דוח המרכז

382

The Factorial Structure of Written Hebrew and Its Application to AES

Anat Ben-Simon
Yael Safran

September 2012

מרכז ארצי לבחינות ולהערכה (ע"ר)
NATIONAL INSTITUTE FOR TESTING & EVALUATION
المركز القطري للامتحانات والتّقييم
מיסודן של האוניברסיטאות בישראל

דוח מרכז 382

ISBN:978-965-502-166-0

The Factorial Structure of Written Hebrew and Its Application to AES

Anat Ben-Simon & Yael Safran

National Institute for Testing & Evaluation (NITE)

Symposium:
The Structure of Writing Ability across Languages from the Perspective of Automated Scoring

Paper presented at the annual meeting of the
National Council on Measurement in Education (NCME)

April 12-16, 2012, Vancouver, British Columbia, Canada

The Factorial Structure of Written Hebrew and its Application to AES

- Abstract -

In 2000, NITE launched the Hebrew Language Project (HLP), the goal of which is to develop computational tools for the analysis and evaluation of Hebrew texts. The present paper summarizes the initial development, analysis and organization of machine-generated statistical and NLP text features and mapping of the underlying structure of written Hebrew through analysis of the structure of these features. To this end, the paper reports the results of two successive studies.

The purpose of the first study was to examine the characteristics of 133 machine-generated quantified features, to identify the ones most relevant to text difficulty and writing quality and to combine them into empirically based and theoretically meaningful linguistic categories. The study also examined the effect of the text-feature clustering model on the accuracy of the automated score. To attain these goals, a three-stage analysis was carried out using two text corpora and two essay corpora.

The second study focuses on analysis of the factorial structure of writing ability and the validation of machine-generated text features used for its prediction. A factor analysis applied to the selected AES features using five essay-corpora, revealed three AES dimensions: lexical complexity (fluency), topical analysis (content) and vocabulary. However, the AES dimensions failed to align with raters' scores on compatible or close dimensions.

Introduction

Automated Essay Scoring

Automated essay scoring (AES) systems have been in use for the past two decades, producing 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 the scores produced by human raters as accurately as possible. Yet, since the machine-generated features are nothing but proxies for the criteria human raters use to assess writing skills, it is important to establish their relationship to writing characteristics that are grounded in a sound theoretical model.

Among the many procedures used to establish the relationship between machine-generated quantified features and the characteristics of good writing, two complementary validation procedures are commonly applied. The first uses factor analysis techniques to explore the internal structure of the machine-generated features and confirm that features theoretically related to the same language dimension are indeed loaded on the same factor, and are thus reflective of an acknowledged writing

dimension. The second procedure examines the relationships between scores obtained on specific features or feature clusters, which allegedly reflect certain writing dimensions, and scores given by human raters on those same dimensions

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 some of these cases, systems developed in and for a given language are applied to other languages while typically using 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. Given the unique characteristic of the Hebrew language (Cohen, Ben-Simon, & Hovav, 2003), such a practice is not recommended as it might well lead to invalid assessment of writing ability.

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 various tool were developed include a dictionary, corpora and computational algorithms and software such as: tokenizer, morphological analyzer, automatic morphological disambiguator, content analyzer, semantic disambiguator and an automated essay scoring program (NiteRater).

Study objectives

The study summarizes an initial attempt to map the underlying structure of written Hebrew through the analysis of relationships among a large number of statistical and NLP text features and the relationship between these features and the judged quality of writing samples.

The study reports the results of a two-stage factor analysis applied to various corpora, including: edited texts, 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.

Corpora

Texts Corpora

1. M1-TX: a corpus consisting of over 1 million Hebrew words from 644 full texts of various genres. The corpus is used for research and development purposes.
2. NR-TX: a corpus consisting of 144 narrative texts taken from text books used in K-12. The text difficulty level ("text level") is determined by the grade level in which they are used. The NR-TX corpus is a sub-corpus of corpus M1.

Essay corpora

3. G8-L1: 1314 essays written by 8th-grade native Hebrew speakers who took the Hebrew language test of the *Meitzav* – Hebrew acronym for "Growth and Efficiency Measures of Schools (Israeli national assessment of educational progress). Of the 1314 students who took the test, 665 wrote a summary of a given text (Topic 1) and 649 wrote an argumentative essay (Topic 2). Each essay was scored by a single rater on three writing dimensions: content (0-10), organization (0-4), and grammar (0-6). The scale for the total score was 0-20.
4. 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 to the same prompt. Each essay was scored by two expert raters on 25 highly specific scoring dimensions, each on a scale of 1-4. The scoring dimensions were grouped into four super-dimensions: content, relevance to topic, awareness of addressee and grammar. The scale for the total score was 25-100.

5. YA-L2: 980 young adult 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 offered to all students who take the Psychometric Entrance Test (PET) in languages other than Hebrew. Of the 980 students who took the test, 484 wrote essays in response to one prompt (Topic 1) and 496 wrote essays in response to another prompt (Topic 2). Both essays were of the argumentative type. Each essay was scored by two expert raters on four writing dimensions: content, organization, word choice & style, and grammar. The scoring scale for each writing dimension was 1-7. The scale for the total score was 4-28.

