Vocabulary Profile as a Measure of Vocabulary Sophistication

Su-Youn Yoon, Suma Bhat*, Klaus Zechner
Educational Testing Service, 660 Rosedale Road, Princeton, NJ, USA
{syoon,kzechner}@ets.org
* University of Illinois, Urbana-Champaign, IL, USA
sumapramod@gmail.com

Abstract

This study presents a method that assesses ESL learners' vocabulary usage to improve an automated scoring system of spontaneous speech responses by non-native English speakers. Focusing on vocabulary sophistication, we estimate the difficulty of each word in the vocabulary based on its frequency in a reference corpus and assess the mean difficulty level of the vocabulary usage across the responses (vocabulary profile).

Three different classes of features were generated based on the words in a spoken response: coverage-related, average word rank and the average word frequency and the extent to which they influence human-assigned language proficiency scores was studied. Among these three types of features, the *average word frequency* showed the most predictive power. We then explored the impact of vocabulary profile features in an automated speech scoring context, with particular focus on the impact of two factors: genre of reference corpora and the characteristics of item-types.

The contribution of the current study lies in the use of vocabulary profile as a measure of lexical sophistication for spoken language assessment, an aspect heretofore unexplored in the context of automated speech scoring.

1 Introduction

This study provides a method that measures ESL (English as a second language) learners' competence in vocabulary usage.

Spoken language assessments typically measure multiple dimensions of language ability. Overall proficiency in the target language can be assessed by testing the abilities in various areas including fluency, pronunciation, and intonation; grammar and vocabulary; and discourse structure. With the recent move toward the objective assessment of language ability (spoken and written), it is imperative that we develop methods for quantifying these abilities and measuring them automatically.

A majority of the studies in automated speech scoring have focused on fluency (Cucchiarini et al., 2000; Cucchiarini et al., 2002), pronunciation (Witt and Young, 1997; Witt, 1999; Franco et al., 1997; Neumeyer et al., 2000), and intonation (Zechner et al., 2011). More recently, Chen and Yoon (2011) and Chen and Zechner (2011) have measured syntactic competence in speech scoring. However, only a few have explored features related to vocabulary usage and they have been limited to type-token ratio (TTR) related features (e.g., Lu (2011)). In addition, Bernstein et al. (2010) developed vocabulary features that measure the similarity between the vocabulary in the test responses and the vocabulary in the pre-collected texts in the same topic. However, their features assessed content and topicality, not vocabulary usage.

The speaking construct of vocabulary usage comprises two sub-constructs: sophistication and precision. The aspect of vocabulary that we intend to measure in this paper is that of lexical sophistication, also termed lexical diversity and lexical richness in second language studies. Measures of lexical sophistication attempt to quantify the degree to which a varied and large vocabulary is used (Laufer and Nation, 1995). In order to assess the degree of lexical sophistication, we employ a vocabulary profile-based approach (partly motivated from the results of a previous study, as will be explained in Section 2).

By a vocabulary profile, it is meant that the frequency of each vocabulary item is calculated from a reference corpus covering the language variety of the target situation. The degree of lexical sophistication is captured by the word frequency - low frequency words are considered to be more difficult, and therefore more sophisticated. We then design features that capture the difficulty level of vocabulary items in test takers' responses. Finally, we perform correlation analyses between these new features and human proficiency scores and assess the feature's importance with respect to the other features in an automatic scoring module. The novelty of this study lies in the use of vocabulary profile in an automatic scoring set-up to assess lexical sophistication.

This paper will proceed as follows: we will review related work in Section 2. Data and experiment setup will be explained in Section 3 and Section 4. Next, we will present the results in Section 5, discuss them in Section 6, and conclude with a summary of the importance of our findings in Section 7.

2 Related Work

Measures of lexical richness have been the focus of several studies involving assessment of L1 and L2 language abilities (Laufer and Nation, 1995; Vermeer, 2000; Daller et al., 2003; Kormos and Denes, 2004). The types of measures considered in these studies can be grouped into quantitative and qualitative measures.

