Computing and Evaluating Syntactic Complexity Features for Automated Scoring of Spontaneous Non-Native Speech

Miao Chen
School of Information Studies
Syracuse University
Syracuse, NY, USA
mchen14@syr.edu

Klaus Zechner
NLP & Speech Group
Educational Testing Service
Princeton, NJ, USA
kzechner@ets.org

Abstract

This paper focuses on identifying, extracting and evaluating features related to syntactic complexity of spontaneous spoken responses as part of an effort to expand the current feature set of an automated speech scoring system in order to cover additional aspects considered important in the construct of communicative competence.

Our goal is to find effective features, selected from a large set of features proposed previously and some new features designed in analogous ways from a syntactic complexity perspective that correlate well with human ratings of the same spoken responses, and to build automatic scoring models based on the most promising features by using machine learning methods.

On human transcriptions with manually annotated clause and sentence boundaries, our best scoring model achieves an overall Pearson correlation with human rater scores of r=0.49 on an unseen test set, whereas correlations of models using sentence or clause boundaries from automated classifiers are around r=0.2.

1 Introduction

Past efforts directed at automated scoring of speech have used mainly features related to fluen cy (e.g., speaking rate, length and distribution of pauses), pronunciation (e.g., using log-likelihood scores from the acoustic model of an Automatic Speech Recognition (ASR) system), or prosody (e.g., information related to pitch contours or syllable stress) (e.g., Bernstein, 1999; Bernstein et al., 2000; Bernstein et al., 2010; Cucchiarini et al., 1997; Cucchiarini et al., 2000; Franco et al., 2000a; Franco et al., 2000b; Zechner et al., 2007, Zechner et al., 2009).

While this approach is a good match to most of the important properties related to low entropy speech (i.e., speech which is highly predictable), such as reading a passage aloud, it lacks many important aspects of spontaneous speech which are relevant to be evaluated both by a human rater and an automated scoring system. Examples of such aspects of speech, which are considered part of the construct[1] of "communicative competence (Bachman, 1990), include grammatical accuracy, syntactic complexity, vocabulary diversity, and aspects of spoken discourse structure, e.g., coherence and cohesion. These different aspects of speaking proficiency are often highly correlated in a non-native speaker (Xi and Mollaun, 2006; Bernstein et al., 2010), and so scoring models built solely on features of fluency and pronunciation may achieve reasonably high correlations with holistic human rater scores. However, it is important to point out that such systems would still be unable to assess many important aspects of the speaking construct and therefore cannot be seen as ideal from a validity point of view.[2]

The purpose of this paper is to address one of these important aspects of spoken language in more detail, namely syntactic complexity. This paper can be seen as a first step toward including

[1] A construct is a set of knowledge, skills, and abilities measured by a test.

[2] "Construct validity" refers to the extent that a test measures what it is designed to measure, in this case, communicative competence via speaking.

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 722–731,
Portland, Oregon, June 19-24, 2011.

features related to this part of the speaking construct into an already existing automated speech scoring system for spontaneous speech which so far mostly uses features related to fluency and pronunciation (Zechner et al., 2009).

We use data from the speaking section of the TOEFL® Practice Online (TPO) test, which is a low stakes practice test for non-native speakers where they are asked to provide six spontaneous speech samples of about one minute in length each in response to a variety of prompts. Some prompts may be simple questions, and others may involve reading or listening to passages first and then answering related questions. All responses were scored holistically by human raters according to pre-defined scoring rubrics (i.e., specific scoring guidelines) on a scale of 1 to 4, 4 being the highest proficiency level.

In our automated scoring system, the first component is an ASR system that decodes the digitized speech sample, generating a time-annotated hypothesis for every response. Next, fluency and pronunciation features are computed based on the ASR output hypotheses, and finally a multiple regression scoring model, trained on human rater scores, computes the score for a given spoken response (see Zechner et al. (2009) for more details). We conducted the study in three steps: (1) finding important measures of syntactic complexity from second language acquisition (SLA) and English language learning (ELL) literature, and extending this feature set based on our observations of the TPO data in analogous ways; (2) computing features based on transcribed speech responses and selecting features with highest correlations to human rater scores, also considering their comparative values for native speakers taking the same test; and (3) building scoring models for the selected sub-set of the features to generate a proficiency score for each speaker, using all six responses of that speaker.

