Assessment of ESL Learners' Syntactic Competence Based on Similarity Measures

Su-Youn Yoon
Educational Testing Service
Princeton, NJ 08541
syoon@ets.org

Suma Bhat
Beckman Institute,
Urbana, IL 61801
spbhat2@illinois.edu

Abstract

This study presents a novel method that measures English language learners' syntactic competence towards improving automated speech scoring systems. In contrast to most previous studies which focus on the length of production units such as the mean length of clauses, we focused on capturing the differences in the distribution of morpho-syntactic features or grammatical expressions across proficiency. We estimated the syntactic competence through the use of corpus-based NLP techniques. Assuming that the range and sophistication of grammatical expressions can be captured by the distribution of Part-of-Speech (POS) tags, vector space models of POS tags were constructed. We use a large corpus of English learners' responses that are classified into four proficiency levels by human raters. Our proposed feature measures the similarity of a given response with the most proficient group and is then estimates the learner's syntactic competence level.

Widely *outperforming* the state-of-the-art measures of syntactic complexity, our method attained a significant correlation with human-rated scores. The correlation between human-rated scores and features based on manual transcription was 0.43 and the same based on ASR-hypothesis was slightly lower, 0.42. An important advantage of our method is its robustness against speech recognition errors not to mention the simplicity of feature generation that captures a reasonable set of learner-specific syntactic errors.

1 Introduction

This study provides a novel method that measures ESL (English as a second language) learners' competence in grammar usage (syntactic competence). Being interdisciplinary in nature, it shows how to combine the core findings in the ESL literature with various empirical NLP techniques for the purpose of automated scoring.

Grammar usage is one of the dimensions of language ability that is assessed during non-native proficiency level testing in a foreign language. Overall proficiency in the target language can be assessed by testing the abilities in various areas including fluency, pronunciation, and intonation; grammar and vocabulary; and discourse structure. Testing rubrics for human raters contain descriptors used for the subjective assessment of several of these features. With the recent move towards the objective assessment of language ability (spoken and written), it is imperative that we develop methods for quantifying these abilities and measuring them automatically.

Ortega (2003) indicated that "the *range* of forms that surface in language production and the degree of *sophistication* of such forms" were two important areas in grammar usage and called the combination of these two areas "syntactic complexity." Features that measure syntactic complexity have been frequently studied in ESL literature and have been found to be highly correlated with students' proficiency levels in writing.

Studies in automated speech scoring have focused on fluency (Cucchiarini et al., 2000; Cucchiarini et al., 2002), pronunciation (Witt and Young, 1997;

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 600–608, Jeju Island, Korea, 12–14 July 2012.

Witt, 1999; Franco et al., 1997; Neumeyer et al., 2000), and intonation (Zechner et al., 2009), and relatively fewer studies have been conducted on grammar usage. More recently, Lu (2010), Chen and Yoon (2011) and Chen and Zechner (2011) have measured syntactic competence in speech scoring. Chen and Yoon (2011) estimated the complexity of sentences based on the average length of the clauses or sentences. In addition to these length measures, Lu (2010) and Chen and Zechner (2011) measured the parse-tree based features such as the mean depth of parsing tree levels. However, these studies found that these measures did not show satisfactory empirical performance in automatic speech scoring (Chen and Yoon, 2011; Chen and Zechner, 2011) when the features were calculated from the output of a speech recognition engine.

This study considers new features that measure syntactic complexity and is novel in two important ways. First, in contrast to most features that infer syntactic complexity based upon the length of the unit, we directly measure students' sophistication and range in grammar usage. Second, instead of rating a student's response using a scale based on native speech production, our experiments compare it with a similar body of learners' speech. Eliciting native speakers' data and rating it for grammar usage (supervised approach) can be arbitrary, since there can be a very wide range of possible grammatical structures that native speakers utilize. Instead, we proceed in a semi-supervised fashion. A large amount of learners' spoken responses were collected and classified into four groups according to their proficiency level. We then sought to find how distinct the proficiency classes were based on the distribution of POS tags. Given a student's response, we calculated the similarity with a sample of responses for each score level based on the proportion and distribution of Part-of-Speech using NLP techniques.

