The Hebrew Language Project:
Automated Essay Scoring & Readability Analysis

Anat Ben-Simon&Yoav Cohen

National Institute for Testing & Evaluation (NITE), Israel

In 2000, NITE launched the Hebrew Language Project (HLP). The goal of the project is to develop computational tools for the analysis and evaluation of Hebrew texts.

The current paper reports the results of two studies.

The first study examined the differential contribution of quantified text features to the automated scoring of essays elicited in three different contexts: essays written by 8th-grade native Hebrew-speakers who took part in the Israeli National Assessment of Educational Progress (n=1413); essays written by 12th-grade indigenous students in an instructional writing program (n=662); and essays written by applicants to higher education who took the YAEL test of Hebrew as a foreign language (n=980). The study also examined the effects of the size of the training sample used to develop the prediction model, and the effect of the text-featureclustering model on the precision of the automated score.

The second study examined the feasibility of assessing the difficulty (readability) of reading comprehension passages using statistical, morphological and lexical text features. A total of 7 sets of 10 passages, taken from various tests administered by NITE, were used in the study. Each set was given to three expert judges who were asked to evaluate the difficulty of the texts on a 1-10 scale. The average of the judges' difficulty estimates for each passage yielded a single difficulty measure. Next, a linear prediction model of passage difficulty was developed. 17 of the 50 text features examined were found to be significantly correlated (.23-.41) with the difficulty level obtained from the expert judges. The correlation between the predicted score and the difficulty measure was .80, indicating that about 65% of the variance in text difficulty can be explained by quantified text features.

The Hebrew language project

In 2000, NITE launched the Hebrew Language Project (HLP). The goal of the project is to develop computational tools for the analysis and evaluation of Hebrew texts. Among the various uses of these tools are: linguistic comparison of texts, quantitative analysis of specific properties and features of texts, evaluation of text difficulty (readability) including identification of the sources of difficulty, and Automated Essay Scoring (AES). To attain these goals,a number of tools were developed, among them corpora, computational algorithms, and a dictionary.

The current paper reports the results of two studies.

The first study examined the differential contribution of quantified text features to the automated scoring of essays elicited in three different contexts: essays written by 8th-grade native Hebrew-speakers, essays written by 12th-grade native Hebrew-speakers and essays written by young adults who are non-native Hebrew-speakers. The study also examined the effects of the size of the training sample used to develop the prediction model, and the effect of the text-featureclustering model on the precision of the automated score.

The second study examined the feasibility of assessing the difficulty (readability) of reading comprehension passages taken from various tests using statistical, morphological and lexical text features.

Study 1: The differential contribution of quantified text features to the automated scoring of various essay types

Introduction

Automated Essay Scoring

Automated essay scoring (AES) systems have been in use for the past two decades and have proven to yield reliable and valid measures of writing ability (Shermis& Burstein, 2003; Ben-Simon & Bennett, 2007). In a typical system, a large number of statistical and natural language processing (NLP) features are extracted from a substantial corpus of student essays. The most useful features are identified by correlating the features with human scores and a scoring model is developed. Almost all AES systems attempt to mimic, as closely as possible, the scores produced by human raters. Yet, since the machine-generated features are all but proxies to the criteria used by human raters to assess writing skills, it is important to establish their relationship to writing characteristics that are grounded in a sound theoretical model.

Several commercial essay scoring systems have been developed in the past two decades. The four leading systems are; PEG -- Project Essay Grade (page 2003), IntelliMetric (1997), IEA -- the Intelligent Essay Assessor (1997), and e-rater® (1997). All four systems were developed predominantly for the analysis of texts in the English language, though some of them have also been applied to texts in other languages. In such cases, where systems developed in and for a given language are applied to other languages, they typically use statistical (surface) features rather than natural language processing (NLP) features, which are contingent on the specific lexical, morphological syntactic and discourse features of a given language.

