Shallow Analysis Based Assessment of Syntactic Complexity for Automated Speech Scoring

Suma Bhat
Beckman Institute,
University of Illinois,
Urbana, IL
spbhat2@illinois.edu

Huichao Xue
Dept. of Computer Science
University of Pittsburgh
Pittsburgh, PA
hux10@cs.pitt.edu

Su-Youn Yoon
Educational Testing Service
Princeton, NJ
syoon@ets.org

Abstract

Designing measures that capture various aspects of language ability is a central task in the design of systems for automatic scoring of spontaneous speech. In this study, we address a key aspect of language proficiency assessment – syntactic complexity. We propose a novel measure of syntactic complexity for spontaneous speech that shows optimum empirical performance on real world data in multiple ways. First, it is both *robust* and *reliable*, producing automatic scores that agree well with human rating compared to the state-of-the-art. Second, the measure makes sense theoretically, both from algorithmic and native language acquisition points of view.

1 Introduction

Assessment of a speaker's proficiency in a second language is the main task in the domain of automatic evaluation of spontaneous speech (Zechner et al., 2009). Prior studies in language acquisition and second language research have conclusively shown that proficiency in a second language is characterized by several factors, some of which are, fluency in language production, pronunciation accuracy, choice of vocabulary, grammatical sophistication and accuracy. The design of automated scoring systems for non-native speaker speaking proficiency is guided by these studies in the choice of pertinent objective measures of these key aspects of language proficiency.

The focus of this study is the design and performance analysis of a measure of the syntactic complexity of non-native English responses for use in automatic scoring systems. The state-of-the art automated scoring system for spontaneous speech (Zechner et al., 2009; Higgins et al., 2011) currently uses measures of fluency and pronunciation (acoustic aspects) to produce scores that are in reasonable agreement with human-rated scores of proficiency. Despite its good performance, there is a need to extend its coverage to higher order aspects of language ability. Fluency and pronunciation may, by themselves, already be good indicators of proficiency in non-native speakers, but from a construct validity perspective[1], it is necessary that an automatic assessment model measure higher-order aspects of language proficiency. Syntactic complexity is one such aspect of proficiency. By "syntactic complexity", we mean a learner's ability to use a wide range of sophisticated grammatical structures.

This study is different from studies that focus on capturing grammatical errors in non-native speakers (Foster and Skehan, 1996; Iwashita et al., 2008). Instead of focusing on grammatical errors that are found to be highly representative of language proficiency, our interest is in capturing the *range* of forms that surface in language production and the degree of *sophistication* of such forms, collectively referred to as *syntactic complexity* in (Ortega, 2003).

The choice and design of objective measures of language proficiency is governed by two crucial constraints:

1. Validity: a measure should show high discriminative ability between various levels of language proficiency, and the scores produced by the use of this measure should show high agreement with human-assigned scores.
2. Robustness: a measures should be derived automatically and should be robust to errors in the measure generation process.

A critical impediment to the robustness constraint in the state-of-the-art is the multi-stage au-

[1]Construct validity is the degree to which a test measures what it claims, or purports, to be measuring and an important criterion in the development and use of assessments or tests.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 1305–1315,
Baltimore, Maryland, USA, June 23-25 2014.

tomated process, where errors in the speech recognition stage (the very first stage) affect subsequent stages. Guided by studies in second language development, we design a measure of syntactic complexity that captures patterns indicative of proficient and non-proficient grammatical structures by a shallow-analysis of spoken language, as opposed to a deep syntactic analysis, and analyze the performance of the automatic scoring model with its inclusion. We compare and contrast the proposed measure with that found to be optimum in Yoon and Bhat (2012).

Our primary contributions in this study are:

- We show that the measure of syntactic complexity derived from a shallow-analysis of spoken utterances satisfies the design constraint of high discriminative ability between proficiency levels. In addition, including our proposed measure of syntactic complexity in an automatic scoring model results in a statistically significant performance gain over the state-of-the-art.
- The proposed measure, derived through a completely automated process, satisfies the robustness criterion reasonably well.
- In the domain of native language acquisition, the presence or absence of a grammatical structure indicates grammatical development. We observe that the proposed approach elegantly and effectively captures this presence-based criterion of grammatical development, since the feature indicative of presence or absence of a grammatical structure is optimal from an algorithmic point of view.