All essays included in the three essay corpora were hand-written; the essays were transcribed and double-checked for typing errors.

Table 1 presents the three essay corpora and gives the characteristics of each corpus.

Table 1: Description of the essay corpora used in the study

		G8-L1 8th-grade native Hebrew speakers		**G12-L1** 12th-grade native Hebrew speakers		**YA-L2** Young adults non-native Hebrew speakers	
		Topic 1	Topic 2	Pre	Post	Topic 1	Topic 2
Scoring dimensions		1. Content 2. Organization 3. Grammar		1. Content 2. Relevance to topic 3. Awareness of addressee 4. Grammar		1. Content 2. Organization 3. Word choice & style 4. Grammar	
Scoring scale for total score		0-20		25-100		4-28	
Essay length	*Mean*	77	69	289	421	108	123
	SD	32	25	152	206	40	29
Total Score	*Mean*	11.0	12.3	58.4	71.3	19.4	17.5
	SD	5.76	5.87	13.5	12.9	4.88	5.93
N		665	649	368	294	484	496

Instruments

NiteRater (2007): This automated essay scoring software extracts quantified linguistic features from any given text, including statistical, morphological, lexical, morpho-syntactic, and discourse features. The software is used for text analysis and for essay scoring. In its application to essay scoring, the software constructs a prediction model for any given essay corpus based on a training sample, tests the validity of the model on a "test sample" and applies the model to the remaining essays. The software allows the user to select or define the text features or dimensions to be included in the prediction model, to choose the prediction and cross validation method, and to define the size and characteristics of the training and test samples. The prediction model can be based on pre-determined weights, on the weight extracted from a standard regression, or on weight obtained by stepwise linear regression in which the features used in the scoring model are those that make a significant contribution to the prediction.

The current version of NiteRater includes 179 micro-features. The default version uses 16 dimensions, which encompass 31 features.

The paper reports the results of two studies. The first study describes the developmental process of the final set of features used for essay scoring. The second study examines the factorial structure of writing ability from the perspective of machine-generated text features used for its prediction. This study uses the final set of features obtained in study 1.

Study 1 – Preliminary analysis of text features

The purpose of this study was to examine the characteristics of 133 machine-generated quantified features (micro-features), identify those features most relevant to text difficulty and writing quality, and assign them empirically based and theoretically meaningful linguistic categories. To attain these goals, a three-stage analysis was carried out using the two text corpora (NR-TX and M1-TX) and the two essay-corpora available at the time, G12-L1 and YA-L2 (Topic 1).

Stage 1: Analysis of feature characteristics

Since the current study is the first to examine the structure of the Hebrew language in terms of interrelated machine-generated quantified text features (micro-features), preliminary analysis of these features was a prerequisite. Therefore, the first stage of the analysis focused on the examination of the statistical characteristics of 133 available micro-features, their internal structure and their contribution to the prediction of text difficulty and writing quality.

At this stage, two procedures were applied to the four corpora. First, means and standard deviations were calculated for all 133 micro-features in each text and essay corpus. This analysis was used to detect features with undesirable distributions (e.g., low variability) and features which produce unusually low or high values. Next, all 133 micro-features generated from the NR-TX corpus and the two essay-corpora were correlated with text level and total essay score respectively. Based on the results obtained from these analyses, 72 micro-features were eliminated from further use in prediction models of readability and AES (due to lack of variability, very high correlations with other features or very low to no correlation with text level and essay scores) and few other features were redefined. The resulting set comprised 61 features that were subjected to further analysis.

Stage 2: Factor analysis and feature clustering

In stage 2, corpus M1 and the G12-L1 essay corpora were randomly divided into two sub-samples (S1 and S2). A factor analysis procedure using the 61 features was then applied to each sub-sample and to the YA-L2 corpus. In addition, the 61 features extracted from the NR-TX corpus and the two G12-L1 sub-corpora and the YA-L2 corpus were correlated with text level and with total essay score respectively. An illustration of the type of data obtained from stage 2 analysis of a sample of features is presented in Table 2. The factor analyses yielded 15 factors (EV>1.0) for all the corpora, which explained 61%-72% of the variance. Of the 15 factors obtained for each corpus, the first three were common to all corpora. They were: (1) vocabulary and content density; (2) lexical diversity & text length; and (3) sentence complexity. The loadings of the features on the first three factors proved stable both across sub-samples within a corpus (see figure 1) and across different corpora (see figure 2).

Correlations between loadings ranged from .79-.97 within corpus and from .52-.90 between corpora.