This study reports the results of software designed for automated scoring of essays in the Hebrew language-- NiteRater (2007). To examine the performance of NiteRater, the system was applied to three corpora of student essays. A cross-validation analysis was performed to assess the quality and stability of the prediction model and a sensitivity analysis was carried out to assess the effects of various aspects of the procedure.

Study objectives

The objectives of the study are twofold:
(1) to examine the application of NiteRater to the scoring of essays written in Hebrew, with special emphasis on the differential contribution of the quantified text features to the automated scoring of essays obtained in three different contexts: essays written by 8th-grade native Hebrew-speakers, essays written by 12th-grade native Hebrew-speakers and essays written by young adults who are non-native Hebrew-speakers;
(2) to examine the effects of the size of the training sample used to develop the prediction model and the effect of the text-featureclustering model on the precision of the automated score.

Method

Sample

Three essay corpora were used in the study. Since all essays were hand-written, the essay responses were transcribed and double-checked for typing errors.

1. **G8-L1:** 1314 essays written by 8th-grade native Hebrewspeakers who took the Hebrew language test component of the Israeli National Assessment of Educational Progress (Maytzav). Of the 1314 students who took the test, 665 wrote a summary of a given text (prompt 1) and 649 wrote an argumentative essay (prompt 2). Each essay was scored by a single rater on three writing dimensions: content (0-10), rhetorical structure (0-4), and grammar (0-6). The total score ranged from 0 to 20.

2. **G12-L1**: 662 12th-grade native Hebrew-speakers who participated in an experimental instructional writing program. The program required students to write an argumentative essay in response to a given prompt at the beginning of the program (pre) and again at the end (post). Both essays (pre and post) were written using the same prompt. Each essay was scored by two expert raters according to 25 highly specific writing-dimensions. The scoring scale of each dimension was 1-4. The total score ranged from 25 to100.

3. **YA-L2**: 980 young adults who were non-native speakers of Hebrew, who took the YAEL test of Hebrew as a foreign language. The YAEL test includes three sub-tests, one of which is a writing assignment. The Yael test is administered to all students who take the Psychometric Entrance Test (PET) in languages other than Hebrew. Of the 980 students who took the YAEL test, 484 wrote essays in response to prompt 1 and 496 wrote essays in response to prompt 2. Both essays were of the argumentative type. Each essay was scored by a two expert raters according to four writing dimensions: content, rhetorical structure, lexical richness and grammar. The scoring scale for each writing dimension was 1-7. The total score ranged from 4-28.

Table 1 presents the three essay corpora and gives the mean and standard deviation of essay length for each corpus and sub-corpus.

Table 1: Mean and standard deviation of essaylength by corpus and sub-corpus

		G8-L1 8th-grade native Hebrewspeakers		**G12-L1** 12th-grade native Hebrewspeakers		**YA-L2** Young adults non-native Hebrewspeakers	
		Prompt 1	Prompt 2	Pre	Post	Prompt 1	Prompt 2
Essay length	Mean	**77**	**69**	**289**	**421**	**108**	**123**
	SD	32	25	152	206	40	29
N		665	649	368	294	484	496
		1314		662		980	

Instruments

NiteRater: NiteRater (2007) is an automated essay scoring program. The program extracts about 130 quantified linguistic features from a given text, including statistical, morphological, lexical, syntactic, and discourse features. Following extensive theoretical and empirical research (Safran& Bar-Siman-Tov, 2011) the features were arranged in three main clusters: 61 specific features, 38 combined features and 15 main factors. The program builds a prediction model using user-determined feature-clusters, training sample size and characteristics. The prediction model is typically based on linear stepwise regression. The features used in the final scoring model are those that contribute significantly to the prediction. Their weights are derived empirically.

Procedure

The study procedure involved three stages. First, NiteRater was applied to the three essay corpora to examine the effect of the feature clustering model. Second, NiteRater was applied to the three essay corpora to examine the effect of the size of the training sample. Finally, in the third stage we computed the correlations between the 38 combined features and the human rater score for each of the three corpora and sub-corpora in order to examine the differential functioning of the text features across the three corpora.