2 Related Work

Speaking in a non-native language requires diverse abilities, including fluency, pronunciation, intonation, grammar, vocabulary, and discourse. Informed by studies in second language acquisition and language testing that regard these factors as key determiners of spoken language proficiency, some researchers have focused on the objective measurement of these aspects of spoken language in the context of automatic assessment of language ability. Notable are studies that have focused on assessment of 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., 2009). The relative success of these studies has yielded objective measures of acoustic aspects of speaking ability, resulting in a shift in focus to more complex aspects of assessment of grammar (Bernstein et al., 2010; Chen and Yoon, 2011; Chen and Zechner, 2011), topic development (Xie et al., 2012), and coherence (Wang et al., 2013).

In an effort to assess grammar and usage in a second language learning environment, numerous studies have focused on identifying relevant quantitative measures. These measures have been used to estimate proficiency levels in English as a second language (ESL) writing with reasonable success. Wolf-Quintero et al. (1998), Ortega (2003), and Lu (2010) found that measures such as mean length of T-unit[2] and dependent clauses per clause (henceforth termed as length-based measures) are well correlated with holistic proficiency scores suggesting that these quantitative measures can be used as objective indices of grammatical development.

In the context of spoken ESL, these measures have been studied as well but the results have been inconclusive. The measures could only broadly discriminate between students' proficiency levels, rated on a scale with moderate to weak correlations, and strong data dependencies on the participant groups were observed (Halleck, 1995; Iwashita et al., 2008; Iwashita, 2010).

With the recent interest in the area of automatic assessment of speech, there is a concurrent need to assess the grammatical development of ESL students automatically. Studies that explored the applicability of length-based measures in an automated scoring system (Chen and Zechner, 2011; Chen and Yoon, 2011) observed another important drawback of these measures in that setting. Length-based measures do not meet the constraints of the design, that, in order for measures to be effectively incorporated in the automated speech scoring system, they must be generated in a fully automated manner, via a multi-stage automated process that includes speech recognition, part of speech (POS) tagging, and parsing.

A major bottleneck in the multi-stage process of an automated speech scoring system for second language is the stage of automated speech recognition (ASR). Automatic recognition of non-native speakers' spontaneous speech is a challenging task as evidenced by the error rate of the state-of-the-

[2]T-units are defined as "the shortest grammatically allowable sentences into which writing can be split." (Hunt, 1965)

art speech recognizer. For instance, Chen and Zechner (2011) reported a 50.5% word error rate (WER) and Yoon and Bhat (2012) reported a 30% WER in the recognition of ESL students' spoken responses. These high error rates at the recognition stage negatively affect the subsequent stages of the speech scoring system in general, and in particular, during a deep syntactic analysis, which operates on a long sequence of words as its context. As a result, measures of grammatical complexity that are closely tied to a correct syntactic analysis are rendered unreliable. Not surprisingly, Chen and Zechner (2011) studied measures of grammatical complexity via syntactic parsing and found that a Pearson's correlation coefficient of 0.49 between syntactic complexity measures (derived from manual transcriptions) and proficiency scores, was drastically reduced to near non-existence when the measures were applied to ASR word hypotheses. This suggests that measures that rely on deep syntactic analysis are unreliable in current ASR-based scoring systems for spontaneous speech.

In order to avoid the problems encountered with deep analysis-based measures, Yoon and Bhat (2012) explored a shallow analysis-based approach, based on the assumption that the level of grammar sophistication at each proficiency level is reflected in the distribution of part-of-speech (POS) tag bigrams. The idea of capturing differences in POS tag distributions for classification has been explored in several previous studies. In the area of text-genre classification, POS tag distributions have been found to capture genre differences in text (Feldman et al., 2009; Marin et al., 2009); in a language testing context, it has been used in grammatical error detection and essay scoring (Chodorow and Leacock, 2000; Tetreault and Chodorow, 2008). We will see next what aspects of syntactic complexity are captured by such a shallow-analysis.

3 Shallow-analysis approach to measuring syntactic complexity

The measures of syntactic complexity in this approach are POS bigrams and are not obtained by a deep analysis (syntactic parsing) of the structure of the sentence. Hence we will refer to this approach as 'shallow analysis'. In a shallow-analysis approach to measuring syntactic complexity, we rely on the distribution of POS bigrams at every proficiency level to be representative of the range and sophistication of grammatical constructions at that level. At the outset, POS-bigrams may seem too simplistic to represent any aspect of true syntactic complexity. We illustrate to the contrary, that they are indeed able to capture certain grammatical errors and sophisticated constructions by means of the following instances. Consider the two sentence fragments below taken from actual responses (the bigrams of interest and their associated POS tags are bold-faced).