The quantitative measures give insight into the number of words known, but do not distinguish them from one another based on their category or frequency in language use. They have evolved to make up for the widely applied measure type-token-ratio (TTR). However, owing to its sensitivity to the number of tokens, TTR has been considered as an unstable measure in differing proficiency levels of language learners. The Guiraud index, Uber index, and Herdan index (Vermeer, 2000; Daller et al., 2003; Lu, 2011) are some measures in this category mostly derived from TTR as either simpler transformations of the TTR or its scaled versions to ameliorate the effect of differing token cardinalities.

Qualitative measures, on the other hand, distinguish themselves from those derived from TTR since they take into account distinctions between words such as their parts of speech or difficulty levels. Adding a qualitative dimension gives more insight into lexical aspects of language ability than the purely quantitative measures such as TTR-based measures. Some measures in this category include a derived form of the limiting relative diversity (LRD) given by $\sqrt{D(verbs)/D(nouns)}$ using the D-measure proposed in (Malvern and Richards, 1997), Lexical frequency profile (LFP) (Laufer and Nation, 1995) and P-Lex (Meara and Bell, 2003).

LFP uses a vocabulary profile (VP) for a given body of written text or spoken utterance and gives the percentage of words used at different frequency levels (such as from the one-thousand most common words, the next thousand most common words) where the words themselves come from a pre-compiled vocabulary list, such as the Academic Word List (AWL) with its associated frequency distribution on words by Coxhead(1998). *Frequency level* refers to a class of words (or appropriately chosen word units) that are grouped based on their frequencies of actual usage in corpora. P-Lex is another approach that uses the frequency level of the words to assess lexical richness. These measures are based on the differing frequencies of lexical items and hence rely on the availability of frequency lists for the language being considered.

These two different types of measures have been used in the analysis of essays written by second language learners of English (ESL). Laufer and Nation (1995) have shown that LFP correlates well with an independent measure of vocabulary knowledge and that it is possible to categorize learners according to different proficiency levels using this measure. In another study seeking to understand the extent to which VP based on students' essays predicted their academic performance (Morris and Cobb, 2004), it was observed that students' vocabulary profile results correlated significantly with their grades. Additionally, VP was found to be indicative of finer distinctions in the language skills of high proficiency nonnative speakers than oral interviews can cover.

Furthermore, these measures have been employed in automated essay scoring. Attali and Burstein

(2006) used average word frequency and average word length in characters across the words in the essay. In addition to the average word frequency measure, the average word length measure was implemented to assess the average difficulty of the word used in the essay under the assumption that the words with more characters were more difficult than the words with fewer characters. These features showed promising performance in estimating test takers' proficiency levels.

In contrast to qualitative measures, quantitative measures did not achieve promising performance. Vermeer (2000) showed that quantitative measures achieve neither the validity nor the reliability of the measures, regardless of the transformations and corrections.

More recently, the relationship of lexical richness to ESL learners' speaking task performance has been studied by Lu (2011). The comprehensive study was aimed at measuring lexical richness along the three dimensions of lexical density, sophistication, and variation, using 25 different metrics (belonging to both the qualitative and quantitative categories above) available in the language acquisition literature. His results, based on the manual transcription of a spoken corpus of English learners, indicate that a) lexical variation (the number of word types) correlated most strongly with the raters' judgments of the quality of ESL learners' oral narratives, b) lexical sophistication only had a very small effect, and c) lexical density (indicative of proportion of lexical words) in an oral narrative did not appear to relate to its quality.

In this study, we seek to quantify vocabulary usage in terms of measures of lexical sophistication: VP based on a set of reference word lists. The novelty of the current study lies in the use of VP as a measure of lexical sophistication for spoken language assessment. It derives support from other studies (Morris and Cobb, 2004; Laufer and Nation, 1995) but is carried out in a completely different context, that of automatic scoring of proficiency levels in spontaneous speech, an area not explored thus far in existing literature.

Furthermore, we investigate the impact of the genre of the reference corpus on the performance of these lexical measures. For this purpose, three different corpora will be used to generate reference frequency levels. Finally, we will investigate how the characteristics of the item types influence the performance of these measures.