POS tag distribution has been used in various tasks such as text genre classification (Feldman et al., 2009); in a language testing context, it has been used in grammatical error detection (Chodorow and Leacock, 2000; Tetreault and Chodorow, 2008) and essay scoring. Recently, Roark et al. (2011) explored POS tag distribution to capture the differences in syntactic complexity between healthy subjects and subjects with mild cognitive impairment, but no other research has used POS tag distribution in measuring syntactic complexity, to the best of authors' knowledge.

An assessment of ESL learners' syntactic competence should consider the structure of sentences as a whole - a task which may not be captured by the simplistic POS tag distribution. However, studies of Lu (2010) and Chen and Zechner (2011) showed that more complex syntactic features are unreliable in ASR-based scoring system. Furthermore, we show that POS unigrams or bigrams indeed capture a reasonable portion of learners' range and sophistication of grammar usage in our discussion in Section 7.

This paper will proceed as follows: we will review related work in Section 2 and present the method to calculate syntactic complexity in Section 3. Data and experiment setup will be explained in Section 4 and Section 5. The results will be presented in Section 6. Finally, in Section 7, we discuss the levels of syntactic competence that are captured using our proposed measure.

2 Related Work

Second Language Acquisition (SLA) researchers have developed many quantitative measures to estimate the level of acquisition of syntactic competence. Bardovi-Harlig and Bofman (1989) classified these measures into two groups. The first group is related to the acquisition of specific morphosyntactic features or grammatical expressions. Tests of negations or relative clauses - whether these expressions occurred in the test responses without errors - fell into this group (hereafter, the expression-based group). The second group is related to the length of the clause or the relationship between clauses and hence not tied to particular structures (hereafter, the length-based group). Examples of the second group measures include the average length of clause unit and dependent clauses per sentence unit.

These syntactic measures have been extensively studied in ESL writing. Ortega (2003) synthesized 25 research studies which employed syntactic measures on ESL writing and reported a significant relationship between the proposed features and writing proficiency. He reported that a subset of features such as the mean length of the clause unit increased with students' proficiency. More recently, Lu (2010)

has conducted a more systematic study using an automated system. He applied 14 syntactic measures to a large database of Chinese learners' writing samples and found that syntactic measures were strong predictors of students' writing proficiency.

Studies in the area of automated speech scoring have only recently begun to actively investigate the usefulness of syntactic measures for scoring spontaneous speech (Chen et al., 2010; Bernstein et al., 2010). These have identified clause boundaries (identified from manual annotations and automatically) and obtained length-based features. In addition to these conventional syntactic complexity features, Lu (2009) implemented an automated system that calculates the revised Developmental Level (D-Level) Scale (Covington et al., 2006) using natural language processing (NLP) techniques. The original D-Level Scale was proposed by Rosenberg and Abbeduto (1987) based primarily on observations of child language acquisition. They classified children's grammatical acquisition into 7 different groups according to the presence of certain types of complex sentences. The revised D-Level Scale classified sentences into the eight levels according to the presence of particular grammatical expressions. For instance, level 0 is comprised of simple sentences, while level 5 is comprised of sentences joined by subordinating conjunction or nonfinite clauses in an adjunct position. The D-Level Scale has been less studied in the speech scoring. To our knowledge, Chen and Zechner (2011) is the only study that applied the D-Level analyzer to ESL learners' spoken responses.

In contrast to ESL writing, applying syntactic complexity features, both conventional length-based features and D-Level features, presents serious obstacles for speaking. First, the length of the spoken responses are typically shorter than written responses. Most measures are based on sentence or sentence-like units, and in speaking tests that elicit only a few sentences the measures are less reliable. Chen and Yoon (2011) observed a marked decrease in correlation between syntactic measures and proficiency as response length decreased. In addition, speech recognition errors only worsen the situation. Chen and Zechner (2011) showed that the significant correlation between syntactic measures and speech proficiency (correlation coefficient = 0.49) became insignificant when they were applied to the speech recognition word hypotheses. Errors in speech recognition seriously influenced the measures and decreased the performance. Due to these problems, the existing syntactic measures do not seem reliable enough for being used in automated speech proficiency scoring.