In the remainder of the paper, we will address related work in syntactic complexity (Section 2), introduce the speech data sets of our study (Section 3), describe the methods we used for feature extraction (Section 4), provide the experiment design and results (Section 5), analyze and discuss the results in Section 6, before concluding the paper (Section 7).

2 Related Work

2.1 Literature on Syntactic Complexity

Syntactic complexity is defined as "the range of forms that surface in language production and the degree of sophistication of such forms" (Ortega, 2003). It is an important factor in the second language assessment construct as described in Bachman's (1990) conceptual model of language ability, and therefore is often used as an index of language proficiency and development status of L2 learners. Various studies have proposed and investigated measures of syntactic complexity as well as examined its predictiveness for language proficiency, in both L2 writing and speaking settings, which will be reviewed respectively.

Writing

Wolfe-Quintero et al. (1998) reviewed a number of grammatical complexity measures in L2 writing from thirty-nine studies, and their usage for predicting language proficiency was discussed. Some examples of syntactic complexity measures are: mean number of clauses per T-unit[3], mean length of clauses, mean number of verbs per sentence, etc. The various measures can be grouped into two categories: (1) clauses, sentences, and T-units in terms of each other; and (2) specific grammatical structures (e.g., passives, nominals) in relation to clauses, sentences, or T-units (Wolfe-Quintero et al., 1998). Three primary methods of calculating syntactic complexity measures are frequency, ratio, and index, where frequency is the count of occurrences of a specific grammatical structure, ratio is the number of one type of unit divided by the total number of another unit, and index is computing numeric scores by specific formulae (Wolfe-Quintero et al., 1998). For example, the measure "mean number of clauses per T-unit" is obtained by using the ratio calculation method and the clause and T-unit grammatical structures. Some structures such as clauses and T-units only need shallow linguistic processing to acquire, while some require parsing. There are numerous combinations for measures and we need empirical evi-

[3] T-units are defined as "shortest grammatically allowable sentences into which (writing can be split) or minimally terminable units" (Hunt, 1965:20).

dence to select measures with the highest performance.

There have been a series of empirical studies examining the relationship of syntactic complexity measures to L2 proficiency using real-world data (Cooper, 1976; Larsen-Freeman, 1978; Perkins, 1980; Ho-Peng, 1983; Henry, 1996; Ortega, 2003; Lu, 2010). The studies investigate measures that highly correlate with proficiency levels or distinguish between different proficiency levels. Many T-unit related measures were identified as statistically significant indicators to L2 proficiency, such as mean length of T-unit (Henry, 1996; Lu, 2010), mean number of clauses per T-unit (Cooper, 1976; Lu, 2010), mean number of complex nominals per T-unit (Lu, 2010), or the mean number of error-free T-units per sentence (Ho-Peng, 1983). Other significant measures are mean length of clause (Lu, 2010), or frequency of passives in composition (Kameen, 1979).

Speaking

Syntactic complexity analysis in speech mainly inherits measures from the writing domain, and the abovementioned measures can be employed in the same way on speech transcripts for complexity computation. A series of studies have examined relations between the syntactic complexity of speech and the speakers' holistic speaking proficiency levels (Halleck, 1995; Bernstein et al., 2010; Iwashita, 2006). Three objective measures of syntactic complexity, including mean T-unit length, mean error-free T-unit length, and percent of error-free T-units were found to correlate with holistic evaluations of speakers in Halleck (1995). Iwashita's (2006) study on Japanese L2 speakers found that length-based complexity features (i.e., number of T-units and number of clauses per T-unit) are good predictors for oral proficiency. In studies directly employing syntactic complexity measures in other contexts, ratio-based measures are frequently used. Examples are mean length of utterance (Condouris et al., 2003), word count or tree depth (Roll et al., 2007), or mean length of T-units and mean number of clauses per T-unit (Bernstein et al., 2010). Frequency-based measures were used less, such as number of full phrases in Roll et al. (2007).

The speaking output is usually less clean than writing data (e.g., considering disfluencies such as false starts, repetitions, filled pauses etc.). Therefore we may need to remove these disfluencies first before computing syntactic complexity features. Also, importantly, ASR output does not contain interpunctuation but both for sentential-based features as well as for parser-based features, the boundaries of clauses and sentences need to be known. For this purpose, we will use automated classifiers that are trained to predict clause and sentence boundaries, as described in Chen et al. (2010). With previous studies providing us a rich pool of complexity features, additionally we also develop features analogous to the ones from the literature, mostly by using different calculation methods. For instance, the frequency of Prepositional Phrases (PPs) is a feature from the literature, and we add some variants such as number of PPs per clause as a new feature to our extended feature set.