3 Data

The AEST balanced data set, a collection of responses from the AEST, is used in this study. AEST is a high-stakes test of English proficiency, and it consists of 6 items in which speakers are prompted to provide responses lasting between 45 and 60 seconds per item, yielding approximately 5 minutes of spoken content per speaker.

Among the 6 items, two items elicit information or opinions on familiar topics based on the examinees' personal experience or background knowledge. These constitute the *independent* (IND) items. The four remaining items are integrated tasks that include other language skills such as listening and reading. These constitute the *integrated* (INT) items. Both sets of items extract spontaneous and unconstrained natural speech. The primary difference between the two elicitation types is that IND items only provide a prompt whereas INT items provide a prompt, a reading passage, and a listening stimulus. The size, purpose, and speakers' native language information for each dataset are summarized in Table 1. All items extract spontaneous, unconstrained natural speech.

Each response was rated by a trained human rater using a 4-point scoring scale, where 1 indicates a low speaking proficiency and 4 indicates a high speaking proficiency. The scoring guideline is summarized in the AEST rubrics.

Since none of the AEST balanced data was double-scored, we estimate the inter-rater agreement ratio of the corpus by using a large double-scored dataset which used the same scoring guidelines and scoring process; using the 41K double-scored responses collected from AEST, we calculate the Pearson correlation coefficient to be 0.63, suggesting a reasonable agreement. The distribution of scores for this data can be found in Table 2.

4 Experiments

4.1 Overview

In this study, we developed vocabulary profile features. From a reference corpus, we pre-compiled

Corpus name	Purpose	# of speakers	# of responses	Native languages	Size (Hrs)
AEST balanced data	Feature evaluation, Scoring model training and evaluation	480	2880	Korean (15%), Chinese (14%), Japanese (7%), Spanish (9%), Others (55%)	44

Table 1: Data size and speakers' native languages

Size	Score1	Score2	Score3	Score4
Number of files	141	1133	1266	340
(%)	5	40	45	12

Table 2: Distribution of proficiency scores in the dataset

multiple sets of vocabulary lists (e.g., a list of the 100 most frequent words in a reference corpus). Next, for each test response, a transcription was generated using the speech recognizer. For each response with respect to each reference word list, vocabulary profile features were calculated. In addition to vocabulary profile features, type-token ratio (TTR) was calculated as a baseline feature. Despite its instability, TTR has been employed in the automated speech scoring systems such as (Zechner et al., 2009), and its use here allows a direct comparison of the performance of the features with the results of previous studies.

4.2 Vocabulary list generation

The three reference corpora we used in this study are presented in Table 3: The General Service List (GSL), the TOEFL 2000 Spoken and Written Academic Language Corpus (T2K-SWAL) and the AEST data.

Corpus	Genre	Tokens	Types
GSL	Written	-	2,284
T2K-SWAL	Spoken	1,869,346	28,855
AEST data	Spoken	5,520,375	23,165

Table 3: Three reference corpora used in this study

GSL (West, 1953) comprises 2,284 words selected to be of "general service" to learners of English. In this study, we used the version with frequency information from (Bauman, 1995). The original version did not include word frequency and was 'enhanced' by John Bauman and Brent Culligan with the frequency information obtained from the Brown Corpus, a collection of written texts.

T2K-SWAL (Biber et al., 2002) is a collection of spoken and written texts covering a broad language variety and use in the academic setting. In this study, only its spoken texts were used. The spoken corpus included manual transcriptions of discussions, conversations, and lectures that occurred in class sessions, study-group meetings, office hours, and service encounters.

Finally, AEST data is a collection of manual transcriptions of spoken responses from the AEST for non-native English speakers. Although there was no overlap between AEST data and the evaluation data (AEST balanced data), the vocabulary lists in AEST data might be a closer match to the vocabulary lists in the evaluation data since both of them come from the same test products. From a content perspective, this dataset is likely to better reflect characteristics of non-native English speakers than the other two reference corpora.

