Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 67 (2015) 42–57

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Automatic assessment of syntactic complexity for spontaneous speech scoring

Suma Bhat [a,*], Su-Youn Yoon [b]

[a] Beckman Institute for Advanced Science and Technology, University of Illinois, Urbana-Champaign, IL, USA
[b] Educational Testing Service, Rosedale Road, Princeton, NJ 08541, USA

Received 12 November 2013; received in revised form 7 July 2014; accepted 12 September 2014
Available online 26 September 2014

Abstract

Expanding paradigms of language learning and testing prompt the need for developing objective methods of assessing language proficiency from spontaneous speech. In this paper new measures of syntactic complexity for use in the framework of automatic scoring systems for second language spontaneous speech, are studied. In contrast to most existing measures that estimate competence levels indirectly based on the length of production units or frequency of specific grammatical structures, we capture the differences in the distribution of morpho-syntactic features across learners' proficiency levels. We build score-specific models of part of speech (POS) tag distribution from a large corpus of spontaneous second language English utterances and use them to measure syntactic complexity.

Given a speaker's response, we consider its similarity with a set of utterances scored for proficiency by humans. The comparison is made by considering the distribution of POS tags in the response and a score-level. The underlying distribution of POS tags (indicative of syntactic complexity) is represented via two models: a vector-space model and a language model.

Empirical results suggest that the proposed measures of syntactic complexity show a reasonable association with human-rated proficiency scores compared to conventional measures of syntactic complexity. They are also significantly robust against errors resulting from automatic speech recognition, making them more suitable for use in operational automated scoring applications. When used in combination with other measures of oral proficiency in a state-of-the-art scoring model, the predicted scores show improved agreement with human-assigned scores over a baseline scoring model without our proposed features.

Keywords: Language testing; Automated scoring; Speaking proficiency; Computer-aided language learning; Syntactic complexity; Objective measures

1. Introduction

The expansion of natural language and speech processing capabilities have created new areas of application for use with the expanding paradigms of human–computer interaction. Today, language learning is gradually moving away from tutor-based or language-lab based scenarios, to become computer-aided. The obvious advantages of this emerging paradigm are both its potential to make language learning materials accessible to a wider range of learners at reduced costs (as compared to using human tutors) and being more ubiquitous for use on a more flexible schedule. With more opportunities for computer-aided language learning (CALL) interfaces being created today, there is an increased need to endow CALL systems with the ability to assess language ability automatically. While the resulting technology could be used for automated scoring in a testing scenario or for providing diagnostic feedback to the learner, efforts are being made to develop objective methods of assessing language ability from spontaneous speech.

[*] Corresponding author.
E-mail addresses: spbhat2@illinois.edu (S. Bhat), syoon@ets.org (S.-Y. Yoon).

http://dx.doi.org/10.1016/j.specom.2014.09.005
0167-6393/© 2014 Elsevier B.V. All rights reserved.

Overall spoken proficiency in a target language can be assessed by testing the abilities in various areas including fluency, pronunciation and intonation, grammar and vocabulary, and discourse structure. Currently, speech-enabled dialog systems allow learners to practice their speaking and listening with a virtual interlocutor (e.g., SpeakESL), to receive feedback on their pronunciation [e.g., Carnegie Speech, or *Native Accent* (Eskenazi et al., 2007), *EduSpeak* from SRI (Franco et al., 2000)].

These and other spoken response scoring systems work on restricted speaking tasks such as reading a passage or answering questions with a limited range of responses (Bernstein et al., 2000; Balogh et al., 2007). In contrast to these systems that score restricted speech, scoring unstructured, unrestricted, and spontaneous responses poses a much harder problem. In addition, if the systems target learners with diverse levels of second language proficiency and varied first language backgrounds, the difficulty increases substantially.

The state-of-the-art system for scoring spontaneous speech in a testing scenario is SpeechRater^{SM} (Zechner et al., 2009). Although the current capability is sufficiently advanced to allow it to be used for the scoring of TOEFL® Practice Online (TPO), a low-stakes practice test product, there is room for improving its feature set by expanding the coverage of important aspects of speaking proficiency and modifying others. For instance, aspects of grammar and vocabulary sophistication are only being measured indirectly (more details on this later in this paper) and a more direct approach to measuring these aspects is necessary.

Taking the challenges posed in processing spontaneous speech automatically into consideration, we propose a set of measures of grammatical competence. This paper describes the measures and their potential of being used in a state-of-the-art spontaneous scoring system. In Section 2 the problem being studied is placed into the context of previous work done in the related areas of written and spoken language assessment. A description of the measures studied in this paper if found in Section 3. In Section 4, we delve into the details of the implementation of our proposed measures. A description of the data is provided in Section 5. The experimental details comprise the material in Section 6 and the results are presented in Section 7. In Section 8 we discuss the results of data analyses and highlight some extensions to the study. Finally, a brief summary of the major findings of the paper is presented in Section 9.

2. Motivation

2.1. Assessment of syntactic competence in second language learning

Numerous studies in related second language acquisition literature reveal that syntactic complexity and grammar accuracy are regarded as some of the key skills that strongly influence second language proficiency. Thus, the study of measures that reflect language learners' command of these influential skills has been the central theme of various studies in the area of second language acquisition.

In related literature, Ortega (2003) indicates that "the *range* of forms that surface in language production and the degree of *sophistication* of such forms" are two important areas in grammar usage collectively termed, "syntactic complexity". A vast majority of measures of syntactic complexity have been used as indicators of levels of acquisition of syntactic competence, and in turn, are suggestive of proficiency levels in ESL writing (e.g. Wolf-Quintero et al., 1998; Ortega, 2003; Lu, 2010). These measures have been broadly classified into two groups (Bardovi-Harlig and Bofman, 1989). The first group is related to the acquisition of specific grammatical expressions corresponding to various stages of language acquisition. Frequencies of negation or relative clauses – in terms of whether these expressions occurred in the test responses without errors, fall into this group (hereafter, the expression-based group). The second group, not tied to particular structures, is related to length of clauses or the relationship between clauses (hereafter, the length-based group). Representative measures in the second group include the *mean length of clause unit*, the *ratio of dependent clauses to the total number of clauses*, and the *number of verb phrases per clause*.

In contrast with syntactic complexity, grammatical accuracy is the ability to generate sentences without grammatical errors. The measures in this group can be classified into two groups. Global accuracy measures include those that count all errors in sentence production and are calculated as normalized values, e.g., the percentage of error-free clauses among all clauses (Foster and Skehan, 1996). A second group of measures is more focussed on specific types of constructions such as verb tense, third-person singular forms, prepositions, and articles, and calculate the percentage of error-free clauses with respect to these constructions (Robinson, 2006; Iwashita et al., 2008).

In the area of spoken language assessment, researchers have sought the application of measures of syntactic competence and grammatical accuracy. In particular, Halleck (1995)'s study found that in the context of English as a foreign language (EFL) assessment, holistic oral proficiency scores were highly correlated with three quantitative measures (mean length of T-units,[1] mean error-free T-unit length, and percentage of error-free T-units). Again, the results from a similar study that included both English and Japanese foreign language assessment, confirmed the utility of these and other quantitative measures that assess grammatical accuracy and syntactic complexity, in addition to vocabulary, pronunciation, and fluency (Iwashita et al., 2008; Iwashita, 2010). However, the results were inconclusive about the strength of the relationship between the measures and the proficiency scores. Strong data

[1] Hunt (1970) proposed the idea of a T-unit which is a main clause with a subordinate clause and non-clausal units. It is different from a clause since it does not consider a subordinate clause as an independent unit.

dependencies on the participant groups were reported in (Iwashita, 2010) and the discriminative ability of these quantitative measures with respect to proficiency levels was not adequate. With proficiency levels rated on a scale, the measures could only broadly discriminate students' proficiency levels, but failed to make fine-grained distinctions between adjacent levels; there were large variations within a given level and the differences between the proficiency levels were not always statistically significant. It is important to note that in all the ESL-related studies mentioned above, the measures were obtained manually.

Studies in automated speech scoring have focused on measurements of several aspects of speech production, including fluency (Cucchiarini et al., 2000, 2002), pronunciation (Witt and Young, 1997; Witt, 1999; Franco et al., 1997; Neumeyer et al., 2000), and intonation (Zechner et al., 2009). However, research on the measurements related to grammar usage is considerably nascent. Zechner et al. (2009) include a normalized language model score of the speech recognizer as a grammatical measure. This measures the similarity between word distributions in the response with that in the language model,[2] rather than the accuracy and diversity in grammatical expressions.