1. They ***can/MD to/TO*** survive ...
2. They created ***the culture/NN that/WDT now/RB*** is common in the US.

We notice that Example 1 is not only less grammatically sophisticated than Example 2 but also has a grammatical error. The error stems from the fact that it has a modal verb followed by the word "to". On the other hand, Example 2 contains a relative clause composed of a noun introduced by "that". Notice how these grammatical expressions (one erroneous and the other sophisticated) can be detected by the POS bigrams "MD-TO" and "NN-WDT", respectively.

The idea that the level of syntactic complexity (in terms of its range and sophistication) can be assessed based on the distribution of POS-tags is informed by prior studies in second language acquisition. It has been shown that the usage of certain grammatical constructions (such as that of the embedded relative clause in the second sentence above) are indicators of specific milestones in grammar development (Covington et al., 2006). In addition, studies such as Foster and Skehan (1996) have successfully explored the utility of frequency of grammatical errors as objective measures of grammatical development.

Based on this idea, Yoon and Bhat (2012) developed a set of features of syntactic complexity based on POS sequences extracted from a large corpus of ESL learners' spoken responses, grouped by human-assigned scores of proficiency level. Unlike previous studies, it did not rely on the occurrence of *normative* grammatical constructions. The main assumption was that each score level is characterized by different types of prominent grammatical structures. These representative constructions are gathered from a collection of ESL learners' spoken responses rated for overall proficiency. The syntactic complexity of a test spoken response was estimated based on its

similarity to the proficiency groups in the reference corpus with respect to the score-specific constructions. A score was assigned to the response based on how similar it was to the high score group. In Section 4.1, we go over the approach in further detail.

Our current work is inspired by the shallow analysis-based approach of Yoon and Bhat (2012) and operates under the same assumptions of capturing the range and sophistication of grammatical constructions at each score level. However, the approaches differ in the way in which a spoken response is assigned to a score group. We first analyze the limitations of the model studied in (Yoon and Bhat, 2012) and then describe how our model can address those limitations. The result is a new measure based on POS bigrams to assess ESL learners' mastery of syntactic complexity.

4 Models for Measuring Grammatical Competence

We mentioned that the measure proposed in this study is derived from assumptions similar to those studied in (Yoon and Bhat, 2012). Accordingly, we will summarize the previously studied model, outline its limitations, show how our proposed measure addresses those limitations and compare the two measures for the task of automatic scoring of speech.

4.1 Vector-Space Model based approach

Yoon and Bhat (2012) explored an approach inspired by information retrieval. They treat the concatenated collection of responses from a particular score-class as a 'super' document. Then, regarding POS bigrams as terms, they construct POS-based vector space models for each score-class (there are four score classes denoting levels of proficiency as will be explained in Section 5.2), thus yielding four score-specific vector-space models (VSMs). The terms of the VSM are weighted by the term frequency-inverse document frequency (tf-idf) weighting scheme (Salton et al., 1975). The intuition behind the approach is that responses in the same proficiency level often share similar grammar and usage patterns. The similarity between a test response and a score-specific vector is then calculated by a cosine similarity metric. Although a total of 4 cosine similarity scores (one per score group) were generated, only $cos_4$from among the four similarity scores, and $cosmax$, were selected as features.

- cos_4: the cosine similarity score between the test response and the vector of POS bigrams for the highest score class (level 4); and,
- $cosmax$: the score level of the VSM with which the given response shows maximum similarity.

Of these, $cos_4$was selected based on its empirical performance (it showed the strongest correlation with human-assigned scores of proficiency among the distance-based measures). In addition, an intuitive justification for the choice is that the score-4 vector is a grammatical "norm" representing the *average* grammar usage distribution of the most proficient ESL students. The measure of syntactic complexity of a response, cos_4, is its similarity to the highest score class.

The study found that the measures showed reasonable discriminative ability across proficiency levels. Despite its encouraging empirical performance, the VSM method of capturing grammatical sophistication has the following limitations.

First, the VSM-based method is likely to overestimate the contribution of the POS bigrams when highly correlated bigrams occur as terms in the VSM. Consider the presence of a grammar pattern represented by more than one POS bigram. For example, both "NN-WDT" and "WDT-RB" in Sentence 2 reflect the learner's usage of a relative clause. However, we note that the two bigrams are correlated and including them both results in an over-estimation of their contribution. The VSM set-up has no mechanism to handle correlated features.