2.2 Devising the Initial Feature Set

Through this literature review, we identified some important features that were frequently used in previous studies in both L2 speaking and writing, such as length of sentences and number of clauses per sentence. In addition, we also collected candidate features that were less frequently mentioned in the literature, in order to start with a larger field of potential candidate features. We further extended the feature set by inspecting our data, described in the following section, and created suitable additional features by means of analogy. This process resulted in a set of 91 features, 11 of which are related to clausal and sentential unit measurements (frequency-based) and 80 to measurements within such units (ratio-based). From the perspective of extracting measures, in our study, some measures can be computed using only clause and sentence boundary information, and some can be derived only if the spoken responses are syntactically parsed. In our feature set, there are two types of features: clause and sentence boundary based (26 in total) and parsing based (65). The features will be described in detail in Section 4.

3 Data

Our data set contains (1) 1,060 non-native speech responses of 189 speakers from the TPO test (NN set), and (2) 100 responses from 48 native speakers that took the same test (Nat set). All responses were verbatim transcribed manually and scored

holistically by human raters. (We only made use of the scores for the non-native data set in this study, since we purposefully selected speakers with perfect or near perfect scores for the Nat set from a larger native speech data set.) As mentioned above, there are four proficiency levels for human scoring, levels 1 to 4, with higher levels indicating better speaking proficiency.

The NN set was randomly partitioned into a training (NN-train) and a test set with 760 and 300 responses, respectively, and no speaker overlap.

Data Set	**Responses**	**Speakers**	**Responses per Speaker (average)**
NN-train	760	137	5.55
	Description: used to train sentence and clause boundary detectors, evaluate features and train scoring models		
1: NN-test-1-Hum	300	52	5.77
	Description: human transcriptions and annotations of sentence and clause boundaries		
2: NN-test-2-CB	300	52	5.77
	Description: human transcriptions, automatically predicted clause boundaries		
3: NN-test-3-SB	300	52	5.77
	Description: human transcriptions, automatically predicted sentence boundaries		
4: NN-test-4-ASR-CB	300	52	5.77
	Description: ASR hypotheses, automatically predicted clause boundaries		
5: NN-test-5-ASR-SB	300	52	5.77
	Description: ASR hypotheses, automatically predicted sentence boundaries		

Table 1. Overview of non-native data sets.

A second version of the test set contains ASR hypotheses instead of human transcriptions. The word error rate (WER[4]) on this data set is 50.5%. We used a total of five variants of the test sets, as described in Table 1. Sets 1-3 are based on human transcriptions, whereas sets 4 and 5 are based on ASR output. Further, set 1 contains human annotated clause and sentence boundaries, whereas the other 4 sets have clause or sentence boundaries predicted by a classifier.

All human transcribed files from the NN data set were annotated for clause boundaries, clause types, and disfluencies by human annotators (see Chen et al. (2010)).

For the Nat data set, all of the 100 transcribed responses were annotated in the same manner by a human annotator. They are not used for any training purposes but serve as a comparative reference for syntactic complexity features derived from the non-native corpus.

The NN-train set was used both for training clause and sentence boundary classifiers, as well as for feature selection and training of the scoring models. The two boundary detectors were machine learning based Hidden Markov Models, trained by using a language model derived from the 760 training files which had sentence and clause boundary labels (NN-train; see also Chen et al. (2010)).

Since a speaker's response to a single test item can be quite short (fewer than 100 words in many cases), it may contain only very few syntactic complexity features we are looking for. (Note that much of the previous work focused on written language with much longer texts to be considered.) However, if we aggregate responses of a single speaker, we have a better chance of finding a larger number of syntactic complexity features in the aggregated file. Therefore we joined files from the same speaker to one file for the training set and the five test sets, resulting in 52 aggregated files in each test set. Accordingly, we averaged the response scores of a single speaker to obtain the total speaker score to be used later in scoring model training and evaluation (Section 5).[5]

While disfluencies were used for the training of the boundary detectors, they were removed afterwards from the annotated data sets to obtain a tran-

[4] Word error rate (WER) is the ratio of errors from a string between the ASR hypothesis and the reference transcript, where the sum of substitutions, insertions, and deletions is divided by the length of the reference. To obtain WER in percent, this ratio is multiplied by 100.0.