With the recent foray into the realm of automated assessment of spontaneous spoken language (Zechner et al., 2009, 2011) there is a concurrent need to develop quantitative measures of syntactic complexity and grammatical accuracy for use in such an environment. Two important factors govern the use of any measure in this realm. The first has to do with the discriminative ability of the measure with respect to the target. This aspect is related to the utility of the measure. The second has to do with the way in which the measures are obtained from the data. By the very nature of automated assessment, it is expected that these measures be obtained automatically.

Studies have only recently begun to actively investigate the usefulness of syntactic measures in the realm of automated scoring of spontaneous speech (Bernstein et al., 2010; Chen and Yoon, 2011; Chen and Zechner, 2011). These studies have used measures such as the average length of the clauses or sentences (Chen and Yoon, 2011) previously studied in the context of writing assessment. In addition to these length-based measures, Chen and Zechner (2011) used parse-tree based features such as the mean depth of parsing tree levels in addition to length-based measures. Using measurements from manual annotations as well as hypotheses generated by the speech recognition engine, these measures were used to predict the oral proficiency scores of the learner given his/her response. These studies found that including length-based and parse-tree based features in an automated scoring model resulted in substantially degraded performance (as compared to manually derived features from manually transcribed responses). This degradation has been attributed to the errors in automatic prediction of clause and sentence boundaries as well as those from the ASR. This raises natural questions regarding the utility of those measures for use in speech scoring.

A combination of multiple difficulties encountered in the context of processing spoken responses may help explain the issue. Most measures used in these studies were based on production units such as clauses and T-units. This being the case, the task of identifying them in speech that is naturally endowed with frequent occurrences of fragments and ellipses is naturally hard. Additionally, speech from language learners tends to include frequent grammatical errors which only increase the difficulty of identifying the units. Foster et al. (2000) listed multiple examples of difficult cases such as phrases without subjects or verbs and showed that having reliable units representing portions of speech to consistently assess features such as accuracy or complexity is not easy to accomplish. A third difficulty is due to the length of spoken responses which are typically shorter than written responses. Most measures based on sentence or sentence-like units, found to be very reliable in measuring syntactic complexity from written responses, are rendered less reliable for use in speaking tasks that elicit only a few sentences. Not surprisingly, Chen and Yoon (2011) observed a marked decrease in the correlation between syntactic measures and proficiency as response length decreased.

One is faced with even more obstacles while using syntactic complexity and grammatical accuracy measures in automated speech scoring. Spontaneous speech contains disfluencies such as repairs and repetitions, and these disfluencies need to be processed appropriately before calculating the quantitative measures. For instance, in calculating grammatical accuracy measures such as the mean length of error-free T-units, only the corrected parts found in the repairs should be considered and the repetitive parts should be excluded. This requires a high performing automated disfluency detector. However, such a tool shows suboptimal performance even with native speech. In addition, speech recognition errors only worsen the situation. Chen and Zechner (2011) showed that a moderate correlation between the score for syntactic complexity and speech proficiency (correlation coefficient = 0.49) was drastically reduced when used with automated speech recognition (ASR) outputs. This reduction in correlation was found to be due to speech recognition errors. Due to these problems, the existing syntactic complexity and grammatical accuracy measures do not seem reliable enough for being used in automated speech proficiency scoring.

In this study, we address the need for measures of grammatical ability in the sense defined by Ortega (2003) to encompass range and sophistication of surface forms in speech. The requirements are: (a) the measures should correspond to differences in proficiency levels in non-native spontaneous speech and (b) the measures should be reliable

[2] The language model was trained on non-native students' speech and broadcast news. Consequently, the grammar errors occurring in non-native students' speech may result in a good language model score, while the grammatically correct but rare expressions may result in a bad score.

for use in the context of automated speech scoring. In (Yoon and Bhat, 2012), inspired from vector-space modeling in the area of information retrieval, a new measure of grammatical competence that is based on comparing the similarity of a given response to a body of learner responses, was studied. This measure was found to be relatively robust against speech recognition errors and was reliable for use with short responses, making it usable for automated scoring of spontaneous speech in typical language assessment scenarios. In contrast to recent studies focusing on length-based and grammatical structure-based features (as outlined above), the focus was on capturing the differences in the distribution of grammatical expressions across proficiency levels.

In this paper, we extend the study in (Yoon and Bhat, 2012) and propose a new measure that is also similarity-based and derived from a language modeling approach to natural language processing. Comparing the two similarity-based measures side-by-side, we first study their degree of association with proficiency scores as an indication of their utility as measures of syntactic complexity. We then study the performance of an automatic scoring model that uses these features alongside features representing other aspects of language ability. Subsequently, we compared the performance of such a scoring model with that using conventional measures of syntactic complexity that have been found to be useful in the context of automated scoring of written language assessment.

3. Measures of syntactic complexity

With the eventual goal of automatically scoring overall spoken proficiency levels covering various aspects of speaking proficiency including grammatical accuracy, syntactic complexity, vocabulary diversity, and discourse structure, our immediate goal is to measure the grammatical competence of a given second language utterance. As will be described, conventional (or typical) measures of syntactic competence use sentence or clause length-based measures. In contrast to this, our approach will be to classify the syntactic complexity of an utterance as being similar to a particular category of learner responses (proficiency class) by constructing representative models that are score-class specific.

3.1. Measures from written language assessment

We use a set of fourteen measures of syntactic complexity studied in (Lu, 2010), found to be highly predictive of syntactic complexity indices for grading second language learner essays. Although the study points out that the results cannot be readily extended to productions with a large portion of grammatically incomplete sentences (such as those produced by beginner-level learners), we chose these measures in order to get a baseline since no other measures of syntactic complexity for spoken utterances are as yet available in related studies.

In (Lu, 2010), these measures of syntactic complexity were chosen based on the literature pertaining to second language development studies. They represent a fairly complete picture of the repertoire of measures that second language development researchers draw from. They can be categorized into five types and are listed in Table 1.[3]

With the goal of having measures that normalize the effect of the varying length of utterances, we choose only the ratio-based measures (types II, III, IV and IV) and exclude measures of the first type.

3.2. Proposed measures

Our proposed syntactic complexity measures have the following two characteristics. First, in contrast to most methods that consider scores of syntactic complexity as a combination of measures of the length of sentence-based units or frequency of specific grammatical structures, we directly measure students' sophistication and range in grammar usage based on the distribution of syntactic constructions.

Second, instead of rating a learner's response using a scale based on native speech production, our experiments compare it with a similar body of learners' responses. Considering the variety of grammatical structures that native speakers produce, a comparison of learners' constructions to that of native speakers implies searching in a large space of possible constructions. We instead collected a large amount of learners' spoken responses and classify them into four groups according to their proficiency level (as assigned by professional raters). We then examined how distinct the proficiency classes were based on the distribution of part-of-speech (POS) tags. We hypothesized that the level of acquired grammatical proficiency is signaled by the distribution of the POS tags. Hence, given a student's response, we obtained its similarity with score-specific models that capture the distribution of POS tags in each score level. The intuition here is that the similarity of POS distribution of a response to that of the representative response of a given score class is governed by the proficiency level of the response.

Working with POS tags as features has the advantage of a much lower dimensionality than would be needed when using lexical information. Moreover, rather than focusing on specific grammatical constructions, we utilize the more fine-grained information at the level of sequences of POS tags. This in turn provides us a convenient way to counter the effect of differences in topics in speaker responses, as would be the case in several scenarios of practical importance.

[3] The study further defines the various production units and syntactic structures used in the measures and since we feel that the definitions and the associated details are beyond the scope of the current study, we omit them from our discussion and direct the interested reader to the paper for more details.

Table 1
Conventional measures of syntactic complexity found useful in written language assessment.

Type	Measures
I	Mean length of clauses (MLC), mean length of sentences (MLS), and mean length of T-units (MLT)
II	Sentence complexity ratio (clauses per sentence, or C/S)
III	T-unit complexity ratio (clauses per T-unit, or C/T), complex T-unit ratio (complex T-units per T-unit, or CT/T), dependent clause ratio (dependent clauses per clause, or DC/C), and dependent clauses per T-unit (DC/T)
IV	No. of coordinate phrases per clause (CP/C), No. of coordinate phrases per T-unit (CP/T), and, sentence coordination ratio (T-units per sentence, or T/S)
V	No. of complex nominals per clause (CN/C), No. of complex nominals per T-unit (CN/T), and, No. of verb phrases per T-unit (VPT)

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 (e.g. 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), as features based on the frequency and types of errors. More recently, Roark et al. (2011) found that the POS tag distribution effectively captured the differences in syntactic complexity between normal subjects and subjects with mild cognitive impairment. Prompted by the utility of POS tag distributions in these areas, our approach is to approximate the differences in syntactic complexity between proficiency levels by the differences in POS tag distributions.