For T2K-SWAL and AEST, all transcriptions were normalized; all the tokens were further decapitalized and removed of all non-alphanumeric characters except for dash and quote. The morphological variants were considered as different words. All words were sorted by the word occurrences in the corpus, and a set of 6 lists were generated: top-100 words (TOP1), word frequency ranks 101-300 (TOP2), ranks 301-700 (TOP3), ranks 701-1500 (TOP4), ranks 1501-3000 (TOP5), and all other words with ranks of 3001 and above (TOP6). For GSL, a set of 5 lists was generated; TOP6 was not generated since GSL only included about 2200 words.

Compared to written texts, speakers tended to use a much smaller vocabulary in speech. For instance, the percentage of words within the top-1000 words on the total word types of AEST data responses was over 90% on average, and they were similar across proficiency levels. This is the reason why we subclassified the top 1000 words into three lists, unlike the vocabulary profile features using top-1000 words as one list like (Morris and Cobb, 2004), which did not have any power to differentiate between proficiency levels.

4.3 Transcription generation for evaluation data

A Hidden Markov Model (HMM) speech recognizer was trained on the AEST dataset, approximately 733 hours of non-native speech collected from 7872 speakers. A gender independent triphone acoustic model and a combination of bigram, trigram, and four-gram language models was used. The word error rate (WER) on the held-out test dataset was 27%. For each response in the evaluation partition, an ASR-based transcription was generated using the speech recognizer.

4.4 Feature generation

Each response comprised less than 60 seconds of speech with an average of 113 word tokens. Due to the short response length, there was wide variation in the proportion of low-frequency word types for the same speaker. In order to address this issue, for each speaker, two responses from the same item-type (IND/INT) were concatenated and used as one large response. As a result, three concatenated responses (one IND response and two INT responses) were generated for each speaker, yielding a total of 480 concatenated responses for IND items and 960 concatenated responses for INT items for our experiment.

First, a list of word types was generated from the ASR hypothesis of each concatenated response. IND items provide only a one-sentence prompt, while INT items provide stimuli including a prompt, a reading passage, and a listening stimulus. In order to minimize the influence of the vocabulary in the stimuli on that of the speakers, we excluded the content words that occurred in the prompts or stimuli from the word type list[1].

Table 4: List of features.

Feature	# of features	Feature type	Description
TTR	1	Ratio	Type-token ratio
TOPn	5 or 6[a]	Listrel	Proportion of types that occurred both the response and TOPn list in the total types of the response.
aRank	1	Rank	Avg. word rank[b]
aFreq	1	Freq	Avg. word freq.[c]
lFreq	1	Freq	Avg. log(word freq)[d]

[a] For GSL, five different features were created using TOP1-TOP5 lists, but TOP6 was not created. For T2K-SWAL and AEST data, six different features were created using TOP1-TOP6 lists separately.

[b] "rank" is the ordinal number of words in a list that is sorted in descending order of word frequency; words not present in the reference corpus get the default rank of RefMaxRank+1.

[c] Avg. word frequency is the sum of the word-frequencies of word types in the reference corpus divided by the total number of words in the reference corpus; words not in the reference corpus get assigned a default frequency of 1.

[d] Same as feature **aFreq**, but the logarithm of the word frequency is taken here

Next, we generated five types of features using three reference vocabulary lists. A maximum of 10 features were generated for each reference list. The feature-types are tabulated in Table 4.

All features above were generated from word types, not word tokens, i.e., multiple occurrences of the same word in a response were only counted once.

Below we delineate the step-by-step process with a sample response that leads to the feature generation outlined in Table 5.

- Step 1: Generate ASR hypothesis for the given speech response. e.g: *Every student has different perspective about how to relax. Playing xbox.*

- Step 2: Generate type list from ASR hypothesis. For the response above we get the list - *about, how, different, xbox, to, relax, every, perspective, student, has, playing*.