Second, the tf-idf weighting scheme for relatively rare POS bigrams does not adequately capture their underlying distribution with respect to score groups. Grammatical expressions that occur frequently in one score level but rarely in other levels can be assumed to be characteristic of a specific score level. Therefore, the more uneven the distribution of a grammatical expression across score classes, the more important that grammatical expression should be as an indicator of a particular score class. However, the simple idf scheme cannot capture this uneven distribution. A pattern that occurs rarely but uniformly across different score groups can get the same weight as a pattern which is unevenly distributed to one score group. Martineau and Finin (2009) observed this weakness of the tf-idf weighting in the domain of sentiment

analysis. When using tf-idf weighting to extract words that were strongly associated with positive sentiment in a movie review corpus (they considered each review as a document and a word as a term), it was found that a substantial proportion of words with the highest tf-idf were rare words (e.g., proper nouns) which were not directly associated with the sentiment.

We propose to address these important limitations of the VSM approach by the use of a method that accounts for each of the deficiencies. This is done by resorting to a maximum entropy model based approach, to which we turn next.

4.2 Maximum Entropy-Based model

In order to address the limitations discussed in 4.1, we propose a classification-based approach. Taking an approach different from previous studies, we formulate the task of assigning a score of syntactic complexity to a spoken response as a classification problem: given a spoken response, assign the response to a proficiency class. A classifier is trained in an inductive fashion, using a large corpus of learner responses that is divided into proficiency scores as the training data and then used to test data that is similar to the training data. A distinguishing feature of the current study is that the measure is based on a comparison of characteristics of the test response to models trained on large amounts of data from each score point, as opposed to measures that are simply characteristics of the responses themselves (which is how measures have been considered in prior studies).

The inductive classifier we use here is the maximum-entropy model (MaxEnt) which has been used to solve several statistical natural language processing problems with much success (Berger et al., 1996; Borthwick et al., 1998; Borthwick, 1999; Pang et al., 2002; Klein et al., 2003; Rosenfeld, 2005). The productive feature engineering aspects of incorporating features into the discriminative MaxEnt classifier motivate the model choice for the problem at hand. In particular, the ability of the MaxEnt model's estimation routine to handle overlapping (correlated) features makes it directly applicable to address the first limitation of the VSM model. The second limitation, related to the ineffective weighting of terms via the the tf-idf scheme, seems to be addressed by the fact that the MaxEnt model assigns a weight to each feature (in our case, POS bigrams) on a per-class basis (in our case, score group), by taking every instance into consideration. Therefore, a MaxEnt model has an advantage over the model described in 4.1 in that it uses four different weight schemes (one per score level) and each scheme is optimized for each score level. This is beneficial in situations where the features are not evenly important across all score levels.

5 Experimental Setup

Our experiments seek answers to the following questions.

1. To what extent does a MaxEnt-score of syntactic complexity discriminate between levels of proficiency?
2. What is the effect of including the proposed measure of syntactic complexity in the state-of-the-art automatic scoring model?
3. How robust is the measure to errors in the various stages of automatic generation?

5.1 Tasks

In order to answer the motivating questions of the study, we set-up two tasks. In the first task, we compare the extent to which the VSM-based measure and the MaxEnt-based measure (outlined in 4.1 and 4.2 above) discriminate between levels of syntactic complexity. Additionally, we compare the performance of an automatic scoring model of overall proficiency that includes the measures of syntactic complexity from each of the two models being compared and analyze the gains with respect to the state-of-the-art. In the second task, we study the measures' robustness to errors incurred by ASR.

5.2 Data

In this study, we used a collection of responses from an international English language assessment. The assessment consisted of questions to which speakers were prompted to provide spontaneous spoken responses lasting approximately 45-60 seconds per question. Test takers read and/or listened to stimulus materials and then responded to questions based on the stimuli. All questions solicited spontaneous, unconstrained natural speech.

A small portion of the available data with inadequate audio quality and lack of student response was excluded from the study. The remaining responses were partitioned into two datasets: the ASR set and the scoring model training/test (SM)

set. The ASR set, with 47,227 responses, was used for ASR training and POS similarity model training. The SM set, with 2,950 responses, was used for feature evaluation and automated scoring model evaluation. There was no overlap in speakers between the ASR set and the SM set.

Each response was rated for overall proficiency by trained human scorers using a 4-point scoring scale, where 1 indicates low speaking proficiency and 4 indicated high speaking proficiency. The distribution of proficiency scores, along with other details of the data sets, are presented in Table 1.