Table 2: Illustration of the type of data obtained from stage 2 analysis

	Correlation with text level / essay score			**Factor No.**							**Factor loadings**						
	NR-TX	**G12**	**YA**	**M1**			**G12**			**YA**	**M1-TX**			**G12**			**YA**
				All	S-1	S-2	All	S-1	S-2	All	All	S-1	S-2	All	S-1	S-2	All
No. of words	.69	.64	.66	2	2	2	2	2	2	3	.95	.97	.95	1	1	1	1
Type/token ratio	-.45	-.39	-.10	2	2	2	2	2	2	3	-.35	-.34	-.41	-.76	-.55	-.69	-.33
Lexical diversity	.79	.73	.74	2	2	2	2	2	2	3	.89	.92	.88	.73	.80	.74	.88
Avg. sentnc. length	.33	-.02	.01	3	3	3	3	3	3	4	.87	.82	.88	1	1	1	1
SD of sentnc. length	.43	-.07	.21	3	3	3	3	3	3	3	1	1	1	1	.98	.96	.45
Avg. word length	.39	.07	.53	1	1	1	1	1	1	2	-.98	-.97	-.83	.94	.98	-.80	.92
Avg. lexeme freq. (token)	-.26	-.23	-.05	1	1	1	1	1	1	1	.90	.88	.88	-.47	-.41	.71	-.42
Avg. lexeme freq. (type)	-.79	-.60	-.12	2	2	2	0	2	0	1	-.68	-.64	-.72	0	-.30	0	-.71

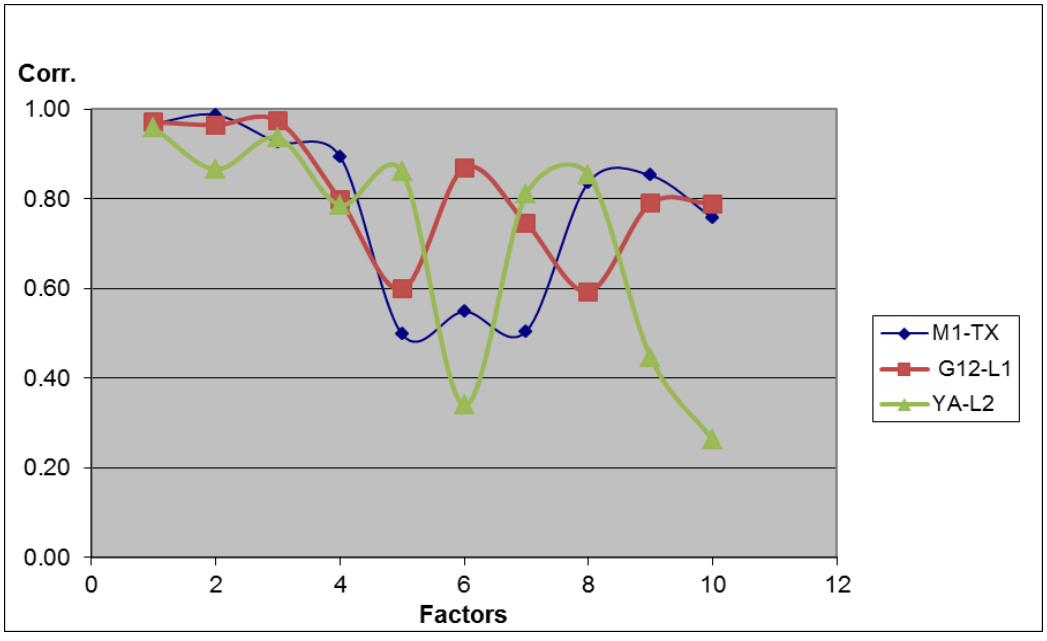

Figure 1: Correlations of feature loadings between corpus sub-samples.

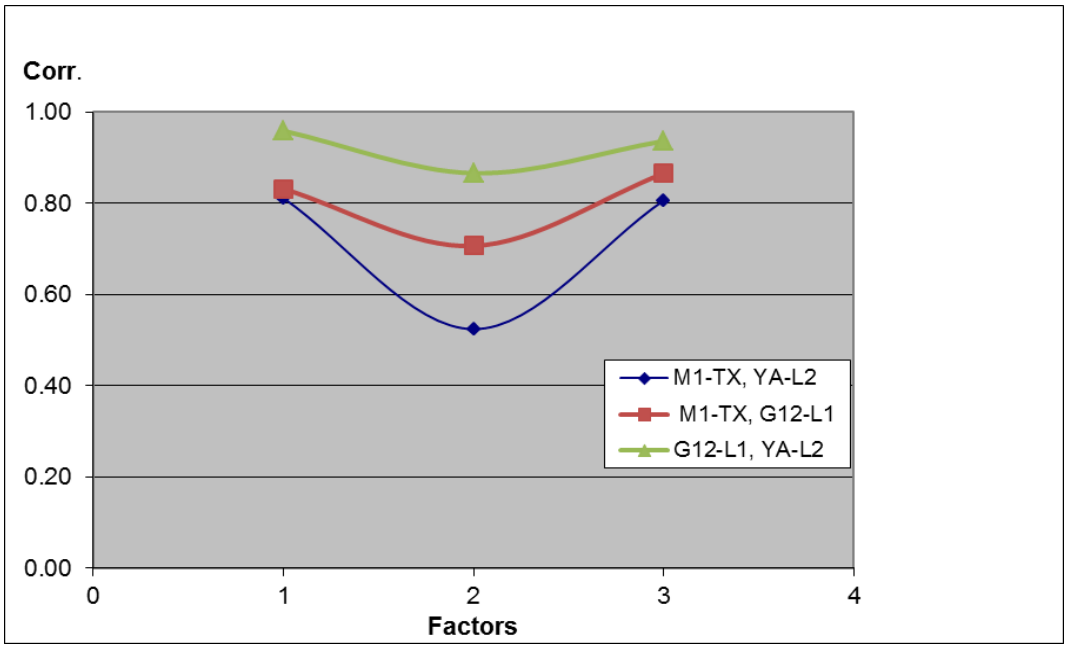

Figure 2: Correlations of feature loadings between the three corpora for the first three factors