Inspired by the approach to word-level modeling in the area of information retrieval (IR) and language models in the area of natural language processing, we model the POS tag distributions in two ways:

1. a POS-based vector space model representation and
2. a POS n-gram language model.

3.3. POS-based vector space model

We resort to the vector-space model (VSM) (Manning et al., 2008) of representing a document (used in IR) and arrive at an analogous vector-space model to represent the proficiency classes of the learner corpus. We begin by representing each proficiency score level as a document whose term vector is formed from the available POS tags (or combinations thereof) of the responses in that score level. As in the case of IR, each term is weighted by the appropriate term frequency-inverse document frequency (tf-idf). The results is four vectors, one per score-level. Such a score-category-based VSM has been used in automated essay scoring to assess the lexical content of an essay by comparing the words in the test essay with the words in sample essays from each score category (Attali and Burstein, 2006). We extend this idea to the assessment of grammar usage using vectors of POS tags. A given test response is treated like a query and converted to a vector on the coordinates defined by the chosen terms of the representative vector of proficiency score classes. Given a test response in its vector form, it is assigned to a score class that has a similar POS distribution, where the similarity is captured by the cosine similarity function – the dot-product of the two normalized vectors.

Cosine similarity is often used to identify documents that are relevant for a given query. This measures the similarity between a given query and a document as the cosine of the angle between the corresponding vectors in a high-dimensional space, where each term in the query (and documents) corresponds to a unique dimension. If a document is relevant to the query, it shares many terms resulting in a small angle. In this study, a term was a single or compound POS tag (unigram, bigram or trigram) weighted by its tf-df, and the document was the response under consideration.

For a given response, we calculated the cosine similarity values of the response with the representative vector of each score level. As an example, cos_1 is the cosine similarity value of the test response to the representative vector of score level 1. Thus, the feature cos_1 captures the similarity of a given response that of score level 1. A total of four similarity values, one with each score level, were calculated. In addition, *cosmax*, the score-level that is most similar to the response is also obtained.

3.4. POS language models

The second approach to modeling the POS distribution of a proficiency level is based on multiple POS language models (LM). Multiple POS LMs with a minimum cross-entropy criterion have been found useful in language identification studies such as (Zissman, 1996), and more recently in (Yoon and Higgins, 2011) for automated speech scoring.

The human-scored responses in the training set are classified into four groups according to their proficiency scores, and score-specific POS n-gram LMs are trained. The key assumption here is that the likelihood of a response given the score specific language model is a measure of the similarity of the response to the set of responses in the score class. Accordingly, the log-likelihood of each response given a score-specific LM constitutes the similarity feature. The resulting similarity value was normalized by the length of the response.

3.5. Features capturing POS distribution

For a given response, a total of ten features (five features per model) are generated. The list of features with a brief description is as follows:

- cos_i: cosine similarity value of the test response with the representative vector of score level $i = 1, \ldots, 4$.
- *cosmax*: the score level with the highest similarity score given the response.
- lm_i: logprob (likelihood) of the LM of score level $i = 1, \ldots, 4$.
- *lmmax*: the score level of the LM with the maximum logprob given the response.

We assume that the score-level having the highest cosine similarity (or log-likelihood) value for a test response is most similar to the underlying syntactic complexity of the test response. In contrast to cos_i, and lm_i which have continuous values, the two *max* features have an advantage in that they can be directly interpreted as the proficiency level of the given response, based on its syntactic complexity.

3.6. Comparison of two models

In VSM, the distributional differences among different score levels are directly modeled through the use of *idf*. Proficient speakers use a set of grammatical expressions potentially more sophisticated than that of beginners, while beginners use simple expressions and sentences with frequent grammatical errors. POS tags (or sequences) capturing these expressions may be seen in corresponding proportions in each score group. The POS tags (or tag sequences) which occur only in specific score levels get a high *idf* value and accordingly have higher impact on the model. On the contrary, those tags which commonly occur across all score levels get a low *idf* value and have a smaller impact on the model. As a result, the inherent differences between score levels are captured in the score-specific VSMs. Similar to the language modeling approach to IR, the LM approach here directly models the idea that a score level is a good match to a response if the score-specific POS language model is most likely to generate the syntactic structure of the response which will in turn happen if the score level contains the grammatical constructions in the response more often.

In a sense, the VSM attempts to model the global characteristics of the POS tag distribution for each score class. The tf-idf weight of terms ensures that only those POS tags relevant to characterizing a score-level are represented. Thus, the representative vector of a score-class in the VSM model looks at the overall distribution of those characteristic POS tags (or tag sequences) and, in some sense, is a global approach.

On the other hand, an LM based model of a score class captures the similarity of a response to a proficiency class based on local features or POS distribution patterns that manifest at the bigram or trigram levels without considering whether the term is relevant to the score-level or not.

4. Implementation details

The transcriptions of the responses were generated manually as well as using a speech recognizer as will be described in Section 4.1. All transcriptions were tagged using the POS tagger described in Section 4.2 and POS tag sequences were extracted. This was followed by the model generation phase where the POS-VSM models and the POS-LM models were obtained from the training data with different sets of tags. Finally, the features were generated for the responses in the test dataset.

4.1. Automatic speech recognizer

An HMM recognizer was trained using approximately 733 h of non-native speech collected from 7872 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 27% on the held-out test dataset was observed.

4.2. POS tagger and model training data preparation

POS tags were generated using the POS tagger implemented in the Open-NLP toolkit.[4] It was trained on the Switchboard (SWBD) corpus. This POS tagger was trained on about 528 K word/tag pairs and achieved a tagging accuracy of 96.3% on a test set of 379 K words. A combination of 36 tags from the Penn Treebank tag set (Marcus et al., 1993) and 6 tags generated for spoken languages were used in the tagger. The POS-tagger was then used to tag the transcribed responses of the training dataset. All responses in the same score level were concatenated into a single score-specific training set, and POS sequences were extracted.

4.3. Compound unit generation using mutual information

Temple (2000) pointed out that proficient learners' speech is characterized by increased automaticity in speech production. These speakers tend to memorize frequently used multi-word sequences as chunks and retrieve the whole chunks as single units. We capture the degree of automaticity by the frequently co-occurring POS sequences. We identified these co-occurring POS tag sequences as those having a high mutual information and included them in our list of terms.

POS bigrams with high mutual information were selected and used as single units. First, all POS bigrams which occurred less than 50 times were filtered out. Next, the remaining POS bigrams were sorted by their mutual information scores, and two different sets (top50 and top100) were selected. Initially, we planned to select 50 bigrams for top50 and 100 bigrams for top100, but we ended up to select 51 bigrams for top50 and 108 bigrams for top100 due to bigrams with tied mutual information scores. The selected POS bigrams were then transformed into compound tags and used in conjunction with the original Penn Treebank tag set. As a result, we generated three

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

sets of POS tags: the original POS set without the compound unit (Base), the original set with 51 compound units (Base + mi50), and the original set with 100 compound units (Base + mi100). The number of tags was 42 for Base set, 93 for Base + mi50 set, and 150 for Base + mi100 set.

4.4. Building VSMs

Unigram, bigram, and trigram tags were used in VSM building. For each *n*-gram, three sets of VSMs (using the three tag sets presented in Section 4.3 as terms), yielding four score-specific vectors (one per score-class), were generated, resulting in a total of nine VSMs. For each VSM we have $cos_i, i = 1, \ldots, 4$ and *cosmax* as features, with cos_i taking values in $[0, 1]$ and *cosmax* taking values in $\{1, 2, 3, 4\}$. The results were based on the each *n*-gram model separately and we did not combine any models.

4.5. Training of POS n-gram LM

Unigram, bigram, trigram, 4-gram and 5-gram LMs were used in LM building. For each *n*-gram, three sets of LMs (using the three tag sets) were trained for each score level using score-specific training data. We used the SRILM toolkit (Stolcke, 2002) for training the models with Witten–Bell smoothing.[5] For each *n*-gram and tag set combination, we obtained $lm_i, i = 1, \ldots, 4$ and *lmmax* as features. While lm_i takes values in $(-\inf, 0], lmmax$ takes values in the set $\{1, 2, 3, 4\}$. As with the VSMs, the results were based on the each *n*-gram model separately and we did not combine any models.