As seen in Table 1, there is a strong bias towards the middle scores (score 2 and 3) with approximately 84-85% of the responses belonging to these two score levels. Although the skewed distribution limits the number of score-specific instances for the highest and lowest scores available for model training, we used the data without modifying the distribution since it is representative of responses in a large-scale language assessment scenario.

Human raters' extent of agreement in the subjective task of rating responses for language proficiency constrains the extent to which we can expect a machine's score to agree with that of humans. An estimate of the extent to which human raters agree on the subjective task of proficiency assessment, is obtained by two raters scoring approximately 5% of data (2,388 responses from ASR set and 140 responses from SM set). Pearson correlation r between the scores assigned by the two raters was 0.62 in ASR set and 0.58 in SM set. This level of agreement will guide the evaluation of the human-machine agreement on scores.

5.3 Stages of Automatic Grammatical Competence Assessment

Here we outline the multiple stages involved in the automatic syntactic complexity assessment. The first stage, ASR, yields an automatic transcription, which is followed by the POS tagging stage. Subsequently, the feature extraction stage (a VSM or a MaxEnt model as the case may be) generates the syntactic complexity feature which is then incorporated in a multiple linear regression model to generate a score.

The steps for automatic assessment of overall proficiency follow an analogous process (either including the POS tagger or not), depending on the objective measure being evaluated. The various objective measures are then combined in the multiple regression scoring model to generate an overall score of proficiency.

5.3.1 Automatic Speech Recognizer

An HMM recognizer was trained using ASR set (approximately 733 hours of non-native speech collected from 7,872 speakers). A gender independent triphone acoustic model and combination of bigram, trigram, and four-gram language models were used. A word error rate (WER) of 31% on the SM dataset was observed.

5.3.2 POS tagger

POS tags were generated using the POS tagger implemented in the Open-NLP toolkit[3]. It was trained on the Switchboard (SWBD) corpus. This POS tagger was trained on about 528K word/tag pairs. A combination of 36 tags from the Penn Treebank tag set and 6 tags generated for spoken languages were used in the tagger.

The tagger achieved a tagging accuracy of 96.3% on a Switchboard evaluation set composed of 379K words, suggesting high accuracy of the tagger. However, due to substantial amount of speech recognition errors in our data, the POS error rate (resulting from the combined errors of ASR and automated POS tagger) is expected to be higher.

5.3.3 VSM-based Model

We used the ASR data set to train a POS-bigram VSM for the highest score class and generated cos_4 and $cosmax$ reported in Yoon and Bhat (2012), for the SM data set as outlined in Section 4.1.

5.3.4 Maximum Entropy Model Classifier

The input to the classifier is a set of POS bigrams (1366 bigrams in all) obtained from the POS-tagged output of the data. We considered binary-valued features (whether a POS bigram occurred or not), occurrence frequency, and relative frequency as input for the purpose of experimentation. We used the maximum entropy classifier implementation in the MaxEnt toolkit[4]. The classifier was trained using the LBFGS algorithm for parameter estimation and used equal-scale gaussian priors for smoothing. The results that follow are based on MaxEnt classifier's parameter settings initialized to zero. Since a preliminary

[3] http://opennlp.apache.org

[4] `http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html`.

Data set	No. of responses	No. of speakers	Score		Score distribution			
			Mean	SD	1	2	3	4
ASR	47,227	7,872	2.67	0.73	1,953 4%	16,834 36%	23,106 49%	5,334 11%
SM	2,950	500	2.61	0.74	166 6%	1,103 37%	1,385 47%	296 10%

Table 1: Data size and score distribution

analysis of the effect of varying the feature (binary or frequency) revealed that the binary-valued feature was optimal (in terms of yielding the best agreement between human and machine scores), we only report our results for this case. The ASR data set was used to train the MaxEnt classifier and the features generated from the SM data set were used for evaluation.

One straightforward way of using the maximum entropy classifier's prediction for our case is to directly use its predicted score-level – 1, 2, 3 or 4. However, this forces the classifier to make a coarse-grained choice and may over-penalize the classifier's scoring errors. To illustrate this, consider a scenario where the classifier assigns two responses A and B to score level 2 (based on the maximum *a posteriori* condition). Suppose that, for response A, the score class with the second highest probability corresponds to score level 1 and that, for response B, it corresponds to score level 3. It is apparent that the classifier has an overall tendency to assign a higher score to B, but looking at its top preference alone (2 for both responses), masks this tendency.