As noted above, the third purpose of the study was to combine the micro-features into empirically based and theoretically meaningful linguistic categories. To this end, the 61 features were collapsed into three clustering models on the basis of the previous analyses: (1) 61 single features (micro-features); (2) 38 single and combined features; and (3) 15 factors. The aggregation of features into clusters was based on theoretical and empirical proximity; where the empirical proximity was determined by a similar loading pattern on the main factors and a similar correlation pattern with text level or essay total score across corpora.

To examine the effect of the feature clustering model on the prediction of essay scores, *NiteRater* was applied to each of the two essay corpora using each of the three clustering models. The prediction model was a stepwise linear regression with the average score of two raters as a criterion. Each corpus was randomly divided into two samples, training and test. The training sample was used to build the prediction model, while the test sample was used for cross validation (CV). To enhance the precision of the prediction accuracy, the procedure was repeated five times for five randomly selected training and test (cross-validation) samples.

Table 3 presents prediction accuracy for the three clustering models and two corpora. The multiple R reported in the table is the average correlation obtained across five

iterations of model development and cross validation. Results indicated that the effect of the clustering model used for development of the prediction model was negligible in both essay corpora. The average prediction accuracy (CV) ranged from .72-.74 for the G12-L1 corpus, and from .80 to .81 for the YA-L2 corpus. The respective inter-rater correlations for the G12-L1 and YA-L2 corpora were .80 and .88. However, comparison of the correlations obtained for the training sample vs. the test sample indicated a slight tendency towards over-fitting of models based on a larger number of features.

Table 3
Prediction accuracy (multiple-R) by clustering model for the three essay corpora

Essay corpus		61 features	38 combined features	15 factors	Inter-rater correlation
G12-L1 (N=662)	Training sample	.77	.74	.74	
	Test sample (CV)	.72	.73	.74	.80
	No. of features in model	8	6	6	
YA-L2 Topic 1 (N=484)	Training sample	.85	.84	.81	
	Test sample (CV)	.80	.80	.81	.88
	No. of features in model	17	12	13	

Stage 3: Feature clustering update

The Hebrew Language Project is an ongoing project and new text features are constantly being developed. These are periodically integrated into the feature set obtained thus far. By the completion of stage 2, 16 new NLP features had been developed and needed to be integrated. In addition, three new essay corpora were collected: G12-L1 (Topic 2), and G8-L1 (Topics 1 & 2) allowing for further exploration of the features. These developments called for re-analysis of the new set of 54 text features (38+16). To integrate the new features, all features were inter-correlated and the correlation between the 54 features and essay rater scores were computed for each of the three essay corpora and topics. Following this analysis, several features were eliminated from use in prediction models and a few were replaced with more valid and theoretically sound features. Finally, the remaining 31 features were condensed into 16 features. The re-aggregation of features into clusters

was, once again, based on theoretical and empirical proximity. The addition of the new features to the prediction model, and its reorganization, increased the prediction accuracy of the model for both the G12-L1 essays and the YA-L2 essays by approximately 0.05 points.

The final set of 16 features currently used by NiteRater for essay scoring is presented in Table 4. Of the 16 features, 14 are routinely used for essay scoring. These features are classified into five theoretical writing dimensions: *grammar, organization & development, topical analysis, word complexity and essay length*. To allow for comparison with E-rater, the terminology used in labeling the writing dimensions is based on that used in the E-rater V2.0 (Attali & Burstein, 2006). Two additional features (supplementary features) are used by NiteRater: *essay irregularity* is used to flagging of essays with irregular pattern and *prompt related vocabulary* is used differentially in accordance with essay genre and prompt length.

Of the five writing dimensions, features assessing the *topical analysis* and *word complexity* dimensions are fully developed, while features assessing the *grammar* and *organization & development* dimensions are still in development and give only partial coverage.