5. Data

The data used in this study, was a proprietary collection of responses from the Test of English as a Foreign Language® internet-based test (TOEFL® iBT), an international English language assessment required for studying in an English medium university environment. The assessment consisted of six items per speaker, where they were prompted to provide responses lasting between 45 and 60 s per item, resulting in approximately 5.5 min of speech per speaker. Among the six items, two items were “independent items” that asked examinees to provide information or opinions on familiar topics based on their personal experience or background knowledge. The four remaining items were “integrated items” that required reading or listening in addition to speaking skill. Test takers read and/or listen to some stimulus materials and then respond to a question based on them. All items extracted spontaneous, unconstrained natural speech. While there were important differences between these task types from an assessment point of view (i.e. whether the item required listening and/or reading skills), they were not differentiated in this study, owing to the fact that both items elicited unconstrained speech and did not have specific target grammar expressions.

Approximately 50,000 responses were collected and split into two datasets: the ASR set and the scoring model training/test (SM) set. The ASR set comprised of about 733 h of speech and was used for ASR training and POS similarity model training. The SM set comprised of 44 h of speech and was used for feature evaluation and automated scoring model evaluation. There was no overlap in items between the ASR set and SM set. Each response was rated for proficiency by trained human scorers using a 4-point scoring scale, where 1 indicated low speaking proficiency and 4 indicated high speaking proficiency. The speaker, item information, and distribution of proficiency scores are presented in Table 2.

As seen in Table 2, there is a strong bias towards the middle scores (score 2 and 3) with approximately 83–84% of the responses belonging to these two score levels. Although the skewed distribution of students at the low and high score levels increased the difficulty of feature development and model training, we used the data without modifying the distribution since this constitutes a typical distribution of proficiency scores in a large-scale language assessment scenario. The mean scores in the two datasets were 2.67 and 2.63 respectively, and the overall score distribution of the two datasets was comparable.

773 Responses in ASR set and 4 responses in SM set were not scoreable due to sub-optimal response characteristics such as technical difficulties (e.g, equipment or transmission errors, loud background noise) and lack of student’s response. These problematic responses were removed from the data as a result of which the full set of 6 responses was not available for some speakers. The average number of responses per speaker for both data sets was more than 5.9 suggesting that these failures were relatively minor.

In order to evaluate the reliability of the human ratings, approximately 5% of the ASR set (a total of 2388 responses) was scored by two raters. Both the Pearson correlation coefficient *r* and weighted-kappa *k* were 0.62 indicating the level of subjectivity in the task of proficiency scoring.

Table 3 presents the descriptive analysis of the length of production units for each score point. For the analysis, we used a total of 200 responses from TOEFL Practice Online (TPO), an online practice test which allows students to gain familiarity with the format of our main data (TOEFL). Thus the responses are similar to those in our datasets. The manual transcriptions were annotated for locations of clause boundaries (hereafter, CBs) and disfluencies such as repetition, repair, and sentence fragments by two annotators with a linguistics background.[6] Disfluencies frequently occur in non-native speakers’ spontaneous speech

[5] This smoothing was chosen to accommodate for the non-zero frequency of POS tags which prevents the application of Good-Turing smoothing.

[6] Approximately 15% of data was double annotated; κ was 0.95 for CBs and 0.75 for disfluencies.

Table 2
Data size and score distribution.

Data set	No. of responses	No. of speakers	No. of items	Score		Score distribution			
				Mean	SD	1	2	3	4
ASR	47,227	7872	24	2.67	0.73	1953 4%	16,834 36%	23,106 49%	5,334 11%
SM	2,876	240	12	2.63	0.75	141 5%	1132 39%	1263 44%	340 12%

Table 3
Means and standard deviations of the number of production units for each score point.

Level	*N*	Words per response		CB per response		T-units per response	
		M	SD	M	SD	M	SD
1.00	50.00	41.78	27.56	5.38	3.58	3.20	2.01
2.00	50.00	71.38	31.48	7.38	3.31	4.86	2.18
3.00	50.00	85.28	32.00	9.22	3.84	5.50	2.43
4.00	50.00	94.22	38.98	9.48	3.93	5.94	2.97

and they should not be considered to be parts of a sentence. This is because their inclusion will result in inflated lengths of production units. Therefore, disfluencies were removed, based on manual annotations. Secondly, the transcriptions were segmented into clauses, T-units, and sentences based on the human annotations. Finally, the means and standard deviations of each production unit were calculated for each score level.

6. Experiments

As such, we need a training set to obtain the underlying score-specific models which we then use to generate the feature values using responses from the evaluation set. Towards this end, we used the data in two modes. In the manual mode, we used the manual transcriptions (henceforth termed *manual*) and in the automated mode (henceforth termed *auto*), we use the output of the speech recognizer. This set-up allowed for an understanding of the impact of ASR on the feature generation process as well on the feature performance. Finally, we use the features obtained in the automated mode in a multiple regression scoring model and compare its performance with that using conventional features of syntactic complexity (also obtained in the automated mode).

6.1. Dataset combination

The experiments were conducted with the following combinations of datasets. We vary the nature of the transcriptions available (auto or manual) in the training and test sets, yielding the following (train, test) combinations for experimentation. Such a split in the dataset, permits us to make the following observations on the performance of the proposed features in various operating scenarios.

- Set 1: (Manual, Manual) – in this mode, we have the best case performance of the model being considered, with manual transcriptions available for both training and evaluation.
- Set 2: (Manual, auto) – in this set-up we observe the effects of ASR output on the performance of a model that is trained with manually transcribed data and tested on the ASR output.
- Set 3: (Auto, auto) – this set-up provides the level of performance in a true operational scenario where it is difficult to obtain manually transcribed responses.

6.2. Association between proposed measures, automatic scores and human ratings

We determine the utility of the features by computing each feature's correlation with human-assigned proficiency scores as has been done in prior related studies. These features are then used in a multiple regression scoring model (as studied in Zechner et al. (2009)) and the resulting automatic scores are compared with the human-rated scores. As a baseline, we use the model studied in Zechner et al. (2009), which is then augmented with the features we propose in this study as well as features using prototypical measures of syntactic complexity. We then compare the resulting scoring models based on the correlations of the predicted values with that of the human scores.

6.3. Factors for practical consideration

A primary requirement of automated scoring systems that make measurements on automatically recognized spoken responses is that the measures be immune to the ASR errors inherent in the process. The utility of a set of automatically derived measures is better analyzed in the context of a parallel analysis that considers measurements obtained using transcriptions obtained manually and automatically. The results will aid the assessment of the overall utility of the measures with respect to a change in the mode of measurement. Our experiments and analyses will reflect this factor of practical significance.

Another factor of importance from a practical standpoint is that of generalizability. In this respect, we are interested in seeing the extent to which the size of the training data affects model performance. Along the lines of generalizability, we also explore the extent to which our method is applicable to scoring items unseen in the training set.

6.4. Comparison with prototypical measures from written language assessment

We compare the utility of the proposed measures with that of the most representative syntactic complexity measures used in (Lu, 2010) and described in Section 3.1. As rationalized in Section 2.1, there is reason to believe that the sentence-length based measures are inherently unsuitable for use with automatic spoken language assessment. In order to test this, the performance of our features with the features of the syntactic complexity proposed in (Lu, 2010) were compared. Towards this, the clause boundaries of the ASR hypotheses were automatically detected using the automated clause boundary detection method used in (Chen and Zechner, 2011).[7] The utterances were then parsed using the Stanford Parser (Klein and Manning, 2003), and a total of 14 features including both length-related features and parse-tree based features were generated using (Lu, 2012). Finally, the Pearson correlation coefficients between these features and human proficiency scores were calculated.

7. Results

7.1. POS features

We will first review the results of VSM-based features, followed by LM-based features.

The VSMs and the features were obtained using both modes (manual and automated). Table 4 shows feature-score correlations for the features *cosmax* and cos_4 (features cos_1, cos_2 and cos_3 were excluded since they showed lower correlation with human scores and were highly correlated with cos_4). From the table, we make the following observations:

- The most correlated feature with human scores of proficiency was cos_4 that used the *base* set with *bigrams*. It achieved the best correlation of 0.43 when manual transcriptions were used in model building as well as in evaluation (Set 1). The drop in score-feature correlation when used with ASR output (Set 2 and Set 3) was not statistically significant.
- Overall, cos_4 not only outperformed *cosmax*, but also was more robust against ASR errors.
- Bigram-based features outperformed both unigram-based and trigram-based features.
- The inclusion of compound tags (refer Section 4.3) did not result in an increased correlation for cos_4. However, it increased the correlation (statistically significant at level 0.01) in the case of *cosmax*, when obtained from unigrams.