We thus capture the classifier's finer-grained scoring tendency by calculating the *expected value* of the classifier output. For a given response, the MaxEnt classifier calculates the conditional probability of a score-class given the response, in turn yielding conditional probabilities of each score group given the observation – p_i for score group $i \in \{1, 2, 3, 4\}$. In our case, we consider the predicted score of syntactic complexity to be the expected value of the class label given the observation as, $mescore = 1 \times p_1 + 2 \times p_2 + 3 \times p_3 + 4 \times p_4$. This permits us to better represent the score assigned by the MaxEnt classifier as a relative preference over score assignments.

5.3.5 Automatic Scoring System

We consider a multiple regression automatic scoring model as studied in Zechner et al. (2009; Chen and Zechner (2011; Higgins et al. (2011). In its state-of-the-art set-up, the following model uses the features – HMM acoustic model score (global normalized), speaking rate, word types per second, average chunk length in words and language model score (global normalized). We use these features by themselves (**Base**), and also in conjunction with the VSM-based feature (**cva4**) and the MaxEnt-based feature (**mescore**).

5.4 Evaluation Metric

We evaluate the measures using the metrics chosen in previous studies (Zechner et al., 2009; Chen and Zechner, 2011; Yoon and Bhat, 2012). A measure's utility has been evaluated according to its ability to discriminate between levels of proficiency assigned by human raters. This is done by considering the Pearson correlation coefficient between the feature and the human scores. In an ideal situation, we would have compared machine score with scores of grammatical skill assigned by human raters. In our case, however, with only access to the overall proficiency scores, we use scores of language proficiency as those of grammatical skill.

A criterion for evaluating the performance of the scoring model is the extent to which the automatic scores of overall proficiency agree with the human scores. As in prior studies, here too the level of agreement is evaluated by means of the weighted kappa measure as well as unrounded and rounded Pearson's correlations between machine and human scores (since the output of the regression model can either be rounded or regarded as is). The feature that maximizes this degree of agreement will be preferred.

6 Experimental Results

First, we compare the discriminative ability of measures of syntactic complexity (VSM-model based measure with that of the MaxEnt-based measure) across proficiency levels. Table 2 summarizes our experimental results for this task. We

Features	Manual Transcriptions	ASR
$mescore$	0.57	0.52
cos_4	0.48	0.43
$cosmax$	-	0.31

Table 2: Pearson correlation coefficients between measures and holistic proficiency scores. All values are significant at level 0.01. Only the measures cos_4 and $mescore$ were compared for robustness using manual and ASR transcriptions.

notice that of the measures compared, $mescore$ shows the highest correlation with scores of syntactic complexity. The correlation was approximately 0.1 higher in absolute value than that of cos_4, which was the best performing feature in the VSM-based model and the difference is statistically significant.

Seeking to study the robustness of the measures derived using a shallow analysis, we next compare the two measures studied here, with respect to the impact of speech recognition errors on their correlation with scores of syntactic complexity. Towards this end, we compare $mescore$ and $cos_4$when POS bigrams are extracted from manual transcriptions (ideal ASR) and ASR transcriptions.

In Table 2, noticing that the correlations decrease going along a row, we can say that the errors in the ASR system caused both $mescore$ and $cos_4$to under-perform. However, the performance drop (around 0.05) resulting from a shallow analysis is relatively small compared to the drop observed while employing a deep syntactic analysis. Chen and Zechner (2011) found that while using measures of syntactic complexity obtained from transcriptions, errors in ASR transcripts caused over 0.40 drop in correlation from that found with manual transcriptions[5]. This comparison suggests that the current POS-based shallow analysis approach is more robust to ASR errors compared to a syntactic analysis-based approach.

The effect of the measure of syntactic complexity is best studied by including it in an automatic scoring model of overall proficiency. We compare the performance gains over the state-of-the-art with the inclusion of additional features (VSM-based and MaxEnt-based, in turn). Table 3 shows the system performance with different grammar sophistication measures. The results reported are averaged over a 5-fold cross validation of the multiple regression model, where 80% of the SM data set is used to train the model and the evaluation is done using 20% of the data in every fold.

As seen in Table 3, using the proposed measure, $mescore$, leads to an improved agreement between human and machine scores of proficiency. Comparing the unrounded correlation results in Table 3 we notice that the model **Base+mescore** shows the highest correlation of predicted scores with human scores. In addition, we test the significance of the difference between two dependent correlations using Steiger's Z-test (via the `paired.r` function in the R statistical package (Revelle, 2012)). We note that the performance gain of **Base+mescore** over **Base** as well as over **Base + cos4** is statistically significant at level = 0.01. The performance gain of **Base+cos4** over **Base**, however, is not statistically significant at level = 0.01. Thus, the inclusion of the MaxEnt-based measure of syntactic complexity results in improved agreement between machine and human scores compared to the state-of-the-art model (here, **Base**).