Table 4: The final set of features currently used by NiteRater for essay scoring

Dimension	*Feature*	*Description*
Grammar	Mechanics	Spelling errors (letter-string & lexemes)
Word complexity	Vocabulary	Average frequency of lexemes based on a large corpus of texts
	Lexical diversity	Letter-string & lexeme diversity
	Conjunction diversity	Conjunction diversity
	Complement diversity	Subordinate & preposition diversity
	Tense diversity	Tense diversity
Organization & Development	Verb pattern	Usage of verb patterns
	Style	Possession/patient suffix
	Punctuation	Based on proportion of very long sentences, and punctuation types and diversity
	Syntax complexity	Preposition & adjective to noun ratios
Topical analysis	PCA semantic rank	Based on Principal Component Analysis of the semantic similarities (values of cosine correlations) based on the vocabulary of the essays corpus
	Semantic proximity to top essays	Based on similarity (values of cosine correlations) of essay vocabulary to prompt-specific vocabulary of top scored essays
	Score of semantically proximate essays	Average score of the K most similar essays. The similarity is computed using LSA, based on the vocabulary of prompt-specific essays
Essay length	Essay length	Log of no. of words / No. of letter-string types
Supplementary features	Prompt-related vocabulary	Based on the overlap of essay vocabulary and the essays prompt (letter-string & lexeme)
	Essay irregularity	No. of deviant features

Note, appendix A gives the correlations of the features with the total essay scores for each of the five essay corpora included in the study.

Study 2: Validation of NiteRater's AES System

The purpose of this study was to expose the underling internal structure of NiteRater's AES features and to examine the degree to which they correspond with scoring dimensions used by raters.

More specifically, the study addresses the following questions:

1. Do the AES features form a consistent structure across different essay corpora?
2. Can the AES factors be meaningfully interpreted from a cognitive-linguistic perspective?
3. Do the AES factors align with compatible writing dimensions used by raters to score the essays?

Method

All the analyses carried out in this study used the final set of NiteRater's 16 features described in Table 4 and an additional feature – *average word length* – used by E-rater. This feature was added to allow for a better comparison of the structures of NiteRater and E-rater, the latter as reported by Attali & Powers (2008).

The following methods were applied:

Apropos the first research question, an exploratory factor analysis was applied to the 17 AES features. Given the substantial difference between the essay genres (Topics 1 & 2), the G8-L1 and the YA-L2 essay corpora, and the differences in student populations in these samples, this analysis was applied separately to the following five corpora: G8-L1(T1), G8-L1(T2), G12-L1, YA-L2(T1) and YA-L2(T2). A detailed description of the five corpora appears in Table 1. However, given the rather small essay samples comprising each corpora and the fairly large number of features included in the analysis, the results of this analysis may not be stable. Thus, only consistent patterns across the five corpora, allowing for generalization of the results, will be considered evidence of the factorial structure of the NiteRater features.

With respect to the second research question, promax rotation was applied to the retained factors. This procedure was expected to facilitate differentiation between the factors.

Finally, with regard to the third research question, AES factors were correlated with raters scores provided on 3-4 writing dimensions.

Results

NiteRater features: Structure

To answer questions 1 and 2 a preliminary principal component analysis was applied to the each of the five essay corpora: G8-L1(T1), G8-L1(T2), G12-L1, YA-L2(T1) and YA-L2(T2). The eigenvalue >1.0 criterion was used to determine the number of factors to be retained for further analysis. Accordingly, four factors were identified for the G12-L1corpora and for both YA-L2 sub-corpora and five factors for both G8-L1 sub-corpora. These factors accounted for 61%-67% of total EVs across the five corpora, with the first four factors accounting for 60%-66% of total EVs. The corresponding EVs are presented in Table 5.

The distribution pattern of the EVs for the 17 factors (see Figure 3) was similar for the five corpora.

To further differentiate between the factors, the factors were rotated using the oblique rotation procedure. To allow for comparison across the five corpora, only four factors were retained for each corpus. The correlations between the first four factors are presented in Table 6. Table 7 presents the loading of the 17 features on the four factors. Examination of the matrix pattern obtained for the five corpora suggests three main feature clusters. The interpretation of these clusters was guided by the features with the highest loadings on the factor.

Lexical complexity – The following six features are grouped together on this factor: lexical diversity, tense diversity, conjunction diversity, complement diversity, essay length and essay irregularity. With one exception, this pattern is consistent across all five corpora. For some corpora, a few other features are also loaded on this factor, though not in a consistent pattern. This factor reflects the diversity of the lexicon and perhaps the fluency of the writing as expressed in the easiness at which the writer picks the words.

Topical Analysis (content) – Two features are grouped together almost consistently on this factor: semantic proximity to essay, and score of semantically proximate essay. Both features are anchored to essay scores on the content dimension. The third

topical analysis feature, PCA semantic rank, did not show any consistent pattern across the five corpora.

Vocabulary – four features tend to be grouped together on this factor: vocabulary, verb pattern, style and average word length. This factor clearly reflects the writer's vocabulary or, in other words, the register of the written product.

The remaining four features: punctuation, spelling errors, development and prompt-related vocabulary did not form any consistent pattern. This result is not surprising in light of the fact that each one of them addresses different aspects of writing ability.