The unigrams had good coverage but limited power in distinguishing between different score levels. Their ability to distinguish between levels was augmented with the inclusion of co-occurring POS tags. On the other hand, trigrams had the opposite characteristics – they captured more structure, but did not have good coverage because of data sparseness. Bigrams seemed to strike a balance in both coverage and complexity (from among the three LMs considered here) and may thus have resulted in the best performance with both the features in both manual and ASR modes.

It is worthwhile to emphasize that the performance of ASR-based features was comparable to that of transcription-based features. The best performing feature among ASR-based features was using the bigram and *base* set, with correlations nearly the same as the best performing feature among the transcription-based features. Seeing how close the correlations were in the case the manual transcription-based and ASR-hypothesis based features, we conclude that the proposed measure is robust to ASR errors.

Table 5 shows correlations between the LM-similarity features and expert-rated proficiency scores for experiments with *n*-grams $n = 1, \ldots, 3$. Since no improved correlation was observed by increasing *n*-gram size beyond trigrams, 4-gram and 5-gram results are excluded from this table. Additionally, features lm_1, lm_2 and lm_3 were excluded since they showed lower correlation with human scores and were in turn highly correlated with lm_4. From the table, we make the following observations.

- The feature most correlated with human scores was *lmmax* using the *base* set with *bigrams*. It achieved a correlation of 0.38 when both model and evaluation were based on manual transcriptions (Set 1). There was a substantial drop in correlation when used with the ASR output; the correlation for Set 3 (both ASRs) was 0.33 (a 0.05 drop in absolute correlation).
- As expected, features based on set 1 of the dataset combinations (manually transcribed training and evaluation data) outperformed those obtained using Set 2 and Set 3. However, when used with ASR output, Set 3 was better than Set 2. In Set 2, there was a

[7] The automated clause boundary detection method in this study was a Maximum Entropy Model based on word bigrams, POS tag bigrams, and pause features. The method achieved an F-score of 0.60 for the non-native speakers' ASR hypotheses. A detailed description of the method is presented in (Chen and Yoon, 2011).

Table 4
Pearson correlation coefficients between VSM-based features and expert proficiency scores. Here we consider the features cos_4 and *cosmax* in the VSM model. All correlations are significant at the 0.01 level.

Feature	Set	Unigram			Bigram			Trigram		
		Base	Base + mi50	Base + mi100	Base	Base + mi50	Base + mi100	Base	Base + mi50	Base + mi100
cosmax	(M, M)	0.08	0.18	0.18	0.34	0.33	0.34	0.34	0.33	0.34
	(M, A)	0.12	0.17	0.17	0.26	0.24	0.26	0.25	0.26	0.26
	(A, A)	0.13	0.18	0.19	0.30	0.30	0.30	0.31	0.28	0.27
cos4	(M, M)	0.30	0.30	0.33	**0.43**	0.36	0.37	0.40	0.32	0.30
	(M, A)	0.25	0.27	0.30	**0.42**	0.35	0.35	0.37	0.31	0.28
	(A, A)	0.30	0.27	0.30	**0.41**	0.32	0.32	0.34	0.28	0.26

The values are in bold to indicate that they correspond to the best performance.

Table 5
Pearson correlation coefficients between LM-based features and expert proficiency scores. All correlations are significant at the 0.01 level.

Feature	Set	Unigram			Bigram			Trigram		
		Base	Base + mi50	Base + mi100	Base	Base + mi50	Base + mi100	Base	Base + mi50	Base + mi100
lmmax	(M, M)	0.31	0.32	0.32	0.38	0.34	0.34	0.36	0.26	0.25
	(M, A)	0.26	0.28	0.28	0.29	0.28	0.29	0.29	0.24	0.26
	(A, A)	0.30	0.31	0.31	0.33	0.32	0.3	0.29	0.29	0.29
lm4	(M, M)	0.06	0.05	0.07	0.19	0.17	0.18	0.23	0.19	0.2
	(M, A)	0.05	0.03	0.02	0.15	0.15	0.14	0.18	0.19	0.17
	(A, A)	0.06	0.04	0.05	0.16	0.12	0.14	0.18	0.14	0.15

discrepancy in the train/evaluation data condition; the train data was based on the manual transcription while the evaluation data was based on the ASR. This discrepancy would have resulted in the additional performance drop.

- Feature *lmmax* outperformed lm_4 overall.
- Bigram-based features using the *base* tag set showed better correlation than the other *n*-gram features. However, for lm_4 the 4-gram showed the best correlation (though not too different from trigram based lm_4), with different tag sets and dataset combinations.
- The inclusion of compound tags did not result in an increased correlation for lm_4.

The LM-based features behaved differently from VSM-based features. First, the features were more susceptible to ASR errors, and there was a substantial performance drop in the best performing features when used with the ASR outputs. Furthermore, from Tables 4 and 5 we observe that the feature cos_4 was the better performing feature in VSM method and the feature *max* was the better performing in LM method.

7.2. Comparison of features

Table 6 shows the performance of the prototypical measures of syntactic competence from written language assessment.

We observe that these measures have lower correlations with human-scores compared to our proposed features. We then include the features in a scoring model for subsequent performance comparison.

As seen in the previous section, cos_4 emerged as the feature of choice when using the VSM, and *lmmax* the feature of choice when using the POS-LM. The two features seem to represent different aspects of grammatical competence. What will follow next, will be a comparison of the two features.

- In the case of cos_4, the feature, in a way, captures how far the given POS tag distribution is with respect to the representative vector of score-level 4 in a high-dimensional space. While this is useful, the feature value only gives the similarity with the score-level. Unlike the feature cos_4, *lmmax* gives a score-level value to a test-response, which may be construed as a score of grammatical competence.
- Although somewhat unrelated to the correlation with human-assigned proficiency scores, the relative ease of interpretation seems to be an advantage of the POS-LM model. However, taking into account the correlation with human judgment, the VSM model seems to be a better option.

We list the correlation coefficients between the conventional measures of syntactic complexity and the scores of proficiency in Table 6. From Table 6, we note that the best performing feature is “DCC” (mean number of dependent clauses per clause) and the correlation *r* was 0.14. In addition, “DCT” had a correlation that was also a statistically significant, but the correlation *r* was even weaker than “DCC”. Our best performing feature (bigram-based cos_4) *widely outperformed* the best of Lu (2010)’s features (with correlations approximately 0.3 apart).

Table 6
Pearson correlation coefficients between the prototypical measures of syntactic complexity and expert proficiency scores for the ASR output. Only correlations marked (*) were significant at $\alpha = 0.01$ level.

C/S	C/T	CT/T	DC/C	DC/T	CP/C	CP/T	T/S	CN/C	CN/T	VPT
0.014	0.006	0.017	0.138*	0.060*	0.015	0.026	0.027	0.028	0.031	−0.008

For the purpose of comparing the features and the models side-by-side, we chose the best features (with the highest correlations with human judgments) from the models studied here (prototypical features from writing assessment, VSM-based and POS-LM based models). For the sake of completion, we included both cases (the models trained and evaluated with manual transcriptions as well as models trained and evaluated with automatic transcriptions). We observed that each of these methods had an associated best performing feature – the feature cos_4 was the better performing feature in the VSM and the feature *max* was the better performing in POS-LM. The comparison of features is summarized in Table 7 below.

A plausible explanation for the poor performance of Lu (2010)'s features is that the features were generated using a multi-stage automated process, and the errors in each stage propagated to the next, resulting in a low feature performance. For instance, the errors in the automated clause boundary detection may result in a serious drop in the performances. With the spoken responses being particularly short (a typical response in the dataset had 10 clauses on average), even one error in clause boundary detection can seriously affect the reliability of the features.

7.3. Effect of the training data size on model performance

For the purpose of VSM-based and LM-based POS-model constructions, we used approximately 47 K responses, which may be unavailable in many practical scenarios. To understand the effect of training data size on the resulting POS-model, we consider varying the available data size for POS-model building in the case of the better performing VSM-based scenario. VSM-based models are built using variable-sized training data of manual transcriptions (sampled to preserve the underlying score distribution) and the Pearson's correlation coefficients of the feature cos_4 (derived from ASR-based transcriptions of the responses in the SM data) with human scores of proficiency for every sample are noted. From the results tabulated in Table 8 we notice that even with 500 responses sampled from the training data, we are able to achieve the feature-score correlation of 0.41, which approaches that shown by using a model that was trained on 47 K responses.