In this study, we propose novel syntactic measures which are relatively robust against speech recognition errors and are reliable in short responses. In contrast to recent studies focusing on length-based features, we focus on capturing differences in the distribution of morphosyntactic features or grammatical expressions across proficiency levels. We investigate the distribution of a broader class of grammatical forms through the use of corpus-based NLP techniques.

3 Method

Many previous studies, that assess syntactic complexity based on the distribution of morphosyntactic features and grammatical expressions, limited their experiments to a few grammatical expressions. Covington et al. (2006) and Lu (2009) covered all sentence types, but their approaches were based on expert observation (supervised rubrics), and descriptions of each level were brief and abstract. It is important to develop a more detailed and refined scale, but developing scales in a supervised way is difficult due to the subjectivity and the complexity of structures involved.

In order to overcome this problem, we employed NLP technology and a corpus-based approach. We hypothesize that the level of acquired grammatical forms is signaled by the distribution of the POS tags, and the differences in grammatical proficiency result in differences in POS tag distribution. Based on this assumption, we collected large amount of ESL learners' spoken responses and classified them into four groups according to their proficiency levels. The syntactic competence was estimated based on the similarity between the test responses and learners' corpus.

A POS-based vector space model (VSM), in which the response belonging to separate proficiency levels were converted to vectors and the similarity between vectors were calculated using cosine

similarity measure and tf-idf weighting, was employed. Such a score-category-based VSM has been used in automated essay scoring. Attali and Burstein (2006) to assess the lexical content of an essay by comparing the words in the test essay with the words in a sample essays from each score category. We extend this to assessment of grammar usage using vectors of POS tags.

Proficient speakers use complicated grammatical expressions, 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. These distributional differences are captured by inverse-document frequency.

In addition, we identify frequent POS tag sequences as those having high mutual information and include them in our experiments. Temple (2000) pointed out that the proficient learners are characterized by increased automaticity in speech production. These speakers tend to memorize frequently used multi-word sequences as a chunk and retrieve the whole chunk as a single unit. The degree of automaticity can be captured by the frequent occurrence of POS sequences with high mutual information.

We quantify the usefulness of the generated features for the purpose of automatic scoring by first considering its correlation with the human scores. We then compare the performance of our features with those in Lu (2011), where the features are a collection of measures of syntactic complexity that have shown promising directions in previous studies.

4 Data

Two different sets of data were used in this study: the AEST 48K dataset and AEST balanced dataset. Both were collections of responses from the AEST, a high-stakes test of English proficiency and had no overlaps. The AEST assessment consists of 6 items in which speakers are prompted to provide responses lasting between 45 and 60 seconds per item. In summary, approximately 3 minutes of speech is collected per speaker.

Among the 6 items, two items are tasks that ask examinees to provide information or opinions on familiar topics based on their personal experience or background knowledge. The four remaining items are integrated tasks that include other language skills such as listening and reading. All items extract spontaneous, unconstrained natural speech. The size, purpose, and speakers' native language information for each dataset is summarized in Table 1.

Each response was rated by trained human raters using a 4-point scoring scale, where 1 indicates a low speaking proficiency and 4 indicates a high speaking proficiency. In order to evaluate the reliability of the human ratings, the data should be scored by two raters. Since none of the AEST balanced data was double scored the inter-rater agreement ratio was estimated using a large (41K) double-scored dataset using the same scoring guidelines and scoring process. The Pearson correlation coefficient was 0.63 suggesting a reasonable inter-rater agreement. The distribution of the scores for this data can be found in Table 2.

We used the AEST 48K dataset as the training data and the AEST balanced dataset as the evaluation data.

5 Experiments

5.1 Overview

Our experimental procedure is as follows. All transcriptions were tagged using the POS tagger described in Section 5.3 and POS tag sequences were extracted. Next, the POS-based VSMs (one for each score class) were created using the AEST 48K dataset. Finally, for a given test response in the AEST balanced dataset, similarity features were generated.