The prediction model is based on linear regression. The criterion used by the prediction models was the average of scoresgiven by two humanraters. The accuracy of the prediction model was determined by the correlation between the machine-predicted score and the average human score.
Since the essays in the G8-L1 corpus were scored by only one rater, they were excluded from the analysis in stages 1 and 2.

Stage 1: examining the effect of the feature-clustering model

In the first stage NiteRater was applied to the three essay corpora for the purpose of examining the effect of the feature-clustering model. Three clustering models were examined: (1) 61 specific text features; (2) 38 combined features; and (3) 15 factors.

Each corpus was randomly divided into two samples: the first (training) sample was used to build the prediction mode, while the second sample was used for cross-validation. The cross-validation sample consisted of those essays not employed for training. Essays from the cross-validation sample were scored using the parameters derived from the training sample. The procedure was repeated five times for each corpus, with different training and cross-validation samples.

Stage 2: examining the effect of the training sample

In the second stage, NiteRater was applied to the three essay corpora for the purpose of examining the effect of the training sample size. Three sizes were examined: 20%, 50% and 80% of the full sample. As in stage 1, the cross-validation sample consisted of those essays not employed for training. and these were scored using the parameters derived from the training sample. This procedure was repeated five times for each corpus, with different training and cross-validation samples. The clustering model used to build the prediction model was that of 38 combined features.

Stage 3: examining the differential functioning of the linguistic features

In the third stage, correlations between the 38 combined features and the human rater score were computed for each corpus in order to examine the differential functioning of the text features across the three corpora.

Results

The effect of the feature-clustering model

Tables 2 and 3 give the multiple correlations between the model score and the average human score for samples G12-L1 and YA-L2. The multiple R reported is the average correlation obtained across five iterations of model development and cross validation.

Fairly similar correlations were obtained within each essay corpus across the three clustering models, for both the training and cross-validation samples. The average correlation for the G12-L1 corpus ranged from .74-77 for the training sample and from .72-.74 for the cross-validation sample. The cross-validation correlations were .03-.06 lower than the inter-rater correlation (.80). The average correlation for the YA-L2 corpus ranged from .81 to .85 for the training sample, and from .80 to .81 for the cross-validation sample. The cross-validation correlations were .07-.08 lower than the inter-rater correlation (.88).

In both corpora, the effect of the clustering model used for the development of the prediction model was negligible. However, comparison of the correlations obtained for the training sample with those obtained for the cross-validation sample indicated a slight tendency towards over-fitting in prediction models based on a larger number of features.

Table 2:Prediction accuracy (multiple-R) by clustering model for the G12-L1 essay corpus

G12-L1 (N=662)	61 features	38 combined features	15 factors
Training sample	.77	.74	.74
Cross validation sample (split sample)	.72	.73	.74
No. of features in the model	8	6	6
Inter-rater correlation	.80		

Table 3: Prediction accuracy (multiple-R) by clustering model for the YA-L2 essay corpus

YA-L2 (N=980)	61 features	38 combined features	15 factors
Training sample	.85	.84	.81
Cross validation sample (split sample)	.80	.80	.81
No. of features in the model	17	12	13
Inter-rater correlation	.88		

The effect of training sample size

Tables 4 and 5 give the multiple correlations between the model score and the average human score for samples G12-L1 and YA-L2. The prediction model used the 38 combined features. The multiple R reported is the average correlation obtained across five iterations of model development and cross validation.

Fairly similar correlations were obtained within each essay corpus across the three sampling conditions for the training sample, yet the cross-validation correlations decreased slightly with the decrease in size of the training sample.

The average correlations for the G12-L1 corpus for the training sample were: .75, .74, and .76 for the 80%, 50% and 20% conditions respectively. The average cross-validation correlations were: .76, .73, and .71 for the 80%, 50% and 20% conditions respectively. The cross-validation correlations for the three sampling size conditions were .04-.09 lower than the inter-rater correlation (.80).