Table 7
Comparison of feature-score correlations.

Study	Feature	Condition	Correlation
(Lu, 2010)	DCC	Trans	0.14
		ASR	0.14
Current study	cos_4	Trans	0.43
		ASR	0.41
Current study	*lmmax*	Trans	0.38
		ASR	0.33

In order to further highlight the generalization performance of proposed models, it may be worth noting that since the test data had no item-overlap with the training set, the feature generalizes well to unseen items.

7.4. Automatic scoring model comparison

The effect of the inclusion of the proposed features is best understood by studying them in an automatic scoring model. We consider a multiple linear regression scoring model that approximates the human scores by a linear combination of the proposed features that represent various constructs of oral language proficiency. We first build and compare stand-alone multiple regression models of syntactic complexity. We then compare scoring models that include measurements of syntactic complexity in addition to the other constructs of proficiency currently being measured in the state-of-the-art scoring model (SpeechRater). For the purpose of this comparison, we perform 5-fold cross validation on the SM data described in Section 5 ($N = 2876$) using a training set (80%) and a held-out test set (20%) to parametrize the models and to assess the models' performance respectively in every run. The results are averaged over the 5 runs.

7.4.1. Scoring model for syntactic complexity

Although the human assigned scores of proficiency are holistic scores of overall proficiency, in the absence of stand-alone scores of syntactic complexity, we assume that the scores of proficiency are well correlated with scores of syntactic complexity. Under this assumption, we first construct automatic scoring models for syntactic complexity

Table 8
Comparison of cos_4-score correlations, r, by varying training data size (number of responses) for VSM-based POS-model training.

Data size (in 1000 s)	0.1	0.2	0.5	1.0	2.4	4.7	9.4	24	47
r	0.35	0.38	0.41	0.42	0.41	0.42	0.41	0.42	0.42

using the three sets of features studied here – the prototypical measures from written language assessment, the VSM-based features and the LM-based features. The grammar scoring model using only the prototypical features described in Section 3.1 will serve as a baseline (**GramLu**). We then compare scoring models using VSM- and LM-based features (**GramVSM** and **GramLM** respectively).

We avoid collinearity in the scoring models by excluding correlated features ($r \geqslant 0.8$). The resulting **GramLu** model includes the features CN/C, CN/T, CP/C, C/T, T/S, CT/T, **GramVSM** includes *cosmax* and cos_4 while **GramLM** includes *lmmax* and lm_4. The agreement between predicted scores and human scores in terms of Pearson's correlation coefficient (unrounded and rounded) as well as weighted kappa are tabulated in the Table 9.

From the Table 9, we observe that the correlation is highest for a scoring model using VSM-based features and that the correlation is considerably higher (0.36) than that of the baseline model using conventional features for written language assessment. The correlations of the scoring model using LM-based features is also reasonably better than the baseline (0.24 higher). The agreement results of **GramLu** here is in line with the results in (Chen and Zechner, 2011), where it was observed that measures of syntactic complexity were highly sensitive to errors in ASR and are hence not reliable for use in speech scoring.

7.4.2. Scoring model for oral proficiency

We next consider an evaluation of the syntactic complexity features studied here in an augmented scoring model that includes features from the SpeechRater that represent constructs of fluency, pronunciation, vocabulary and grammar. Towards this, we again consider a multiple regression model that includes the features global normalized HMM acoustic model score (**amscore**), speaking rate (**wpsec**), types per second (**tpsecutt**), average chunk length in words (**wdpchk**) and global normalized language model score (**lmscore**) (Zechner et al., 2009) by themselves (**Base**) as also alongside the features considered in this study. Such an arrangement permits us to not only compare the relative gains in scoring model performance compared to the base model but also to assess the relative importance of the features considered in the model.

Table 9
Comparison of scoring model performances only using features for grammar competence. All values are significant at level 0.01 except the one marked with (*).

Evaluation method	GramLu	GramLM	GramVSM
Weighted κ	0.05	0.25	**0.33**
Correlation (unrounded)	0.06	0.30	**0.42**
Correlation (rounded)	0.04*	0.23	**0.34**

The values are in bold to indicate that they correspond to the best performance.

Here, in addition to a base model with the features found in SpeechRater, we study augmented scoring models using the three sets of features studied here – the prototypical measures from written language assessment, the VSM-based features and the LM-based features. The augmented models will be denoted by **SynLu**, **SynVSM**, and **SynLM** respectively. In addition, we will consider a combination model **SynAll**, that will include all features. The correlations (unrounded and rounded) as well as the weighted kappa for the models studied here are found in Table 10.

Comparing the agreement results in Table 10 we notice that the model **SynVSM** shows the highest (0.607) correlation of predicted scores with human scores among **Base**, **SynLu**, **SynLM** and **SynVSM**. Moreover, testing for the significance of the difference between two dependent correlations using Steiger's Z-test, we notice that the difference in correlations (Base–SynVSM, SynLu–SynVSM and SynLM–SynVSM) are statistically significant at level = 0.01. Combining all features in the **SynAll** model, although the correlation appears to be marginally higher than that with **SynVSM** the difference is not statistically significant. Thus, inclusion of measures of syntactic complexity shows improved correlation between machine and human scores compared to the state-of-the-art model (here, **Base**).

The agreement between human and machine scores for the scoring model **SynVSM** (Pearson correlation coefficient 0.607) may not seem impressive by itself, but taken in the context of the subjectivity of the task of assigning a proficiency score (as mentioned in Section 5), the correlation coefficient is seen to approach near-human agreement of 0.62.

Our next focus is the relative importance of the features in an overall scoring model such as **SynVSM** that includes all available features for measuring the various constructs of oral proficiency. In Table 11 we list the relative importance of the predictors, as the R^2 contribution averaged over orderings among predictors (Grömping, 2006). We notice that the feature cos_4 is the *third* most important predictor of oral proficiency next only to *tpsecutt* and *wdpchk* which are existing features in the SpeechRater module.

Table 10
Comparison of scoring model performances using features of syntactic complexity studied in this paper along with those available in SpeechRater. Here, **Base** is the scoring model using just the features in SpeechRater. All correlations are significant at level 0.01.

Evaluation method	Base	SynLu	SynLM	SynVSM	SynAll
Weighted κ	0.540	0.540	0.560	0.560	0.570
Correlation (unrounded)	0.587	0.589	0.595	0.607	0.610
Correlation (rounded)	0.500	0.503	0.509	0.526	0.522

Table 11
Comparison of the relative importance of the features in the augmented scoring model **SynVSM**.

Feature	Relative importance (%)
tpsecutt	35.03
wdpchk	23.90
cos_4	**16.29**
amscore	9.64
wpsec	6.38
cosmax	5.92
lmscore	2.81

The relative importance in bold highlights the fact that the proposed new feature (cos_4) is relatively important compared to the other features available in prior studies.

8. Discussion

The performance of the features considered in the experiment may be viewed in the light of the following case-wise analyses of the VSM-based features and the LM-based features.

8.1. VSM-based similarity assessment

The measure of syntactic competence that we studied here may be viewed as a simplified version of overall syntactic competence, without the consideration of specific constructions. We analyze the results further with the intention of casting light on the level of detail of syntactic competence that can be explained using our measure. Furthermore, we will show that bigram POS sequences can yield significant information on the range and sophistication of grammar usage in the specific assessment context (spontaneous speech with only declarative sentences).

ESL speakers with high proficiency scores are expected to use more complicated grammatical expressions that result in a high proportion of POS tags related to these expressions in that score group. The distribution of POS tags was analyzed in detail in order to investigate whether there were systematic distributional changes according to the proficiency levels. For illustration purposes, we restrict our discussion to the analysis using unigrams (base and compound tags). For each score group, the POS tags were sorted based on the frequencies in training data and the rank orders were calculated. The more frequent the POS tag, the higher its rank.

A total of 150 POS tags, including the original POS tag set and top 100 compound tags, were classified into 5 classes:

- Absent-in-low-proficiency (ABS): The group of POS tags that appeared in all score groups except the lowest proficiency group.
- Increase (INC): The group of POS tags whose ranks increased consistently as proficiency increased.
- Decrease (DEC): The group of POS tags whose ranks decreased consistently as proficiency increased.
- Constant (CON): The group of POS tags whose ranks remained same despite change in proficiency.
- Mix: The group of POS tags of with no consistent patterns in the ranks.

Table 12 presents the number of POS tags in each class.