7 Discussions

We now discuss some of the observations and results of our study with respect to the following items.

Improved performance: We sought to verify empirically that the MaxEnt model really outperforms the VSM in the case of correlated POS bigrams. To see this, we separate the test set into three subsets A, B, C. Set A contains responses where MaxEnt outperforms VSM; set B contains responses where VSM outperforms MaxEnt; set C contains responses where their predictions are comparable. For each group of responses $s \in \{A, B, C\}$, we calculate the percentage of responses P_s where two highly correlated POS bigrams occur[6]. We found that the percentages follow the order: $P_A = 12.93\% > P_C = 7.29\% >$

[5]Due to differences in the dataset and ASR system, a direct comparison between the current study and the cited prior study was not possible.

[6]We consider two POS bigrams to be highly correlated, when the their pointwise-mutual information is higher than 4.

Evaluation method	Base	Base+cos4	Base+mescore
Weighted $kappa$	0.503	0.524	**0.546**
Correlation (unrounded)	0.548	0.562	**0.592**
Correlation (rounded)	0.482	0.492	**0.519**

Table 3: Comparison of scoring model performances using features of syntactic complexity studied in this paper along with those available in the state-of-the-art. Here, **Base** is the scoring model without the measures of syntactic complexity. All correlations are significant at level 0.01.

$P_B = 4.41\%$. This suggests that when correlated POS bigrams occur, MaxEnt is more likely to provide better score predictions than VSM does.

Feature design: In the case of MaxEnt, the observation that binary-valued features (presence/absence of POS bigrams) yield better performance than features indicative of the occurrence frequency of the bigram has interesting implications. This was also observed in Pang et al. (2002) where it was interpreted to mean that overall sentiment is indicated by the presence/absence of keywords, as opposed to topic of a text, which is indicated by the repeated use of the same or similar terms. An analogous explanation is applicable here.

At first glance, the use of the presence/absence of grammatical structures may raise concerns about a potential loss of information (e.g. the distinction between an expression that is used once and another that is used multiple times is lost). However, when considered in the context of language acquisition studies, this approach seems to be justified. Studies in native language acquisition, have considered multiple grammatical developmental indices that represent the grammatical levels reached at various stages of language acquisition. For instance, Covington et al. (2006) proposed the revised D-level scale which was originally studied by Rosenberg and Abbeduto (1987). The D-Level Scale categorizes grammatical development into 8 levels according to the presence of a set of diverse grammatical expressions varying in difficulty (for example, level 0 consists of simple sentences, while level 5 consists of sentences joined by a subordinating conjunction). Similarly, Scarborough (1990) proposed the Index of Productive Syntax (IPSyn), according to which, the presence of particular grammatical structures, from a list of 60 structures (ranging from simple ones such as including only subjects and verbs, to more complex constructions such as conjoined sentences) is evidence of language acquisition milestones.

Despite the functional differences between the indices, there is a fundamental operational similarity - that they both use the presence or absence of grammatical structures, rather than their occurrence count, as evidence of acquisition of certain grammatical levels. The assumption that a presence-based view of grammatical level acquisition is also applicable to second language assessment helps validate our observation that binary-valued features yield a better performance when compared with frequency-valued features.

Generalizability: The training and test sets used in this study had similar underlying distributions – they both sought unconstrained responses to a set of items with some minor differences in item type. Looking ahead, an important question is the extent to which our measure is sensitive to a mismatch between training and test data.

8 Conclusions

Seeking alternatives to measuring syntactic complexity of spoken responses via syntactic parsers, we study a shallow-analysis based approach for use in automatic scoring.

Empirically, we show that the proposed measure, based on a maximum entropy classification, satisfied the constraints of the design of an objective measure to a high degree. In addition, the proposed measure was found to be relatively robust to ASR errors. The measure *outperformed* a related measure of syntactic complexity (also based on shallow-analysis of spoken response) previously found to be well-suited for automatic scoring. Including the measure of syntactic complexity in an automatic scoring model resulted in statistically significant performance gains over the state-of-the-art. We also make an interesting observation that the impressionistic evaluation of syntactic complexity is better approximated by the presence or absence of grammar and usage patterns (and not by their frequency of occurrence), an idea supported by studies in native language acquisition.