[1] This process prevents to measure the content relevance; whether the response is off-topic or not. However, this is not problematic since the features in this study will be used in the conjunction with the features that measure the accuracy of the aspects of content and topicality such as (Xie et al., 2012)'s features.

	word freq. in reference corpus	word rank in the reference corpus	TOPn
about	25672	30	TOP1
how	8944	96	TOP1
has	18105	53	TOP1
to	218976	2	TOP1
different	5088	153	TOP2
every	2961	236	TOP2
playing	798	735	TOP4
perspective	139	1886	TOP5
xbox	1	20000	No

Table 5: An example of feature calculation.

- Step 3: Generate type list excluding words that occurred in the prompt - *about, how, different, xbox, to, every, perspective, has, playing*.

From the ASR hypotheses (result of Step 1), the corresponding type list was generated (Step 2) and two words ('student', 'relax') were excluded from the final list due to overlap with the prompt. The final word list used in the feature generation has 9 types (Step 3).

Next, for each word in the above type list, if it occurs in the reference corpus (a list of words sorted by frequency), its word frequency, word rank and the TOPn information (whether the word belonged to the TOPn list or not) are obtained. If it did not occur in the reference corpus, the default frequency (1) and the default word rank (20000) were assigned. In 5, the default values were assigned for 'xbox' since it was not in the reference corpus.

Finally, the average of the word frequencies and the average of the the word ranks were calculated (aFreq and aRank). For lFreq, the log value of each frequency was calculated and then averaged. For TOPn features, we obtain the proportion of the word types that belong to the TOPn category. For the above sample, the TOP1 feature value was 0.444 since 4 words belong to TOP1 and the total number of word types was 9 (4/9=0.444).

5 Results

5.1 Correlation

We analyzed the relationship between the proposed features and human proficiency scores to assess their influence on predicting the proficiency score. The reference proficiency score for a concatenated response was estimated by summing up the two scores of the constituent responses. Thus, the new score scale was 2-8. Table 6 presents Pearson correlation coefficients (r).

The best performing feature was **aFreq** followed by **TOP1**. Both features showed statistically significant negative correlations with human proficiency scores. **TOP6** also showed statistically significant correlation with human scores, but it was 10-20% lower than **TOP1**. This suggests that a human rater more likely assigned high scores when the vocabulary of the response was not limited to a few most frequent words. However, the use of difficult words (low-frequency) shows a weaker relationship with the proficiency scores.

Features based on AEST data outperformed features based on T2K-SWAL or GSL. The correlation of the AEST data-based **aFreq** feature was -0.61 for the IND items and -0.51 for the INT items; they were approximately 0.1 higher than the correlations of T2K-SWAL or GSL-based features. A similar tendency was found for the TOP1-TOP6 features, although differences between AEST data-based features and other reference-based features were less salient overall.

For top-performing vocabulary profile features including **aFreq** and **TOP1**, the correlations of INT items were weaker than those of the IND items. In general, the correlations of INT items were 10-20% lower than those of the IND items in absolute value.

aFreq and **TOP1** consistently achieved better performance than TTR across all item-types.

5.2 Scoring model building

To arrive at an automatic scoring model, we included the new vocabulary profile features with other features previously found to be useful in a multiple linear regression (MLR) framework. A total of 80 features were generated by the automated speech proficiency scoring system from Zechner et al. (2009),

	Reference	TTR	TOP1	TOP2	TOP3	TOP4	TOP5	TOP6	aRank	aFreq	lFreq
IND	GSL	-.147	-.347	.027	.078	.000	.053	-	.266	-.501	-.260
	T2K-SWAL	-.147	-.338	.085	.207	.055	.020	.168	.142	-.509	-.159
	ATEST	-.147	-.470	.014	.275	.172	.187	.218	.236	-.613	-.232
INT	GSL	-.245	-.255	-.086	-.019	-.068	-.031	-	.316	-.404	-.318
	T2K-SWAL	-.245	-.225	.010	.094	.047	.079	.124	.087	-.405	-.198
	ATEST	-.245	-.345	-.092	.156	.135	.188	.194	.214	-.507	-.251