The average correlations for the YA-L2 corpus for the training sample were: .83, .84, and .82 for the 80%, 50% and 20% conditions respectively. The average cross-validation correlations were .81, .80, and .79 for the 80%, 50% and 20% conditions respectively. The cross-validation correlations for the three sampling size conditions were .07-.09 lower than the inter-rater correlation (.80).

Figure 1 shows thepredictionaccuracy (R) by the size of the training sample for the G12-L1 and YA-L2 essaycorpora. Results indicate that the prediction accuracy stabilizes at a sample of 150-200 essays.

Table 4: Prediction accuracy (multiple-R) by size of the training sample for the G12-L1 essay corpus

20%	50%	80%		**G12-L1** (N=662)
.76	**.74**	**.75**	Mean	Training sample
.73-.78	.73-.80	.73-.78	Range	
.71	**.73**	**.76**	Mean	Cross validation sample
.69-.73	.71-.76	.72-.78	Range	
.80			Inter-rater correlation	

Table 5: **Prediction accuracy (multiple-R) by size of the training sample for the YA-L2 essay corpus**

20%	50%	80%		**YA-L2** (N=980)
.82	**.84**	**.83**	Mean	Training sample
.79-.84	.81-.86	.82-.85	Range	
.79	**.80**	**.81**	Mean	Cross validation sample
.60-.80	.78-.83	.75-.83	Range	
.88			Inter-rater correlation	

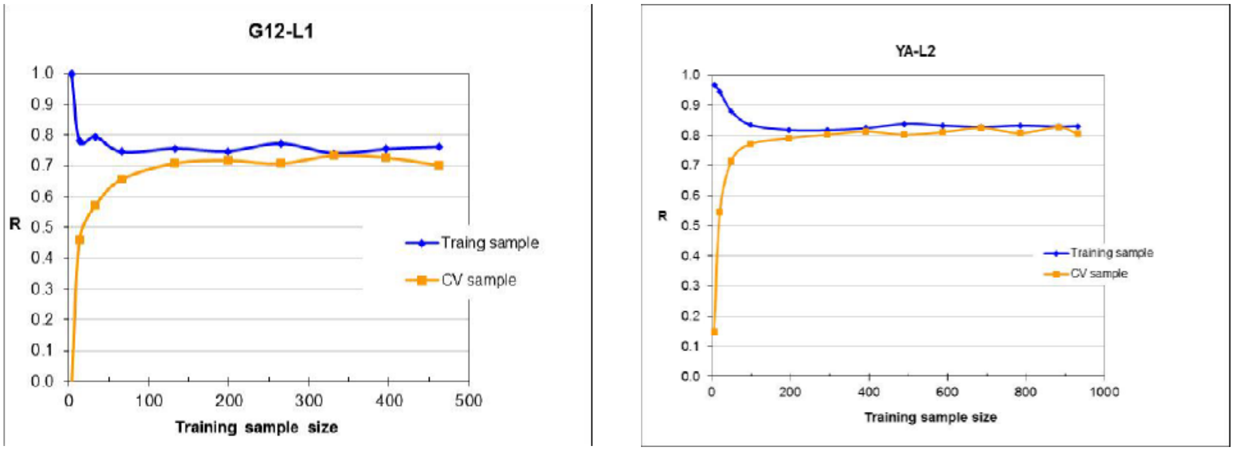

Figure 1: Prediction accuracy (R) by size of the training sample forthe G12-L1 and YA-L2 essaycorpora

The differential functioning of linguistic features across essay corpora

As noted, three essay corpora were used in the study. The corpora varied with regard to the age of the writers and whether Hebrew was their first language (L1) or second language (L2). Table 6 presents the correlations (validity) between a selected group of single and combined features (27) and the total essay score given by the expert raters. The features included in the table are those with the highest correlations (appendix A reports the results by sub-corpus).