References

Adam L Berger, Vincent J Della Pietra, and Stephen A Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71.

Jared Bernstein, Jian Cheng, and Masanori Suzuki. 2010. Fluency and structural complexity as predictors of L2 oral proficiency. In *Proceedings of InterSpeech*, pages 1241–1244.

Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. 1998. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *Proc. of the Sixth Workshop on Very Large Corpora*.

Andrew Borthwick. 1999. *A maximum entropy approach to named entity recognition*. Ph.D. thesis, New York University.

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*, IUNLPBEA '11, pages 38–45, Stroudsburg, PA, USA. Association for Computational Linguistics.

Miao Chen and Klaus Zechner. 2011. Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 722–731.

Martin Chodorow and Claudia Leacock. 2000. An unsupervised method for detecting grammatical errors. In *Proceedings of NAACL*, pages 140–147.

Michael A Covington, Congzhou He, Cati Brown, Lorina Naci, and John Brown. 2006. How complex is that sentence? a proposed revision of the rosenberg and abbeduto d-level scale. *ReVision. Washington, DC http://www. ai. uga. edu/caspr/2006-01-Covington. pdf.(Accessed May 10, 2010.)*.

Catia Cucchiarini, Helmer Strik, and Lou Boves. 2000. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *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.

Sergey Feldman, M.A. Marin, Mari Ostendorf, and Maya R. Gupta. 2009. Part-of-speech histograms for genre classification of text. In *Proceedings of ICASSP*, pages 4781 –4784.

Pauline Foster and Peter Skehan. 1996. The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, 18:299–324.

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

Gene B Halleck. 1995. Assessing oral proficiency: a comparison of holistic and objective measures. *The Modern Language Journal*, 79(2):223–234.

Derrick Higgins, Xiaoming Xi, Klaus Zechner, and David Williamson. 2011. A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language*, 25(2):282–306.

Kellogg W Hunt. 1965. Grammatical structures written at three grade levels. ncte research report no. 3.

Noriko Iwashita, Annie Brown, Tim McNamara, and Sally O'Hagan. 2008. Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, 29(1):24–49.

Noriko Iwashita. 2010. Features of oral proficiency in task performance by efl and jfl learners. In *Selected proceedings of the Second Language Research Forum*, pages 32–47.

Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D Manning. 2003. Named entity recognition with character-level models. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 180–183. Association for Computational Linguistics.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

M.A Marin, Sergey Feldman, Mari Ostendorf, and Maya R. Gupta. 2009. Filtering web text to match target genres. In *Proceedings of ICASSP*, pages 3705–3708.

Justin Martineau and Tim Finin. 2009. Delta tfidf: An improved feature space for sentiment analysis. In *ICWSM*.

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

Lourdes Ortega. 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college–level L2 writing. *Applied Linguistics*, 24(4):492–518.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

William Revelle, 2012. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 1.2.1.

Sheldon Rosenberg and Leonard Abbeduto. 1987. Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults. *Applied Psycholinguistics*, 8:19–32.

Ronald Rosenfeld. 2005. *Adaptive statistical language modeling: a maximum entropy approach*. Ph.D. thesis, IBM.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Hollis S Scarborough. 1990. Index of productive syntax. *Applied Psycholinguistics*, 11(1):1–22.

Joel R. Tetreault and Martin Chodorow. 2008. The ups and downs of preposition error detection in ESL writing. In *Proceedings of COLING*, pages 865–872.

Xinhao Wang, Keelan Evanini, and Klaus Zechner. 2013. Coherence modeling for the automated assessment of spontaneous spoken responses. In *Proceedings of NAACL-HLT*, pages 814–819.

Silke Witt and Steve Young. 1997. Performance measures for phone-level pronunciation teaching in CALL. In *Proceedings of STiLL*, 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.

Kate Wolf-Quintero, Shunji Inagaki, and Hae-Young Kim. 1998. Second language development in writing: Measures of fluency, accuracy, and complexity. Technical Report 17, Second Language Teaching and curriculum Center, The University of Hawai'i, Honolulu, HI.

Shasha Xie, Keelan Evanini, and Klaus Zechner. 2012. Exploring content features for automated speech scoring. In *Proceedings of the NAACL-HLT*, pages 103–111.

Su-Youn Yoon and Suma Bhat. 2012. Assessment of esl learners' syntactic competence based on similarity measures. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 600–608. Association for Computational Linguistics.

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(10):883–895.