The 'ABS' group mostly consists of 'WP' and 'WDT'; more than 50% of tags in this group are related to these two tags. 'WP' is a Wh-pronoun while 'WDT' is a Wh-determiner. Since most sentences in our data are declarative sentences, 'Wh' phrases signal the use of relative clauses. Therefore, the lack of these tags strongly support the hypothesis that the speakers in score group 1 were not proficient in the use of relative clauses or their use in limited situations.

The 'INC' group can be classified into three sub-groups: verbs, comparatives, and relative clauses. The verb group includes infinitive (TO_VB), passive (VB_VBN, VBD_VBN, VBN, VBN_IN, VBN_RP), and gerund forms (VBG, VBG_RP, VBG_TO). Next, the comparative group encompasses comparative constructions. Finally, the relative clause group signals the presence of relative clauses. The increased proportion of these tags reflects the use of more complicated tense forms and modal forms as well as more frequent use of relative clauses. It supports the hypothesis that speakers with higher proficiency scores tend to use more complicated grammatical expressions.

The 'DEC' group can be classified into five sub-groups: noun, simple tense verb, GW and UH, non-compound, and comparative. The noun group is comprised of many noun or proper noun-related expressions, and their high proportions are consistent with the tendency of less proficient speakers to use nouns more frequently. Secondly, the simple tense verb group is comprised of the base form (VB) and simple present and past forms (PRP_VBD, VB, VBD_TO, VBP_TO, VBZ). Note that the expressions in these sub-groups are simpler than those in the 'INC' group.

The 'UH' tag is for interjection and filler words such as 'uh' and 'um', while the 'GW' tag is for word-fragments. These two spontaneous speech phenomena are strongly related to fluency, and they signal problems in speech production. Frequent occurrences of these two tags are evidence of frequent planning problems and their inclusion in the 'DEC' class suggests that instances of speech planning problems decrease with increased proficiency.

Tags in the non-compound group, such as 'DT', 'MD', 'RBS', and 'TO', have related compound tags. The non-compound tags are associated with the expressions that do not co-occur with strongly related words, and they tend to be related to errors. For instance, the non-compound 'MD' tag signals that there is an expression in which a modal verb is not followed by 'VB' (base form) and as seen

Table 12
Tag distribution and proficiency scores.

ABS	INC	DEC	CON	Mix
14	37	33	18	48

in the examples, *‘the project may can change’* and *‘the others must can not be good’*, they are related to grammatical errors.

Finally, the comparative group includes ‘RBR_JJR’. The decrease of the tag ‘RBR_JJR’ is related to the correct acquisition of the comparative form. ‘RBR’ is for comparative adverbs and ‘JJR’ is for comparative adjectives, and a combination of these two tags is strongly related to double-marked errors such as ‘more easier’. In the course of acquiring the comparative form, learners tend to use the double-marked form. The compound tags correctly capture this erroneous stage.

The ‘DEC’ group also includes three Wh-related tags – those in (WDT_NN,WDT_VBP, WRB), but the proportion is much smaller than that found in the ‘INC’ group.

The analysis shows that the combination of original and compound POS tags correctly captures systematic changes in the grammatical expressions according to changes in proficiency levels.

8.2. POS LM-based similarity assessment

In contrast to VSM-based features, ASR errors caused significant performance drops in the LM-based features. In this section, we will analyze the impact of ASR errors on LM-based features in detail with reference to the models based on Set 1 and Set 2 (the manual transcription based model). Since the feature *lmmax* had a better performance than lm_4, we conducted further analysis on *lmmax*. In addition, there was no significant difference between the manual-transcription-based model and the ASR-hypothesis-based model (Set 1 and Set 3). Hence, we limit our discussion to those features obtained using the manual transcription-based model.

The relationship between predicted scores and human scores in the case of LM-based similarity features was further analyzed using a contingency table for the feature *lmmax* with Set 1 and Set 2 (recall that in both these cases, the training data consists of manually transcribed utterances and it is only the mode of evaluation data that changes). Focusing on the best performing model (the bigram LM), the agreement between the human score and the feature was promising; for the manual transcription, the exact agreement between human scores and *lmmax* was 36% and the combination of the exact and adjacent agreement was 86%. Upon examination of the contingency table for the manual transcriptions, we notice that every predicted score level is dominated by responses from the score levels 2 and 3.

However, the feature *lmmax* was strongly influenced by recognition errors. This is seen in the significant drop in the correlation from 0.39 with set 1 to 0.29 with set 2. For ASR hypotheses the exact agreement decreased from 36% to 32%, and the adjacent agreement from 50% to 40%, respectively.

For further analysis, we calculated the mean values of *lmmax* (most similar score class) for each score level and compared ASR-based ones with transcription-based ones. Fig. 1 shows the results. As a comparison, Fig. 2 provides the mean values of cos_4 (cosine similarity value with score level 4) for each score level.

In *lmmax*, the mean values of ASR-based features were lower than those of manual transcription-based features except for the score-class 1. In particular, the drop in feature values for *lmmax* increased with increasing score levels and this resulted in reducing the distinction between score levels. However, no such behavior was found for cos_4. In fact, the mean values of ASR-based features were slightly higher than those of transcription-based features for all score levels, and the distinction between score levels remained steady.

We attribute the drop in performance in the ASR-based features to the following sources of errors:

1. The speech recognizer may mis-recognize grammatical errors in the response, replacing them with the more frequent, correct expressions.
2. The speech recognizer may mis-recognize sophisticated expressions in a response replacing them with more frequent, simpler expressions.
3. The recognizer errors may generate grammatical errors not existing in the actual responses.

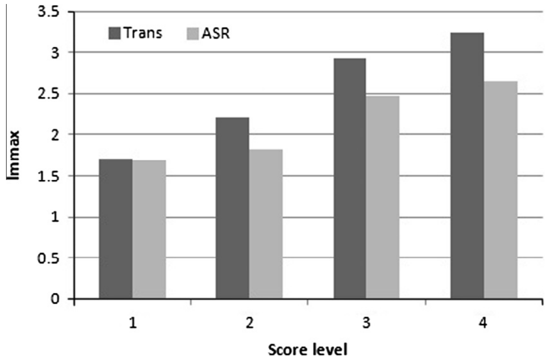

Fig. 1. Comparison of *lmmax* between manual-based features and ASR-based features at each score level.

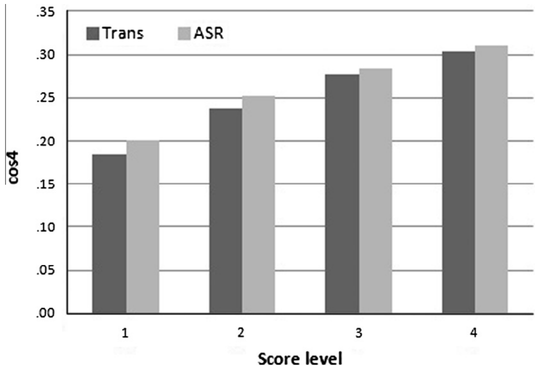

Fig. 2. Comparison of cos_4 between manual-based features and ASR-based features at each score level.

Each of the items above, as well as any combination thereof, results in a loss of characteristics of a response thereby reducing the discriminating power of the features. While 1 results in inflated predicted levels, 2 and 3 result in underestimated proficiency levels. Current results (Fig. 1) supports Hypothesis 2 and 3 since the mean values of ASR-based *lmmax* were generally lower than those of manual transcription-based ones.

In order to identify those bigrams that were heavily influenced by ASR errors, bigram frequencies for manual transcriptions and ASR hypotheses were obtained, and differences were examined. Since the impact of the errors was particularly strong on responses with a score of 4, we focused on responses with that score. We identified the top 10 bigrams for which the ASR-based frequencies were substantially higher than their manual transcription-based frequencies. Among the 10 bigrams, tag repetitions ('DT–DT','IN–IN','VBZ–VBZ') and some determiner-related bigrams ('DT–IN','DT–PRP','NN–DT') were strongly associated with grammatical errors. The analysis lends support to our hypothesis that ASR generates non-existing grammatical errors resulting in underestimated high scores. This results in an underestimated proficiency level for the highly proficient speakers when using the feature *lmmax*.

Surprisingly, the ASR errors that affected the performance of *lmmax* did not influence the performance of cos_4. We found that the 10 bigrams that were highly associated with ASR errors had relatively low *idf* values; they were 85–90% from the top when the bigrams were sorted by their *idf* values. This low *idf* has the effect of reducing the impact of ASR errors on cos_4. This suggests that when compared to *lmmax*, cos_4 seems to be better suited for being used in a fully automated speech scoring system that is based on speech recognition.