Table 6 shows that the following features have the highest correlation with rater scores in all corpora: *lexical diversity* (.54-.74*), essay length* (.52-.66), *text irregularity* (-.33 - -.53), *complement diversity* (.31-.45) and *verb form diversity* (.27-.52). In general, the validity of most features was higher for essays by second-language Hebrew speakers (YA-L2) than for those by Hebrew native speakers (G8-L1 & G12-L1). The most notable differences were found for, *referential pronouns, content density, spelling errors, denominative adjectives, infrequent verb forms, adjectives and word frequency*. The only exception to the abovementioned tendency was knowledge of rare words, which had markedly higher correlations with rater scores in the G12-L1 corpus than in the other two corpora.

Table 6: Correlations between selected single and combined linguistic features and rater scores by corpora

	G8-L1 (n=1314)	G12-L1 (N=662)	YA-L2 (N=980)
Lexical diversity	.54	.63	.74
Essay length	.52	.60	.66
Text irregularity	-.47	-.33	-.53
Complement diversity	.31	.40	.45
Verb form diversity	.33	.27	.52
Punctuation diversity	.28	.24	.33
Conjunction diversity	.18	.24	.35
Referential pronouns(p)	.21	.13	.38
Content density(p)	.19	.14	.37
Spelling errors(p)	-.24	-.07	-.35
Denominative adjectives(p)	.02	.19	.40
Rare words knowledge(p)	.11	.37	.10
Usage of the word 'it'	-.14	-.29	-.14
Infrequent verb forms(p)	.13	.10	.35
Adjectives(p)	.07	.16	.31
Word frequency	-.11	-.01	-.42
Prefixes(p)	.15	.13	.26
Suffixes(p)	.09	.24	.17
Punctuation sentence dividers(p)	.08	.18	.14
Existential words	-.10	-.07	-.23
Causal words(p)	-.16	-.13	-.11
Quantifiers(p)	-.07	-.20	-.07
Connectives(p)	.11	.04	.17
Sentence with multiple negations(p)	-.09	.03	-.20
Third person(p)	.04	.22	-.03
Negation words(p)	-.08	.04	-.17
Present tense verbs(p)	.06	-.20	.08

(p) Denotes a proportional measure

Study2: Assessing the readability of reading comprehension passages usingmachine-generated linguisticfeatures

Introduction

Readability

Readability as defined by George Klare(1963)is “The ease of understanding or comprehension due to the style of writing”.

Readability measures, which reflect various aspects of text difficulty, can be extremely useful in many instructional and assessment contexts, including selection of passages for textbooks and reading comprehension tests.Graesser, Mcnamara&Kulikowich (2011) call this "… for assigning the right texts to the right students at the right time". Such measures can be applied to any given group of texts, rendering a statistical survey of human readers unnecessary.

Most readability research focuses on the development of readability tests. A readability test isa formula that generates a score based on easily extracted statistical characteristics of a given text, such as average word length (in letters or syllables; used as a proxy for semantic difficulty) andaverage sentence length (used as a proxy for syntactic complexity).With the growing prevalence of electronic texts, these features are now automatically generated.

Most formulas are designed for evaluating texts, yet there are also formulas designed for assessing the difficulty of books and speech.Also, formulas can produce either an ease scale, or a grade scale. To validate the scores produced by readability formulas, they are often compared to objective criteria. The most frequently used criteria are: the percentage of correct answers on a reading comprehension test, reading grade level, or expert judgments of text difficulty.

Among the earliest and most thorough works on readability is by Gray & Leary (1935) in the landmark publication *'What Makes aBook Readable'*. The authors used 48 texts of about 100 words each, from various sources and genres. The difficulty of the text was determined by scores on reading comprehension tests. The authors identified 228 featuresassociated with readability and grouped them into four main categories: Content, Style, Format, and Features of organization.64 out of the 228 features were significantly correlated (r>.35) withthe test scores. The highest correlations were found for features associated with sentence length and word frequency. A prediction model based on five features yielded a multiple correlation of .645 with test scores.

Further research confirmed the findings of Gray & Leary and, indeed, most current readability formulas use semantic indices(such as vocabulary difficulty) and syntacticindices (sentence structure – such as average sentence length), which are the best predictors of textual difficulty.