A score-class-specific POS-based VSM was created using POS tags generated from the manual transcriptions. For evaluation, two different types of transcriptions (manual transcription and word hypotheses from the speech recognizer described in Section 5.2) were used in order to investigate the influence of speech recognition errors in the feature performance.

5.2 Speech recognition

An HMM recognizer was trained on AEST 48K dataset - approximately 733 hours of non-native speech collected from 7872 speakers. A gender independent triphone acoustic model and combination

Corpus name	Purpose	Number of speakers	Number of responses	Native languages	Size (Hrs)
AEST 48K data	ASR training and POS model training	7872	47227	China (20%), Korea (19%), Japanese (7%), India (7%), others (46%)	733
AEST balanced data	Feature development and evaluation	480	2880	Korean (15%), Chinese (14%), Japanese (7%), Spanish (9%), Others (55%)	44

Table 1: Data size and speakers native languages

Corpus name	Size	Score1	Score2	Score3	Score4
AEST 48K data	Number of files	1953	16834	23106	5334
	(%)	4	36	49	11
AEST balanced data	Number of files	141	1133	1266	340
	(%)	5	40	45	12

Table 2: Proficiency scores and data sizes

of bigram, trigram, and four-gram language models were used. The word error rate (WER) on the held-out test dataset was 27%.

5.3 POS tagger

POS tags were generated using the POS tagger implemented in the OpenNLP toolkit. It was trained on the Switchboard (SWBD) corpus. This POS tagger was trained on about 528K word/tag pairs and achieved a tagging accuracy of 96.3% on a test set of 379K words. The Penn POS tag set was used in the tagger.

5.4 Unit generation using mutual information

POS bigrams with high mutual information were selected and used as a single unit. First, all POS bigrams which occurred less than 50 times were filtered out. Next, the remaining POS tag bigrams were sorted by their mutual information scores, and two different sets (top50 and top110) were selected. The selected POS pairs were transformed into compound tags. As a result, we generated three sets of POS units by this process: the original POS set without the compound unit (Base), the original set and an additional 50 compound units (Base+mi50), and the original set and an additional 110 units (Base+mi110).

Finally, unigram, bigram and trigram were generated for each set separately. The size of total terms in each condition was presented in table 3.

	Base	Base+mi50	Base+mi110
Unigram	42	93	151
Bigram	1366	4284	9691
Trigram	21918	54856	135430

Table 3: Number of terms used in VSMs

5.5 Building VSMs

For each ngram, three sets of VSMs were built using three sets of tags as terms, yielding a total of nine VSMs. The results were based on the individual model and we did not combine any models.

5.6 Cosine similarity-based features

The cosine similarity has been frequently used in the information retrieval field to identify the relevant documents for the given query. This measures the similarity between a given query and a document by measuring the cosine of the angle between vectors in a high-dimensional space, whereby each term in the query and documents corresponding to a unique dimension. If a document is relevant to the query, then it shares many terms resulting in a small angle. In this study, the term was a single or compound POS tag (unigram,bigram or trigram) weighted by its tf-idf, and the document was the response.

First, the inverse document frequency was calculated from the training data, and each response was treated as a document. Next, responses in the same

	Unigram			Bigram			Trigram		
	Base	Base +mi50	Base +mi110	Base	Base +mi50	Base +mi110	Base	Base +mi50	Base +mi110
Trans-cription	0.301**	0.297**	0.329**	0.427**	0.361**	0.366**	0.402**	0.322**	0.295**
ASR	0.246**	0.272**	0.304**	0.415**	0.348**	0.347**	0.373**	0.311**	0.282**