Table 6: Correlations between features and human proficiency scores

and they were classified into 5 sub-groups: fluency, pronunciation, prosody, vocabulary complexity, and grammar usage. For each sub-group, at least one feature that correlated well with human scores but had a low inter-correlation with other features was selected. A total of following 6 features were selected and used in the base model (**base**):

- **wdpchk** (fluency): Average chunk length in words; a chunk is a segment whose boundaries are set by long silences
- **tpsecutt** (fluency): Number of types per sec.
- **normAM** (pronunciation): Average acoustic model score normalized by the speaking rate
- **phn_shift** (pronunciation): Average absolute distance of the normalized vowel durations compared to standard normalized vowel durations estimated on a native speech corpus
- **stretimdev** (prosody): Mean deviation of distance between stressed syllables in sec.
- **lmscore** (grammar): Average language model score normalized by number of words

We first calculated correlations between these features and human proficiency scores and compared them with the most predictive vocabulary profile features. Table 7 presents Pearson correlation coefficients (r) of these features.

In both item-types, the most correlated features represented the aspect of fluency in production. While **tpsecutt** was the best feature in IND items and the correlation with human scores was approximately 0.66, in INT items, **wdpchk** was the best feature and the correlation was even higher, 0.73.

The performance of **aFreq** was particularly high in IND items; it was the second best feature and only marginally lower than the best feature (by 0.04). **aFreq** also achieved promising performance in INT; it was the fourth best feature. However, the performance was considerably lower than the the best feature, and the difference between the best feature and **aFreq** was approximately 22%.

Features	IND	INT
wdpchk	.538	.729
tpsecutt	.659	.612
normAM	.467	.312
phn_shift	-.503	-.397
stretimemdev	-.442	.429
lmscore	.257	-.535
aFreq	-.613	-.507
TOP1	-.470	-.345
TTR	-.147	-.245

Table 7: Comparison of feature-correlations with human-assigned proficiency scores.

We compared the performances of this **base** model with an augmented model (**base + TTR + all vocabulary profile features**) whose feature set was the **base** augmented with our proposed measures of vocabulary sophistication. Item-type specific multiple linear regression models were trained using five-fold cross validation. The 480 IND responses 960 INT responses were partitioned into five sets, separately. In each fold, an item-type specific regression model was trained using four of these partitions and tested on the remaining one.

The averages of the five-fold models are summarized in Table 8, showing weighted kappa to indicate agreement between automatic scores and human-assigned scores and also the Pearson's correlation (r) of the unrounded (un-rnd) and rounded (rnd) scores with the human-assigned scores. We used the correlation and weighted kappa as performance evaluation measures to maintain the consistency with the previous studies such as (Zechner et al., 2009). In addition, the correlation metric matches better with our goal to investigate the relationship between the predicted scores and the actual scores rather than the difference between the predicted scores and the actual scores.

	Features	un-rnd corr.	rnd corr.	weighted kappa
IND	base	0.66	0.62	0.55
	base + TTR	0.66	0.63	0.56
	base + TTR + all	0.66	0.64	0.57
INT	base	0.76	0.73	0.69
	base + TTR	0.76	0.74	0.70
	base + TTR + all	0.77	0.74	0.70

Table 8: Performance of item-type specific multiple linear regression based scoring models.

The new scores show slightly better agreement with human-assigned scores, but the improvement was small in both item-types, approximately 1%.

6 Discussion

In general, we found that the test takers used a fairly small number of vocabulary items in the spoken responses. On average, the total types used in the responses was 87.21 for IND items and 98.52 for INT items. Furthermore, the proportions of high frequency words on test takers' spoken responses were markedly high. The proportion of top-100 words was almost 50% and the proportion of top-1500 words (summation of TOP1-TOP4) was over 89% on average. This means that only 1500 words represent almost 90% of the active vocabulary of the test takers in their spontaneous speech. Figure 1 presents the average TOP1-TOP6 features across all proficiency levels.