[5] Although in most operational settings, features are derived from single responses, this may not be true in all cases. Furthermore, scores of multiple responses are often combined for score reporting, which would make such an approach easier to implement and argue for operationally.

scription which is "cleaner" and lends itself better to most of the feature extraction methods we use.

4 Feature Extraction

4.1 Feature Set

As mentioned in Section 2, we gathered 91 candidate syntactic complexity features based on our literature review as initial feature set, which is grouped into two categories: (1) Clause and sentence Boundary based features (CB features); and (2) Parse Tree based features (PT features). Clause based features are based on both clause boundaries and clause types and can be generated from human clause annotations, e.g., "frequency of adjective clauses[6] per one thousand words", "mean number of dependent clauses per clause", etc. Parse tree based features refer to features that are generated from parse trees and cannot be extracted from human annotated clauses directly.

We first selected features showing high correlation to human assigned scores. In this process the CB features were computed from human labeled clause boundaries in transcripts for best accuracy, and PT features were calculated from using parsing and other tools because we did not have human parse tree annotations for our data.

We used the Stanford Parser (Klein and Manning, 2003) in conjunction with the Stanford Tregex package (Levy and Andrew, 2006) which supports using rules to extract specific configurations from parse trees, in a package put together by Lu (Lu, 2011). When given a sentence, the Stanford Parser outputs its grammatical structure by grouping words (and phrases) in a tree structure and identifies grammatical roles of words and phrases.

Tregex is a tree query tool that takes Stanford parser trees as input and queries the trees to find subtrees that meet specific rules written in Tregex syntax (Levy and Andrew, 2006). It uses relational operators regulated by Tregex, for example, "A << B" stands for "subtree A dominates subtree B". The operators primarily function in subtree precedence, dominance, negation, regular expression, tree node identity, headship, or variable groups, among others (Levy and Andrew, 2006).

Lu's tool (Lu, 2011), built upon the Stanford Parser and Tregex, does syntactic complexity analysis given textual data. Lu's tool contributed 8 of the initial CB features and 6 of the initial PT features, and we computed the remaining CB and PT features using Perl scripts, the Stanford Parser, and Tregex.

Table 2 lists the sub-set of 17 features (out of 91 features total) that were used for building the scoring models described later (Section 5).

4.2 Feature Selection

We determined the importance of the features by computing each feature's correlation with human raters' proficiency scores based on the training set NN-train. We also used criteria related to the speaking construct, comparisons with native speaker data, and feature inter-correlations. While approaches coming from a pure machine learning perspective would likely use the entire feature pool as input for a classifier, our goal here is to obtain an initial feature set by judicious and careful feature selection that can withstand the scrutiny of construct validity in assessment development.

As noted earlier, the disfluencies in the training set had been removed to obtain a "cleaner" text that looks somewhat more akin to a written passage and is easier to process by NLP modules such as parsers and part-of-speech (POS) taggers.[7] The extracted features partly were taken directly from proposals in the literature and partly were slightly modified to fit our clause annotation scheme. In order to have a unified framework for computing syntactic complexity features, we used a combination of the Stanford Parser and Tregex for computing both clause- and sentence-based features as well as parse-tree-based features, i.e., we did not make use of the human clause boundary label annotations here. The only exception to this

[6] An adjective clause is a clause that functions as an adjective in modifying a noun. E.g., "This cat is a cat that is difficult to deal with."

[7] We are aware that disfluencies can provide valuable clues about spoken proficiency in and of themselves; however, this study is focused exclusively on syntactic complexity analysis, and in this context, disfluencies would distort the picture considerably due to the introduction of parsing errors, e.g.

Name	Type[8]	Meaning	Correlation	Regression
MLS	CB	Mean length of sentences	0.329	0.101
MLT	CB	Mean length of T-units	0.300	-0.059
DC/C	CB	Mean number of dependent clauses per clause	0.291	2.873
SSfreq	CB	Frequency of simple sentences per 1000 words	-.0242	0.001
MLSS	CB	Mean length of simple sentences	0.255	0.040
ADJCfreq	CB	Frequency of adjective clauses per 1000 words	0.253	0.004
Ffreq	CB	Frequency of fragments per 1000 words	-0.386	-0.057
MLCC	CB	Mean length of coordinate clauses	0.224	0.017
CT/T	PT	Mean number of complex T-units per T-unit	0.248	0.908
PP_ling/S	PT	Mean number of linguistically meaningful prepositional phrases (PP) per sentence[9]	0.310	0.423
NP/S	PT	Mean number of noun phrases (NP) per sentence	0.244	-0.411
CN/S	PT	Mean number of complex nominal per sentence	0.325	0.653
VB _ling/T	PT	Mean number of linguistically meaningful[10] verb phrases per T-unit	0.273	-0.780
PAS/S	PT	Mean number of passives per sentence	0.260	1.520
DI/T	PT	Mean number of dependent infinitives per T-unit	0.325	1.550
MLev	PT	Mean number of parsing tree levels per sentence	0.306	-0.134
MPSam	PT	Mean P-based Sampson[11] per sentence	0.254	0.234