Table 4: Pearson correlation coefficients between ngram-based features and expert proficiency scores
** Correlation is significant at the 0.01 level

score group were concatenated, and a single vector was generated for each score group. A total of 4 vectors were generated using training data. For each test response, a similarity score was calculated as follows:

$$\cos(\vec{q}, \vec{d_j}) = \frac{\sum_{i=1}^{n} q_i d_{ji}}{\sum_{i=1}^{n} {q_i}^2 \sum_{i=1}^{n} {d_i}^2}$$

$$q_i \equiv tf(t_i, \vec{q}) \times \log\left(\frac{N}{df(t_i)}\right)$$

$$d_{ji} \equiv tf(t_i, \vec{d_j}) \times \log\left(\frac{N}{df(t_i)}\right)$$

where $\vec{q}$ is a vector of the test response,
$\vec{d_j}$ is a vector of the $scoreGroup_j$,
n is the total number of POS tags,
$tf(t_i, \vec{q})$ is the term frequency of POS tag t_i in the test response,
$tf(t_i, \vec{d_j})$ is the term frequency of POS tag t_i in the $scoreGroup_j$,
N is the total number of training responses,
$df(t_i)$ is the document frequency of POS tag t_i in the total training responses

Finally, a total of 4 cos scores (one per score group) were generated. Among these four values, the $cos4$, the similarity score to the responses in the score group 4, was selected as a feature with the following intuition. $cos4$ measures the similarity of a given test response to the representative vector of score class 4; the larger the value, the closer it would be to score class 4.

6 Results

6.1 Correlation

Table 4 shows correlations between cosine similarity features and proficiency scores rated by experts.

The bigram-based features outperformed both unigram-based and trigram-based features. In particular, the similarities using the *base* tag set with bigrams achieved the best performance. By adding the mutual information-based compound units to the original POS tag sets, the performance of features improved in the unigram models. However, there was no performance gain in either bigram or trigram models; on the contrary, there was a large drop in performance. Unigrams have good coverage but limited power in distinguishing different score levels. On the other hand, trigrams have opposite characteristics. Bigrams seem to strike a balance in both coverage and complexity (from among the three considered here) and may thus have resulted in the best performace.

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

6.2 Comparison with other Measures of Syntactic Complexity

We compared the performance of our features with the features of syntactic complexity proposed in (Lu, 2011). Towards this, the clause boundaries of the ASR hypotheses, were automatically detected using the automated clause boundary detection method[1].

[1]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

The utterances were then parsed using the Stanford Parser, and a total of 22 features including both length-related features and parse-tree based features were generated using (Lu, 2011). Finally, we calculated Pearson correlation coefficients between these features and human proficiency scores.

Study	Feature	Correlation
Current study	bigram based cos4	0.41**
(Lu, 2011)	DCC	0.14**

Table 5: Comparison between (Lu, 2011) and this study
** Correlation is significant at the 0.01 level

As indicated in Table 5, the best performing feature was mean number of dependent clauses per clause (DCC) and the correlation r was 0.14. No features other than DCC achieved statistically significant correlation. Our best performing feature (bigram based cos4) *widely outperformed* the best of Lu (2011)'s features (correlations approximately 0.3 apart).

A logical explanation for the poor performance of Lu (2011)'s features is that the features are generated using multi-stage automated process, and the errors in each process contributes the low feature performance. For instance, the errors in the automated clause boundary detection may result in a serious drop in the performance. With the spoken responses being particularly short (a typical response in the data set had 10 clauses on average), even one error in clause boundary detection can seriously affect the reliability of features.

7 Discussion

While the measure of syntactic competence that we study here is an abstraction of the overall syntactic competence, without consideration of specific constructions, we analyzed the results further with the intention of casting light on the level of details of syntactic competence that can be explained using our measure. Furthermore, this section 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 comprised of 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 proficiency levels. Owing to space constraints, we restrict our discussion to the analysis using unigrams (base and compund). 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 110 compound tags, were classified into 5 classes:

- Absence-of-low-proficiency (ABS): Group of POS tags that appear in all score groups except the lowest proficiency group;
- Increase (INC): Group of POS tags whose ranks increase consistently as proficiency increases;
- Decrease (DEC): Group of POS tags whose ranks decrease consistently as proficiency increases;
- Constant (CON): Group of POS tags whose ranks remain same despite change in proficiency;
- Mix: Group of POS tags of with no consistent pattern in the ranks.