Table 5: Eigenvalues of principal component analysis for factors with EV >1.0

	G12-L1	***G8-L1***		***YA-L2***	
Factor		(T1)	(T2)	(T1)	(T2)
1	6.11	5.72	4.62	6.37	4.84
2	2.05	2.17	2.48	1.98	2.48
3	1.23	1.35	1.93	1.75	1.53
4	1.05	1.10	1.12	1.12	1.09
5		1.02	1.06		
Percent of total EVs	61	67 (61*)	66 (60*)	66	58

* Percent of total EVs for the first 4 factors

Table 6: Correlation matrices for the four factors using promax procedure

	F-1	*F-2*	*F-3*	*F-1*	*F-2*	*F-3*
	G8-L1 (T1)			**G8-L1** (T2)		
Factor 2	.40			-.03		
Factor 3	.18	.32		.08	.06	
Factor 4	.22	.18	.15	.32	.09	.17
	G12-L1 (T1)					
Factor 2	.21					
Factor 3	.34	.04				
Factor 4	-.20	.01	-.13			
	YA-L2 (T1)			***YA-L2 (T2)***		
Factor 2	.42			.33		
Factor 3	.12	.14		.01	-.10	
Factor 4	.25	.35	.03	-.04	-.12	-.04

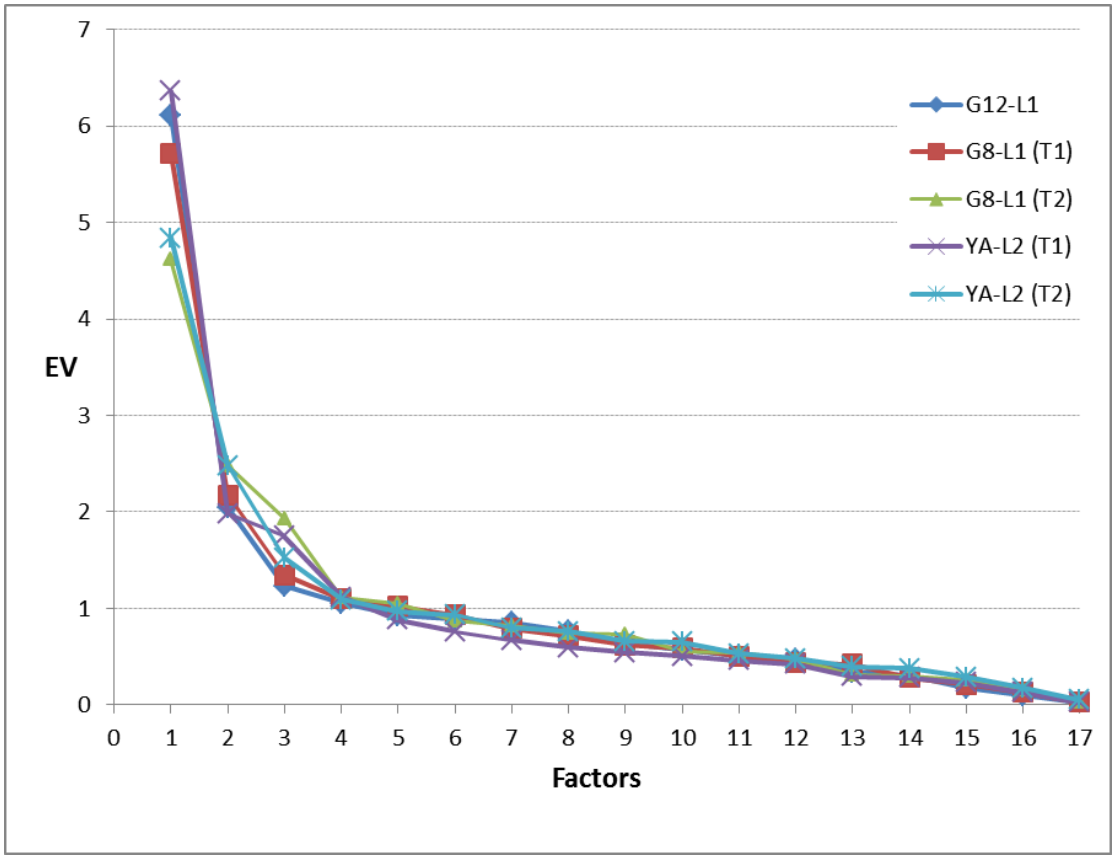

Figure 3: Eigenvalues of 17 factors obtained by principal component analysis for the five corpora.

Table 6: Feature loadings on the first four factors subjected to oblique rotation