Researchers at School Renaissance Institute (1999, 2000, Paul 2003) and Touchstone Applied Science Associates developed the Advantage-TASA Open Standard (ATOS) Readability Formula for Books. The project was one the most extensive readability studies ever conducted. The corpus included 650 norm-referenced reading texts representing 28,000 recently published K-12textbooks. The combination of three variables yielded the best account of text difficulty: words per sentence (r^2 = .897), average grade-level of words (r^2 = .891), and number of characters per word (r^2 = .839). The formula produces grade-level scores.
In a recent publication Graesser, et al., (2011) review the traditional and most popular computer metrics of text ease/difficulty. These approaches include: (1) the Flesch-Kinkade Grade Level or Reading Ease (Klare 1974-75) which is based on the length of words and the length of sentences; (2) the Degree of Reading Power (DRP; Koslin, Zeno &Koslin, 1987) which relates text characteristics to performance in cloze tasksand the Lexile scores (Stenner, 2006). According to Graesser and his associates these three metrics of text difficulty are highly correlated (r=.89-.94).

Readability formulas for Hebrew texts are practically nonexistent. The only documented study reporting readability formulas for prose texts was published in 1957 (Rabin, 1988) and was based on fairly few linguistic features.

The current study is a first attempt to develop readability formula for Hebrew texts used for the assessment of reading comprehension.

Method

Sample

70 reading comprehension passages were assembled from various tests administered by NITE. Included in the corpus were: 9 passages from the analytical section of pre-med admissions tests; 38 passages from the verbal section of the Psychometric Entrance Test (PET); 9 passages from the admissions test to pre-academic preparatory programs; 8 passages from the nursing school admissions test; 7 passages from the admissions test for teacher colleges and 2 passages from the reading comprehension test of MATAL -- a computerized test-battery for the diagnosis of learning disabilities. Passage length ranged from 261 to 515 words with a mean of 406 and standard deviation of 61.8.

Instruments

NiteRater (2007) is an automated essay scoring program that extracts a pre-selected group of text features from any given text and develops prediction models for essay scores or readability. The features extracted include, statistical, lexical, morphological, semantic, and syntactic features.

Procedure

The 70 passages were grouped into seven roughly parallel sets of 10 passages each.Each of the seven sets was given to three expert judges who were asked to evaluate the difficulty of the texts. All expert judges were test developers or test reviewers with extensive experience in the development of verbal tests. The expert judges were instructed to base their evaluations on both the linguistic level of each text and on its content. The difficulty judgments were given on a scale of 1-10. Three texts, one easy (rank=2), one average (rank=5) and one difficult (rank=9) were pre-selected and used as anchors for the rating scale.

The difficulty estimates provided by the expert judges were averaged for each passage to yield a single difficulty score. Finally, all passages were analyzed by NiteRater. The program extracted 50 single and combined quantified text featuresfrom each passage.

Results

Table 7 presents the mean and standard deviation of the passage length and difficulty rating by the source of the passages (the test it was used in).

It should be noted that unlike other readability studies, this study used reading comprehension passages that were fairly homogeneous in both length and difficulty;

all the passages were used to assess reading compression ability in young adults applying to various institutions of higher education.

Table 7: Mean and standard deviation of the passages length and difficulty ratingsby the source of the passage

Source	No. of passages	Mean (SD) of passage length (in words)	Mean (SD) of passage difficulty
Pre-med admissions test	9	442 (62.0)	6.67 (1.21)
PET's verbal sections	38	415 (42.6)	6.36 (1.87)
Teachers college admissions test	7	298 (29.5)	4.10 (.37)
Pre-academic admissions test	9	413 (44.8)	3.81 (1.24)
Nursing school admissions test	8	405 (83.7)	2.90 (.96)
MATAL testbattery	2	336 (63.3)	1.17 (.24)

Inter-rater agreement

The difficulty level of each passage was determined by three expert judges. Table 8 presents the mean and standard deviation of these difficulty ratingsand the inter-rater agreementobtained for each of the seven passage sets.