Table 6 presents the number of POS tags in each class.

ABS	INC	DEC	CON	Mix
14	37	33	18	48

Table 6: Tag distribution and proficiency scores

The 'ABS' class mostly consists of 'WP' and 'WDT'; more than 50% of tags in this class are related to these two tags. 'WP' is a Wh-pronoun while 'WDT' is a Wh-determiner. Since most sentences in

F-score of 0.60 on the non-native speakers' ASR hypotheses. A detailed description of the method is presented in (Chen and Zechner, 2011)

our data are declarative sentences, 'Wh' phrase signals the use of relative clause. Therefore, the lack of these tags strongly support the hypothesis that the speakers in score group 1 showed incompetence in the use of relative clauses or their use in limited situations.

The 'INC' class can be sub-classified into three groups: verb, comparative, and relative clause. Verb group is includes the 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' class can be sub-classified into five 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 that less proficient speakers 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). The expressions in these groups are simpler than those in 'Increase' 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 it signals 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 that a modal verb is not followed by 'VB' (base form) and as seen 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 'RBR_JJR' is related to the correct acquisition of the comparative form. 'RBR' is for comparative adverbs and 'JJR' is for comparative adjectives, and the combination of two tags is strongly related to double-marked errors such as 'more easier'. In the intermediate stage in the acquisition of comparative form, learners tend to use the double-marked form. The compound tags correctly capture this erroneous stage.

The 'Decrease' class also includes three Wh-related tags (WDT_NN, WDT_VBP, WRB), but the proportion is much smaller than the 'Increase' class.

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

The robust performance of our proposed measure to speech recognition errors may be better appreciated in the context of similar studies. Compared with the state-of-the art measures of syntactic complexity proposed in Lu (2011) our features achieve significantly better performance especially when generated from ASR hypotheses. It is to be noted that the performance drop between the transcription-based feature and the ASR hypothesis-based feature was marginal.

8 Conclusions

In this paper, we presented features that measure syntactic competence for the automated speech scoring. The features measured the range and sophistication of grammatical expressions based on POS tag distributions. A corpus with a large number of learners' responses was collected and classified into four groups according to proficiency levels. The syntactic competence of the test response was estimated by identifying the most similar group from the learners' corpus. Furthermore, speech recognition errors only resulted in a minor performance drop. The robustness against speech recognition errors is an important advantage of our method.

Acknowledgments

The authors would like to thank Shasha Xie, Klaus Zechner, and Keelan Evanini for their valuable comments, help with data preparation and experiments.

References

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

Kathleen Bardovi-Harlig and Theodora Bofman. 1989. Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition*, 11:17–34.

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

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

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

Lei Chen, Joel Tetreault, and Xiaoming Xi. 2010. Towards using structural events to assess non-native speech. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 74–79.

Martin Chodorow and Claudia Leacock. 2000. An unsupervised method for detecting grammatical errors. In *In Proceedings of NAACL00*, 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. Technical report, CASPR Research Report 2006-01, Athens, GA: The University of Georgia, Artificial Intelligence Center.

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

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

Sergey Feldman, M.A. Marin, Maria Ostendorf, and Maya R. Gupta. 2009. Part-of-speech histograms for genre classification of text. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4781 –4784, april.

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

Xiaofei Lu. 2009. Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics*, 14(1):3–28.

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

Xiaofei Lu. 2011. L2 syntactic complexity analyze. Retrieved March 17, 2012 from `http://www.personal.psu.edu/xxl13/downloads/l2sca.html/`.

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.

Brian. Roark, Margaret Mitchell, John-Paul. Hosom, Kristy Hollingshead, and Jeffrey Kaye. 2011. Spoken language derived measures for detecting mild cognitive impairment. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7):2081 –2090, sept.

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.

Liz Temple. 2000. Second language learner speech production. *Studia Linguistica*, pages 288–297.

Joel R. Tetreault and Martin Chodorow. 2008. The ups and downs of preposition error detection in esl writing. In *In Proceedings of COLING*.

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

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

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