	G8-L1 (T1)				G8-L1 (T2)				G12-L1				YA-L2 (T1)				YA-L2 (T2)			
	F-1	F-2	F-3	F-4	F-1	F-2	F-3	F-4	F-1	F-2	F-3	F-4	F-1	F-2	F-3	F-4	F-1	F-2	F-3	F-4
Lexical diversity	.94	.01	.05	.00	.86	.01	.00	.04	.87	.02	.02	.00	.68	.04	.04	.01	.62	.14	.00	.01
Tense diversity	.90	.04	.00	-.03	.94	.00	-.03	-.02	.99	.00	.00	.01	.61	.00	.17	-.01	.02	.00	.00	1.0
Conjunction diversity	.70	-.10	.00	.04	.75	-.13	.00	.00	.38	-.08	.02	-.15	1.0	.00	.00	.00	.72	-.04	-.03	.00
Complement diversity	1.0	-.01	.02	.00	1.0	-.01	.00	.01	.64	.00	.04	-.08	.84	.02	.00	.00	1.0	.00	.00	.00
Essay length	.94	.00	.06	.00	.90	.00	.00	.05	.93	.01	.02	.00	.76	.03	.02	.00	.84	.03	.00	.01
Essay irregularity	-.65	-.14	.00	-.03	-.91	.00	.00	-.06	-.01	.00	-.97	.03	-.88	.00	.00	-.02	-.55	-.06	-.09	.00
Semantic proximity to essay	.08	.01	.75	.00	.14	.00	.97	.01	.77	.01	.08	.00	.15	.51	.00	.05	.24	.41	.00	-.03
Score of semantically proximate essay	.05	.12	.48	.01	.06	.02	.00	.73	.76	.05	.03	.00	.08	.60	.01	.04	.08	.64	-.02	-.01
Vocabulary	.00	.72	.05	.03	.00	1.0	.00	.00	.37	.39	.00	.00	.00	.07	1.0	.03	.01	.41	-.31	.00
Verb pattern	.30	.32	.01	.05	.33	.31	-.01	-.02	.08	.54	-.01	-.08	.19	.08	.46	.01	.03	.60	-.11	.00
Style	.01	.78	.04	-.01	.00	.84	.00	.04	.06	.83	.01	.00	.00	.54	.18	-.06	.00	.35	.00	.43
Avg. word length	.00	1.0	.00	.00	-.05	.69	.06	.00	-.02	1.0	.00	.00	.00	1.0	.00	.01	.00	1.0	.00	.00
PCA semantic rank	.15	.00	.68	.00	.17	.00	.94	.00	.06	.00	1.0	.00	.69	.01	-.12	.00	.44	.00	.30	.00
Punctuation	.04	.02	.00	1.0	.19	.25	-.01	.16	.41	.05	.00	-.19	.00	.00	.00	1.0	.31	.15	.00	-.17
Spelling errors	.00	.00	-.25	-.66	.00	.00	-.01	-1.0	.00	.00	.00	1.0	-.03	-.27	.57	.00	-.03	-.07	-.17	.34
Development	.01	.11	.15	-.41	-.10	.00	.89	.04	.00	.43	.19	.09	.00	.97	.00	-.02	.00	.97	.02	.00
Prompt-related vocabulary	.00	.01	1.0	.00	-.02	.01	1.0	-.06	-1.0	.00	.00	-.01	-.24	.00	-.69	.00	.00	-.01	1.0	.00

Compatibility of AES factors with writing dimension scores

To further explore the interpretation of the AES factors, the factors scores were correlated with the writing dimension scores given by human raters. Though such analysis can potentially assist in interpreting the factors, it is limited by the reliability and validity of the raters' scores.

To examine the compatibility of AES factors with the writing dimensions, three AES dimensions were constructed in accordance with the feature clusters obtained from the factor analysis: lexical complexity, topical analysis and vocabulary. The dimension scores were computed by averaging the scores of the main features associated with each dimension. Table 7 presents the correlation between the AES dimensions and the writing dimension for each of the five corpora. Table 8 gives the correlations between the writing dimensions for each essay-corpus.

In general, all three AES dimensions produced medium to high correlations with the total essay score and with scores on most writing dimensions. This finding indicated that these three dimensions indeed reflect important aspects of writing ability. However, the correlations obtained between each of the AES dimensions and the 3-4 writing dimensions were very close within each essay corpus, to the degree that no further interpretation of the AES dimensions was supported. This pattern is probably due to the high correlations observed between the writing dimensions (see Table 8) which reflect the fact that these dimensions are highly interrelated and raters find it difficult to differentiate between them.

In all instances, the topical analysis dimension produced the highest correlations with the total score. This result may be partly due the fact that both features comprising this dimension are based on the content score and thus tied more closely with the essay score.

Table 7: Correlation between AES dimensions and writing dimensions

	Lexical complexity	***Topical analysis***	***Vocabulary***
G8-L1(T1)			
Content	.36	.67	.39
Organization	.31	.53	.32
Grammar	.35	.47	.34
Total score	.40	.67	.41
G8-L1(T2)			
Content	.44	.46	.15
Organization	.40	.42	.17
Grammar	.37	.40	.16
Total score	.48	.51	.19
G12-L1			
Content	.36	.67	.39
Relevance to topic	.31	.53	.32
Awareness to addressee	.35	.47	.34
Grammar	.40	.67	.41
Total score	.36	.67	.39
YA-L2(T1)			
Content	.66	.77	.56
Organization	.66	.79	.59
Word choice & style	.62	.81	.61
Grammar	.59	.80	.56
Total score	.65	.82	.60
YA-L2(T2)			
Content	.47	.69	.50
Organization	.48	.72	.55
Word choice & style	.47	.73	.56
Grammar	.41	.74	.54
Total score	.48	.76	.57