The values of INT items were similar to IND items, but the TOP3-TOP6 values were slightly higher than IND items; INT items tended to include more low frequency words. In order to investigate the impact of the higher proportion of low frequency words in INT items, we selected two features (aFreq and TOP1) and further analyzed them.

Table 9 provides the mean of **aFreq** and **TOP1** for each score level. The features were generated using AEST as a reference.

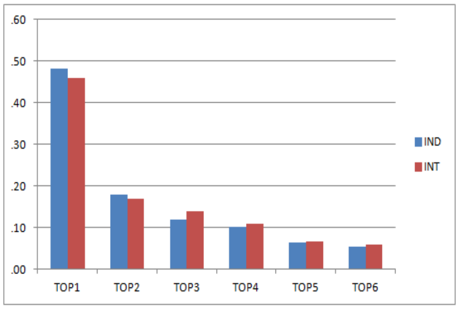

Figure 1: Proportion of top-N frequent words on average

Score	aFreq		TOP1	
	IND	INT	IND	INT
2	43623	36175	.60	.52
3	38165	32493	.55	.49
4	33861	28884	.51	.48
5	30599	27118	.49	.46
6	28485	26327	.46	.45
7	27358	25093	.45	.43
8	26065	24711	.43	.43

Table 9: Mean of vocabulary profile features for each score level

On average, the differences between adjacent score levels in INT items were smaller than those in IND items. The weaker distinction between score levels may result in the lower performance of vocabulary profile features in INT items. Particularly, the differences were smaller in lower score levels (2-4) than in higher score levels (5-8). The relatively high proportion of low frequency words in the low score level reduced the predictive power of vocabulary profile features.

This difference between the item-types strongly supports item-type-specific modeling. We combined the IND and INT item responses and computed a correlation between the features and the proficiency scores over the entirety of data sets. Despite increase in sample sizes, the correlations were lower than both the corresponding correlations of the IND items and the INT items. For instance, the correlation of the T2K-SWAL-based **aFreq** feature was -0.393, and that of the AEST data-based **aFreq** was -0.50, which was approximately 3% lower than the INT items and 10% lower than the IND items. The difference in the vocabulary distributions between the two item-types decreased the performance of the features.

In this study, AEST data-based features outperformed T2K-SWAL-based features. Although no items in the evaluation data overlapped with items in AEST data, the similarity in the speakers' proficiency levels and task types might have resulted in a better match between the vocabulary and its distributions of AEST data with AEST balanced data, finally the AEST data-based features achieved the best performance.

In order to explore the degree to which AEST balanced data (test responses) and the reference corpora matched, we calculated the proportion of word types that occurred in test responses and reference corpora (the coverage of reference list). The ASR hypotheses of AEST balanced data comprised 6,024 word types. GSL covered 73%, T2K-SWAL covered 99%, and AEST data covered close to 100%. Considering the fact that, a) despite high coverage of both T2K-SWAL and AEST data, T2K-SWAL-based features achieved much lower performance than AEST data, and, b) despite huge differences in the coverage between T2K-SWAL and GSL, the performance of features based on these reference corpora were comparable, coverage was not likely to have been a factor having a strong impact on the performance. The large differences in the performance of **TOP1** across reference lists support the possibility of the strong influence of high frequency word types on proficiency; the kinds of word types that were in the TOP1 bins were an important factor that influenced the performance of vocabulary profile features. Finally, genre differences (spoken texts vs. written texts) in reference corpora did not have strong impact on the predictive ability of the features; the performance of features based on written reference corpus (GSL) were comparable to those based on a spoken reference corpus (T2K-SWAL).

Despite the high correlation shown by the individual features (such as **aFreq**), we do not see a corresponding increase in the performance of the scoring model with all the best performing features. The most likely explanation to this is the small training data size; in each fold, only about 380 responses for IND and about 760 responses for INT were used in the scoring model training. Another possibility is overlap with the existing features; the aspect that vocabulary profile features are modeling may be already covered to some extent in existing feature set. In future research, we will further investigate this aspect in details.