The inter-rater reliability between pairs of raters ranged from .22 to .93. The mean inter-rater reliability for the passage sets ranged from .40 to .80 with a mean of .70 and median of .77. The mean inter-rater reliability obtained is quite high in light of the fairly low variability among the passages included in the study.

Table 8: Mean and standard deviation of the difficulty rating of the passages and inter-rater agreementby passage set

Text set	Difficulty rating		Correlation			
	Mean	Std.	R1/R2	R2/R3	R1/R3	**Mean**
1	5.5	2.3	.72	.78	.52	**.69**
2	5.1	2.3	.93	.67	.68	**.80**
3	4.7	2.1	.76	.83	.70	**.77**
4	5.3	1.7	.22	.66	.24	**.40**
5	4.8	2.4	.88	.64	.62	**.74**
6	6.8	2.0	.59	.91	.71	**.77**
7	5.2	1.8	.24	.60	.75	**.56**

Mean						**.70**

Prediction model

Of the 50 features examined, 17 were significantly correlated (.23-.46) with the average passage difficulty as rated by the expert judges. Table 9 lists the correlations for these features. A prediction model applied to the data yielded a correlation of .80 between the predicted score and the average difficulty rating, indicating that about 65% of the variance in text difficulty rating can be explained by quantified text features (see table 10). Only seven of the 50 features had a significant contribution to the prediction model: (1) *sentence length diversity*-- a combined measure based on average sentence length and standard deviation of sentence length; (2) *number of sentences*, (3) *lexical diversity* -- type to token ratio (TTR_D); (4) *negation words*; (5) *sentences with multiple negations*; (6) *determiners*; and (7) *rare wordknowledge* - percentage of highly infrequent words. The last four features are based on the ratio of their occurrencesto total words in the text).

Table 9: Correlations between text features the difficulty rating of reading comprehension passages

Correlation	Text feature
.41	Sentence length diversity
-.41	Content density
-.39	Lexical diversity (TTR-U)
-.38	Lexical diversity (TTR-D)
.33	Lexical diversity (Zipf)
-.33	Referential pronouns
.32	Rare word knowledge(p)
.32	Text irregularity
-.30	Noun to verb ratio(p)
.29	Usage of the word 'it(p)'
.29	Negation words(p)
.29	No. of sentences
.29	Conditional subordinates(p)
-.28	Prefixes(p)
.25	First person(p)
.25	Text length
-.25	Content words (lexemes) diversity

(p) Indicates proportion-based features

Table 10: Linear regression prediction model of the difficulty of reading comprehension passages

	Partial r^2	Model r^2	F	P
Sentence length diversity	.17	.17	14.7	.0003
No. of sentences	.14	.31	14.0	.0004
Lexical diversity (TTR_D)	.08	.39	9.2	.0034
Negation words	.05	.44	6.6	.0125
Sentences with multiple negations	.07	.52	9.7	.0027
Determiners	.09	.60	14.7	.0003
Rare words	.05	.65	8.4	.0052

Reference list

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

Graesser, A. C., Mcnamara, D. S. &Kulikovwich, J.M. ((2011).Coh-Metrix: providing multilevel analysis of text characteristic.*Educational Researcher, Vol. 40,* No. 5, 223-234.

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

Klare, G. R. (1963). *The measurement of readability*. Ames, Iowa: Iowa State University Press.

Klare, G.R. (1974-75). Assessing readability. *Reading Research Quarterly, 10,* 62-102.

Kosline, B.I., Zeno, S. &Koslin, S. (1987). *The DRP: An effective measure in reading*. New York: College Entrance Examination Board.

IntelliMetric Engineer [Computer software]. (1997). Yardley, Advantage Technologies.

NiteRater [Computer software]. (2007). Jerusalem: National Institute for Testing &Evaluation.

Rabin, A. T. (1988). Determining difficulty levels of text written in languages other than English. In B.LZakaluk,& S.J Samuels (Eds),*Readability: Its past, present, and future,* Newark, DE: International Reading Association.

Page, E.B. (2003). Project essay grade: PEG. In M.D. Shermis& J.C.Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective*(pp. 43–54). Mahwah, NJ: Erlbaum.