Table 2. List of syntactic complexity features selected to be included in building the scoring models.

is that we are using human clause and sentence labels to create a candidate set for the clause boundary features evaluated by the Stanford Parser and Tregex, as explained in the following subsection.

[8] Feature type: CB=Clause boundary based feature type, PT=Parse tree based feature type

[9] A "linguistically meaningful PP" (PP_ling) is defined as a PP immediately dominated by another PP in cases where a preposition contains a noun such as "in spite of" or "in front of". An example would be "she stood in front of a house" where "in front of a house" would be parsed as two embedded PPs but only the top PP would be counted in this case.

[10] A "linguistically meaningful VP" (VP_ling) is defined as a verb phrase immediately dominated by a clausal phrase, in order to avoid VPs embedded in another VP, e.g., "should go to work" is identified as one VP instead of two embedded VPs.

[11] The "P-based Sampson" is a raw production-based measure (Sampson, 1997), defined as "proportion of the daughters of a nonterminal node which are themselves nonterminal and nonrightmost, averaged over the nonterminals of a sentence".

Clause and Sentence based Features (CB features)

Firstly, we extracted all 26 initial CB features directly from human annotated data of NN-train, using information from the clause and sentence type labels. The reasoning behind this was to create an initial pool of clause-based features that reflects the distribution of clauses and sentences as accurately as possible, even though we did not plan to use this extraction method operationally, where the parser decides on clause and sentence types. After computing the values of each CB feature, we calculated correlations between each feature and human-rated scores. Then we created an initial CB feature pool by selecting features that met two criteria: (1) the absolute Pearson correlation coefficient with human scores was larger than 0.2; and (2) the mean value of the feature on non-native speakers was at least 20% lower than that for na-

tive speakers in case of positive correlation and at least by 20% higher than for native speakers in case of negative correlation, using the Nat data set for the latter criterion. Note that all of these features were computed without using a parser. This resulted in 13 important features.

Secondly, Tregex rules were developed based on Lu's tool to extract these 13 CB features from parsing results where the parser is provided with one sentence at a time. By applying the same selection criteria as before, except for allowing for correlations above 0.1 and giving preference to linguistically more meaningful features, we found 8 features that matched our criteria:
MLS, MLT, DC/C, SSfreq, MLSS, ADJCfreq, Ffreq, MLCC

All 28 pairwise inter-correlations between these 8 features were computed and inspected to avoid including features with high inter-correlations in the scoring model. Since we did not find any inter-correlations larger than 0.9, the features were considered moderately independent and none of them were removed from this set so it also maintains linguistic richness for the feature set.

Due to the importance of T-units in complexity analysis, we briefly introduce how we obtain them from annotations. Three types of clauses labeled in our transcript can serve as T-units, including simple sentences, independent clauses, and conjunct (coordination) clauses. These clauses were identified in the human-annotated text and extracted as T-units in this phase. T-units in parse trees are identified using rules in Lu's tool.

Parse Tree based Features (PT features)

We evaluated 65 features in total and selected features with highest importance using the following two criteria (which are very similar as before): (1) the absolute Pearson correlation coefficient with human scores is larger than 0.2; and (2) the feature mean value on native speakers (Nat) is higher than on score 4 for non-native speakers in case of positive correlation, or lower for negative correlation. 20 of 65 features were found to meet the requirements.

Next, we examined inter-correlations between these features and found some correlations larger than 0.85.[12] For each feature pair exhibiting high inter-correlation, we removed one feature according to the criterion that the removed feature should be linguistically less meaningful than the remaining one. After this filtering, the 9 remaining PT features are:
CT/T, PP_ling/S, NP/S, CN/S, VP_ling/T, PAS/S, DI/T, MLev, MPSam

In summary, as a result of the feature selection process, a total of 17 features were identified as important features to be used in scoring models for predicting speakers' proficiency scores. Among them 8 are clause boundary based and the other 9 are parse tree based.