7 Conclusions

In this study, we presented features that measure ESL learners' vocabulary usage. In particular, we focused on vocabulary sophistication, and explored the suitability of vocabulary profile features to capture sophistication. From three different reference corpora, the frequency of vocabulary items was calculated which was then used to estimate the sophistication of test takers' vocabulary. Among the three different reference corpora, features based on AEST data, a collections of responses similar to that of the test set, showed the best performance. A total of 29 features were generated, and the average word frequency (**aFreq**) achieved the best correlation with human proficiency scores. In general, vocabulary profile features showed strong correlations with human proficiency scores, but when used in an automatic scoring model in combination with an existing set of predictors of language proficiency, the augmented feature set showed marginal improvement in predicting human-assigned scores of proficiency.

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e–rater R v.2. *The Journal of Technology, Learning, and Assessment*, 4(3).

John Bauman. 1995. About the GSL. Retrieved March 17, 2012 from `http://jbauman.com/gsl.html`.

Jared Bernstein, Jian Cheng, and Masanori Suzuki. 2010. Fluency and structural complexity as predictors of L2 oral proficiency. In *Proceedings of InterSpeech 2010, Tokyo, Japan, September*.

Douglas Biber, Susan Conrad, Randi Reppen, Pat Byrd, and Marie Helt. 2002. Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly*, 36:9–48.

Lei Chen and Su-Youn Yoon. 2011. Detecting structural events for assessing non-native speech. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 38–45.

Miao Chen and Klaus Zechner. 2011. Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. In *Pro-*

ceedings of the 49th Annual Meeting of the Association for Computational Linguistics 2011, pages 722–731.

Catia Cucchiarini, Helmer Strik, and Lou Boves. 2000. Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *The Journal of the Acoustical Society of America*, 107(2):989–999.

Catia Cucchiarini, Helmer Strik, and Lou Boves. 2002. Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *The Journal of the Acoustical Society of America*, 111(6):2862–2873.

Helmut Daller, Roeland van Hout, and Jeanine Treffers-Daller. 2003. Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, 24(2):197–222.

Horacio Franco, Leonardo Neumeyer, Yoon Kim, and Orith Ronen. 1997. Automatic pronunciation scoring for language instruction. In *Proceedings of ICASSP 97*, pages 1471–1474.

Judit Kormos and Mariann Denes. 2004. Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32:145–164.

Batia Laufer and Paul Nation. 1995. Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics*, 16:307–322.

Xiaofei Lu. 2011. The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*.

David D. Malvern and Brian J. Richards. 1997. A new measure of lexical diversity. In *Evolving models of language: Papers from the Annual Meeting of the British Association of Applied Linguists held at the University of Wales, Swansea, September*, pages 58–71.

Paul Meara and Huw Bell. 2003. P_lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Applied Linguistics*, 24(2):197–222.

Lori Morris and Tom Cobb. 2004. Vocabulary profiles as predictors of the academic performance of teaching english as a second language trainees. *System*, 32:75–87.

Leonardo Neumeyer, Horacio Franco, Vassilios Digalakis, and Mitchel Weintraub. 2000. Automatic scoring of pronunciation quality. *Speech Communication*, pages 88–93.

Anne Vermeer. 2000. Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, 17(1):65–83.

Michael West. 1953. *A General Service List of English Words*. Longman, London.

Silke Witt and Steve Young. 1997. Performance measures for phone-level pronunciation teaching in CALL. In *Proceedings of the Workshop on Speech Technology in Language Learning*, pages 99–102.

Silke Witt. 1999. *Use of the speech recognition in computer-assisted language learning*. Unpublished dissertation, Cambridge University Engineering department, Cambridge, U.K.

Shasha Xie, Keelan Evanini, and Klaus Zechner. 2012. Exploring content features for automated speech scoring. In *Proceedings of the NAACL-HLT, Montreal, July*.

Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken english. *Speech Communication*, 51:883–895, October.

Klaus Zechner, Xiaoming Xi, and Lei Chen. 2011. Evaluating prosodic features for automated scoring of non-native read speech. In *IEEE Workshop on Automatic Speech Recognition and Understanding 2011, Hawaii, December*.