Safran, Y. & Bar-Siman-Tov, A. (2011). *The factorial structure of written Hebrew*. A paper presented at the 7th Annual Conference of the Israeli Psychometric Association (ISPA). Jerusalem. (Hebrew).

Shermis M. D. & Burstein J. C. (Eds.) (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ:

Stenner, A.J. (2006). Measuring reading comprehension with the Lexile framework. Durham, NC:Metametrix, Inc. Paper presented at the California Comparability Symposium, October 1996. Retrieved from http://www.lexile.com/DesktopDefault.aspx?view=re

Appendix A:Correlations between selected single and combined linguistic features and essay score by corpus and sub-corpus

	G8-L1 (N=1314)			**G12-L1** (n=662)			**YA-L2** (n=980)		
	All	Prm.1	Prm.2	All	pre	post	All	Prm.1	Prm.2
Lexical Diversity	**.54**	.50	.60	**.63**	.68	.71	**.74**	.80	.71
Essay length	**.52**	.49	.57	**.60**	.67	.67	**.66**	.76	.57
Text irregularity(p)	**-.47**	-.41	-.56	**-.33**	-.30	-.45	**-.53**	-.63	-.44
Complements diversity	**.31**	.28	.36	**.40**	.39	.51	**.45**	.54	.35
Verb form diversity	**.33**	.34	.33	**.27**	.23	.39	**.52**	.55	.49
Punctuation diversity	**.28**	.28	.31	**.24**	.25	.29	**.33**	.39	.27
Conjunction diversity	**.18**	.15	.23	**.24**	.23	.31	**.35**	.46	.24
Referential pronouns(p)	**.21**	.31	.10	**.13**	.09	.22	**.38**	.37	.42
Content density(p)	**.19**	.38	-.03	**.14**	.09	.25	**.37**	.40	.36
Spelling errors(p)	**-.24**	-.26	-.22	**-.07**	-.07	-.09	**-.35**	-.33	-.39
Denominative adjectives(p)	**.02**	.07	-.05	**.19**	.18	.27	**.40**	.37	.46
Rare word knowledge(p)	**.11**	.20	.01	**.37**	.37	.47	**.10**	.03	.20
Usage of the word 'it'	**-.14**	-.14	-.15	**-.29**	-.20	-.47	**-.14**	-.14	-.15
Infrequent verb forms(p)	**.13**	.16	.10	**.10**	.03	.22	**.35**	.30	.42
Adjectives(p)	**.07**	.16	-.03	**.16**	.14	.23	**.31**	.32	.31
Word frequency	**-.11**	-.16	-.04	**-.01**	.05	-.08	**-.42**	-.41	-.45
Prefixes(p)	**.15**	.28	.00	**.13**	.10	.20	**.26**	.38	.13
Suffixes(p)	**.09**	.10	.07	**.24**	.17	.38	**.17**	.15	.19
Punctuation sentence dividers(p)	**.08**	.02	.15	**.18**	.14	.27	**.14**	.11	.19
Existential words	**-.10**	-.18	-.01	**-.07**	-.01	-.18	**-.23**	-.20	-.29
Causal words(p)	**-.16**	-.16	-.18	**-.13**	-.20	-.07	**-.11**	-.03	-.20
Quantifiers(p)	**-.07**	-.16	.03	**-.20**	-.17	-.30	**-.07**	-.09	-.05
Connectives(p)	**.11**	.08	.16	**.04**	.01	.08	**.17**	.14	.22
Sentence with multiple negations(p)	**-.09**	-.12	-.06	**.03**	.05	.03	**-.20**	-.14	-.29
Third person(p)	**.04**	.06	.02	**.22**	.21	.29	**-.03**	.08	-.17
Negation words(p)	**-.08**	-.16	.02	**.04**	.03	.06	**-.17**	-.18	-.16
Present tense verbs(p)	**.06**	.05	.07	**-.20**	-.24	-.19	**.08**	.12	.04

(p) Indicates proportion based features