such as Hebrew, in which there is high proportion of homographs, the number of ways in which the sentence can be parsed grows very fast with the length of the sentence. If, for instance, half of the words in the sentence can be interpreted in two ways or more, a 12 word sentence can be parsed in over 64 ways. On the other hand, it may turn out that in languages which are “highly inflectional”, sentences are shorter and each word conveys more about its relation to other words in the sentence. Further study in what might be termed “comparative NLP” will furnish the answer.

The path to a working system of AES in languages other than English seems quite arduous, but given the head-start in English and the amazing tools of modern NLP, it seems a realistic and achievable goal.

References

Bennett, R. E. (1999) Using New Technologies to Improve Assessment. *Educational Measurement: issues and practice, Vol. 18(3)* 5-12.

Burstein J., Braden-Harder L., Chondrow M., Hua S., Kaplan B., Kukich K., Nolan G., Rock D. ,Lu C. & Wolff S. (1998) *Computer Analysis of Essay Content For Automated Score Prediction.* ETS Report (820).

Burstein, J. & Chondrow, M. (1999). *Automatic Scoring for Nonnative English Speaker*. Joint Symposium of the of the Association of Computational Linguistics and the international Association of Computer Language Learning Technologies , Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Languages Proceeding, College Park ,Maryland.

Burstein, J. & Marcu, D. (2000). *Benefits of Modularity in an Automated Scoring System*. Proceedings of the Workshop on Using Toolsets and Architectures to Build NLP Systems , 18th Intentional Conference on Computational Linguistics , Luxembourg, August 2000.

Burstein, J. (2001). *Automated essay evaluation with natural language processing.* Paper presented at the NCME Annual meeting, Seattle, Washington.

Carroll, J. B. (1967). On Sampling from a lognormal model of word-frequency distribution. In Kucera, H., & Francis, W. N., *Computational Analysis of Present-Day American English.* Providence, RI: Brown University Press.

Carroll, J. B. (1971). The Development of the Corpus. In: Carroll, J.B., Davies, P., Richman, B., *The American Heritage, word frequency book*. New York, NY: American Heritage Publishing Inc.

Chung, G. K., & O'Neill, H. F. (1997). *Methodological approaches to online scoring of essays* (CSE Tech. Rep. No. 461). Los Angeles: UCLA, National Center for Research on Evaluation, Student Standards, and Testing.

Elliot S. (2001) *IntelliMetric*: *From Here to Validity*. Paper presented at the annual meeting of the AERA, Seattle, Washington.

E-rater [computer software]. (1997). Princeton, NJ: Educational Testing Service.

Foltz, P. W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, and Computers, 28(2)*, 197-202.

Foltz, P. W., Kintsch, W.,& Landauer, T. K. (1998). The measurement of textual Coherence with Latent Semantic Analysis. *Discourse Processes, 25,* 285-307.

InQuizit [computer software]. (1998). Santa Monica, CA: InQuizit Technologies.

Intelligent Essay Assessor [computer software]. (1997). Boulder, CO: University of Colorado.

Intellimetric Engineer [computer software] (1997). Yardley, PA: Vantage Technologies.

Itai, A. & Segal, E. (2002) *A Corpus Based Morphological Analyzer for Unvocalized Modern Hebrew*. Unpublished report, Department of Computer Science, Haifa, Israel.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes, 25*, 259-284.

Landauer, T. K., & Laham, D. & Foltz, P. W. (2000). *The intelligent Essay Assessor: Putting Knowledge to the Test.* Knowledge Analysis Technologies.

Larkey, L. S. (1998, August). *Automatic essay grading using text categorization Techniques.* Paper presented at 21st International conference of the Association for computing Machinery-Special Interest Group on Information Retrieval (ACM-SIGIR), Melbourne, Australia.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 48*, 238-243.

Page, E. B. (1994) Computer grading of students scores, using modern concepts and software. *Journal of Experimental Education, 62*, 127-142.

Page, E. B., Poggio, J. P. & Keith, T. Z. (1997) *Computer analysis of student essays: Finding trait differences in the student profile*. Paper presented at the AERA Annual Meeting, Chicago.

Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan, 76*, 561-565.

Peterson, N. S. (1997) *Automated scoring of written essays: Can such scores be valid?* Paper presented at the NCME Annual Meeting, Chicago.

Powers, D. E., Fowles, M. E. & Welsh, C. K. (1999) *Further validation of a writing assessment for graduate admissions*. GRE Board Research Report No. 96-13R and ETS Research Report No. 99-18.

Powers, D. E., Burstein, J., Chodorow, M. Fowles, M. E. & Kukich, K. (2001) *Stumping E-Rater: Challenging the validity of automated essay scoring*. GRE Board Professional Report No. 98-08bP and ETS Research Report No. 01-03.

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z. & Harrington, S. (2002). Trait Rating for Automated Essay Scoring. *Educational and Psychological Measures, 62,* 5-18.

Miller, T. Essay assessment with latent semantic analysis. *Journal of Educational computing Research,* 28(3), 2003.

Vantage Learning (2001). *A Preliminary Study of the Efficacy of IntelliMetric for Use in Scoring Hebrew Assessments*. RB-561.

Vantage Learning (2002). *A Study of IntelliMetric Scoring for Responses Writen in Bahasa Malay.* RB-735.