Table 8: Correlation matrices for writing dimensions

G8-L1(T1)	*Content*	*Organization*	*Grammar*
Organization	.72		
Grammar	.55	.60	
Total score	.93	.86	.79

G8-L1(T2)	*Content*	*Organization*	*Grammar*
Organization	.77		
Grammar	.79	.79	
Total score	.95	.89	.92

G12-L1	*Content*	*Relevance to topic*	*Awareness of addressee*	*Grammar*
Relevance to topic	.88			
Awareness of addressee	.90	.85		
Grammar	.71	.66	.80	
Total score	.95	.89	.97	.86

YA-L2 (T1)	*Content*	*Organization*	*Word choice & style*	*Grammar*
Organization	.95			
Word choice & style	.91	.93		
Grammar	.87	.90	.95	
Total score	.96	.97	.98	.96

YA-L2 (T2)	*Content*	*Organization*	*Word choice & style*	*Grammar*
Organization	.89			
Word choice & style	.83	.89		
Grammar	.79	.89	.92	
Total score	.92	.97	.96	.95

Summary and discussion

The paper reports the results of two studies that examined the internal structure of machine-generated text features developed for automated scoring of texts in the Hebrew language. The first study describes the development process of these features while the second reports the results of a validation study.

In the first study, several procedures were used to select quantified features which are both theoretically relevant to good writing and empirically correlated with essay scores. Both theoretical and empirical considerations were used to combine micro-features into clusters reflecting acknowledged dimensions of writing characteristics. Four corpora were used for this purpose, including edited texts and essay corpora of native and non-native Hebrew-speakers. The final product of this process was a set of 16 features classified into the following dimensions: grammar, word complexity, organization & development, topical analysis, essay length and supplementary features. These features comprise the current version of the NiteRater system which is used for automated essay scoring.

In the second study, factor analysis was applied to the NiteRater features using five essay corpora. The analysis produced three main clusters (AES dimensions): (1) lexical complexity – reflecting the diversity of the lexicon and the fluency of the writing; (2) topical analysis – associated with the essay's content; and (3) vocabulary - reflecting the vocabulary level of the writer. The feature loading pattern on these dimensions was highly consistent across the five corpora included in the study. Of NiteRater's 16 features, 12 are contained in these clusters while the remaining four features – punctuation, spelling errors, development and prompt-related vocabulary – did not form any consistent pattern. This is probably attributable to the fact that each one of them addresses different aspects of writing ability. This analysis follows the work of Attali & Burstein (2006) who, using factor analysis, found three non-content trait scores which generalize across a large variety of essay corpora: *word choice* (measured by vocabulary and word length features), *grammatical conventions within a sentence* (measured by the grammar, usage, and mechanics and features) and *organization* (measured by the style, organization and development features). In a more recent study, Attali & Powers (2008) used factor analysis to validate E-rater's AES features and identified three dimensions: *fluency* (features correlated with essay length), *vocabulary*, and *accuracy* (grammar, usage mechanic and style). The

structure of NiteRater's AES features approximates that reported by Attali & Powers (2008), with the exception of the accuracy dimension, which is under-represented in the NiteRater system. Once these accuracy features are added to NiteRater, it is expected that NiteRater's feature structure will mirror E-rater's structure.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment, 4*(3). Available from http://www.jtla.org

Attali, Y., & Powers, D. (2008). *A developmental writing scale* (ETS RR-08-19). Educational Testing Service: Princeton, NJ.

Ben-Simon, A., & Bennett, R.E. (2007). Toward more substantively meaningful automated essay scoring. *Journal of Technology, Learning, and Assessment, 6*(1). Available from http://www.jtla.org

Cohen, Y., Ben-Simon, A. & Hovav, M. (2003). *The effect of specific language features on the complexity of systems for automated essay scoring.* Report No. 31. Jerusalem: NITE.

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

IntelliMetric Engineer [Computer software]. (1997). Yardley, PA: Vantage Technologies.

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

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

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

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

Appendix 1: Correlations between NiteRater's features and total score by essay corpus

Features	***G8-L1*** *(T1)*	***G8-L1*** *(T2)*	***G12_L1***	***YA-L2*** *(T1)*	***YA-L2*** *(T2)*
Mechanics	-.25	-.22	-.10	-.33	-.39
Vocabulary	.39	.05	.57	.42	.54
Lexical diversity	.51	.60	.73	.80	.71
Complement diversity	.29	.36	.49	.54	.35
Conjunction diversity	.16	.23	.31	.46	.24
Tense diversity	.23	.25	.33	.31	.00
Verb pattern	.30	.27	.25	.49	.47
Style	.15	.14	.37	.21	.20
Punctuation	.20	.32	.29	.30	.30
Syntax complexity	.21	.00	.24	.29	.31
PCA semantic rank	.58	.37	.40	.54	.32
Semantic proximity to top essays	.62	.48	.47	.74	.73
Score of semantically proximate essays	.70	.53	.77	.84	.68
Essay length	.50	.57	.72	.76	.57
Prompt-related vocabulary	.39	-.15	-.50	-.36	-.13
Essay irregularity	-.42	-.56	-.30	-.62	-.43