8.3. Extensions and future directions

In assessing the level of association between the set of features and the human scores for proficiency, the underlying assumption was that the overall proficiency score was also indicative of syntactic competence. In reality, the holistic manual scores (used in this study and otherwise) for overall proficiency capture more aspects of speech than just syntactic complexity. Thus, the use of a more analytic human score of syntactic complexity in place of the overall proficiency score may provide a more accurate picture of the quality of our features.

Future explorations in this direction will include combination of *n*-gram models and interpolated language models with interpolation weights that are specific to a proficiency level, an approach that will alleviate the problem of data sparsity.

9. Conclusions

In this study, we proposed a set of features corresponding to two models - the VSM model and the POS-LM model, for measuring grammatical competence, with the requirement that they be amenable for use with automatically recognized spontaneous speech. The underlying assumption in the choice of features is that the level of acquired grammatical proficiency is signaled by the distribution of the POS tags. We observed that each of these models has an associated best performing feature; the feature cos_4 in the VSM and the feature *lmmax* in POS-LM. The features were found to generalize well with respect to differences in item-type in responses. When used alongside existing features in the state-of-the-art scoring model for oral proficiency assessment in spontaneous speech, the VSM-based features were seen to be important predictors of oral proficiency. Additionally, their inclusion in the scoring model showed improved agreement between machine and human scores compared to the state-of-the-art model.

We also observed that a set of prototypical features based on sentence- or clause-based units are ill-suited for automatic scoring on ASR output which is in-line with previous studies. In addition to the reasonable correlation with human-assigned scores, the key advantage of our proposed features is the relative immunity to ASR errors (compared to conventional measures of syntactic complexity) making them suitable for use in automatic spontaneous speech scoring systems.

Acknowledgements

The authors would like to thank Klaus Zechner, Keelan Evanini, Lei Chen, and Shasha Xie for their valuable comments, help with data preparation and assistance with the experiments. In addition, we gratefully acknowledge the proof-reading assistance offered by Prama Pramod.

References

Attali, Y., Burstein, J., 2006. Automated essay scoring with e–rater R V.2. J. Technol. Learn. Assess. 4.

Balogh, J., Bernstein, J., Cheng, J., Townshend, B., 2007. Automated evaluation of reading accuracy: assessing machine scores. In: Proceedings of SLaTE, pp. 1–3.

Bardovi-Harlig, K., Bofman, T., 1989. Attainment of syntactic and morphological accuracy by advanced language learners. Stud. Second Lang. Acquis. 11, 17–34.

Bernstein, J., Cheng, J., Suzuki, M., 2010. Fluency and structural complexity as predictors of L2 oral proficiency. In: Proceedings of InterSpeech, pp. 1241–1244.

Bernstein, J., DeJong, J., Pisoni, D., Townshend, B., 2000. Two experiments in automated scoring of spoken language proficiency. In: Proceedings of InSTIL, pp. 57–61.

Chen, L., Yoon, S.Y., 2011. Detecting structural events for assessing non-native speech. In: Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications, pp. 38–45.

Chen, M., Zechner, K., 2011. Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. In: Proceedings of ACL, pp. 722–731.

Chodorow, M., Leacock, C., 2000. An unsupervised method for detecting grammatical errors. In: Proceedings of NAACL, pp. 140–147.

Cucchiarini, C., Strik, H., Boves, L., 2000. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. J. Acoust. Soc. Am. 107, 989–999.

Cucchiarini, C., Strik, H., Boves, L., 2002. Quantitative assessment of second language learners' fluency: comparisons between read and spontaneous speech. J. Acoust. Soc. Am. 111, 2862–2873.

Eskenazi, M., Kennedy, A., Ketchum, C., Olszewski, R., Pelton, G., 2007. The native accentTM pronunciation tutor: measuring success in the real world. In: Proceedings of SLaTE, pp. 124–127.

Feldman, S., Marin, M., Ostendorf, M., Gupta, M.R., 2009. Part-of-speech histograms for genre classification of text. In: Proceedings of ICASSP, pp. 4781–4784.

Foster, P., Skehan, P., 1996. The influence of planning and task type on second language performance. Stud. Second Lang. Acquis. 18, 299–324.

Foster, P., Tonkyn, A., Wigglesworth, G., 2000. Measuring spoken language: a unit for all reasons. Appl. Linguist. 21, 354–375.

Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., Butzberger, J., Rossier, R., Cesari, F., 2000. The SRI EduSpeak system: recognition and pronunciation scoring for language learning. In: Proceedings of InSTIL, pp. 123–128.

Franco, H., Neumeyer, L., Kim, Y., Ronen, O., 1997. Automatic pronunciation scoring for language instruction. In: Proceedings of ICASSP, pp. 1471–1474.

Grömping, U., 2006. Relative importance for linear regression in r: the package relaimpo. J. Stat. Softw. 17, 1–27.

Halleck, G.B., 1995. Assessing oral proficiency: a comparison of holistic and objective measures. Mod. Lang. J. 79, 223–234.

Hunt, K., 1970. Syntactic maturity inschool children and adults. Monogr. Soc. Res. Child Develop.

Iwashita, N., 2010. Features of oral proficiency in task performance by EFL and JFL learners. In: Selected Proceedings of the Second Language Research Forum, pp. 32–47.

Iwashita, N., Brown, A., McNamara, T., OHagan, S., 2008. Assessed levels of second language speaking proficiency: how distinct? Appl. Linguist. 29, 24–49.

Klein, D., Manning, C.D., 2003. Accurate unlexicalized parsing. In: Proceedings of ACL, pp. 423–430.

Lu, X., 2010. Automatic analysis of syntactic complexity in second language writing. Int. J. Corpus Linguist. 15, 474–496.

Lu, X., 2012. L2 Syntactic Complexity Analyzer. <http://www.personal.psu.edu/xxl13/downloads/l2sca.html/> (retrieved 17.03.2012).

Manning, C.D., Raghavan, P., Schütze, H., 2008. Introduction to Information Retrieval. Cambridge University Press, New York, USA.

Marcus, M.P., Santorini, B., Marcinkiewicz, M.A., 1993. Building a large annotated corpus of English: the penn treebank. Comput. Linguist. 19, 313–330.

Marin, M., Feldman, S., Ostendorf, M., Gupta, M. R., 2009. Filtering web text to match target genres. In: Proceedings of ICASSP, pp. 3705–3708.

Neumeyer, L., Franco, H., Digalakis, V., Weintraub, M., 2000. Automatic scoring of pronunciation quality. Speech Commun., 88–93.

Ortega, L., 2003. Syntactic complexity measures and their relationship to L2 proficiency: a research synthesis of college-level L2 writing. Appl. Linguist. 24, 492–518.

Roark, B., Mitchell, M., Hosom, J.P., Hollingshead, K., Kaye, J., 2011. Spoken language derived measures for detecting mild cognitive impairment. IEEE Trans. Audio Speech Lang. Process. 19, 2081–2090.

Robinson, P., 2006. Task complexity and second language narrative discourse. Lang. Learn. 45, 99–140.

Stolcke, A., et al., 2002. SRILM-an extensible language modeling toolkit. In: INTERSPEECH.

Temple, L., 2000. Second language learner speech production. Stud. Linguist., 288–297.

Tetreault, J.R., Chodorow, M., 2008. The ups and downs of preposition error detection in ESL writing. in: Proceedings of COLING, pp. 865–872.

Witt, S., 1999. Use of the Speech Recognition in Computer-Assisted Language Learning. Unpublished Dissertation, Cambridge University Engineering Department, Cambridge, UK.

Witt, S., Young, S., 1997. Performance measures for phone-level pronunciation teaching in CALL, in: Proceedings of STiLL, pp. 99–102.

Wolf-Quintero, K., Inagaki, S., Kim, H.Y., 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.

Yoon, S.Y., Bhat, S., 2012. Assessment of esl learners' syntactic competence based on similarity measures. In: Proceedings of EMNLP, pp. 600–608.

Yoon, S.Y., Higgins, D., 2011. Non-english response detection method for automated proficiency scoring system. In: Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications, pp. 161–169.

Zechner, K., Higgins, D., Xi, X., Williamson, D.M., 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. Speech Commun. 51, 883–895.

Zechner, K., Xi, X., Chen, L., 2011. Evaluating prosodic features for automated scoring of non-native read speech. 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, pp. 461–466.

Zissman, M.A., 1996. Comparison of four approaches to automatic language identification of telephone speech. Speech Audio Process. 4.