5 Experiments and Results

In the previous section, we identified 17 syntactic features that show promising correlations with human rater speaking proficiency scores. These features as well as the human-rated scores will be used to build scoring models by using machine learning methods. As introduced in Section 3, we have one training set (N=137 speakers with all of their responses combined) for model building and five testing sets (N=52 for each of them) for evaluation.

The publicly available machine learning package Weka was used in our experiments (Hall et al. 2009). We experimented with two algorithms in Weka: multiple regression (called "LinearRegression" in Weka) and decision tree (called "M5P"in Weka). The score values to be predicted are real numbers (i.e., non-integer), because we have to compute the average score of one speaker's responses. Our initial runs showed that decision tree models were consistently outperformed by multiple regression (MR) models and thus decided to only focus on MR models henceforth.

We set the "AttributeSelectionMethod" parameter in Weka's LinearRegression algorithm to all 3 of its possible values in turn: (Model-1) M5 method; (Model-2) no attribute selection; and (Model-3) greedy method. The resulting three multiple regression models were then tested against the five testing sets. Overall, correlations for all models for the NN-test-1-Hum set were between 0.45 and 0.49, correlations for sets NN-test-2-CB and NN-

[12] The reason for using a lower threshold than above was to obtain a roughly equal number of CB and PT features in the end.

test-3-SB (human transcript based, and using automated boundaries) around 0.2, and for sets NN-test-4-ASR-CB and NN-test-5-ASR-SB (ASR hypotheses, and using automated boundaries), the correlations were not significant. Model-2 (using all 17 features) had the highest correlation on NN-test-1-Hum and we provide correlation results of this model in Table 3.

Test set	Correlation coefficient	Correlation significance ($p < 0.05$)
NN-test-1-Hum	0.488	Significant
NN-test-2-CB	0.220	Significant
NN-test-3-SB	0.170	Significant
NN-test-4-ASR-CB	-0.025	Not significant
NN-test-5-ASR-SB	-0.013	Not significant

Table 3. Multiple regression model testing results for Model-2.

6 Discussion

As we can see from the result table (Table 3) in the previous section, using only syntactic complexity features, based on clausal or parse tree information derived from human transcriptions of spoken test responses, can predict holistic human rater scores for combined speaker responses over a whole test with an overall correlation of r=0.49. While this is a promising result for this study with a focus on a broad spectrum of syntactic complexity features, the results also show significant limitations for an immediate operational use of such features. First, the imperfect prediction of clause and sentence boundaries by the two automatic classifiers causes a substantial degradation of scoring model performance to about r=0.2, and secondly, the rather high error rate of the ASR system (50.5%) does not allow for the computation of features that would result in any significant correlation with human scores. We want to note here that while ASR systems can be found that exhibit WERs below 10% for certain tasks, such as restricted dictation in low-noise environments by native speakers, our ASR task is significantly harder in several ways: (1) we have to recognize non-native speakers'rresponses where speakers have a number of different native language backgrounds; (2) the proficiency level of the test takers varies widely; and (3) the responses are spontaneous and unconstrained in terms of vocabulary.

As for the automatic clause and sentence boundary classifiers, we can observe (in Table 4) that although the sentence boundary classifier has a slightly higher F-score than the clause boundary classifier, errors in sentence boundary detection have more negative effects on the accuracy of score prediction than those made by the clause boundary classifier. In fact, the lower F-score of the latter is mainly due to its lower precision which indicates that there are more spurious clause boundaries in its output which apparently cause little harm to the feature extraction processes.

Among the 17 final features, 3 of them are frequency-based and the remaining 14 are ratio-based, which mirrors our findings from previous work that frequency features have been used less successfully than ratio features. As for ratio features, 5 of them are grammatical structure counts against sentence units, 4 are counts against T-units, and only 1 is based on counts against clause units. The feature set covers a wide range of grammatical structures, such as T-units, verb phrases, noun phrases, complex nominals, adjective clauses, coordinate clauses, prepositional phrases, etc. While this wide coverage provides for richness of the construct of syntactic complexity, some of the features exhibit relatively high correlation with each other which reduces their overall contributions to the scoring model's performance.

Going through the workflow of our system, we find at least five major stages that can generate errors which in turn can adversely affect feature computation and scoring model building. Errors may appear in each stage of our workflow, passing or even enlarging their effects from previous stages to later stages:

1) grammatical errors by the speakers (test takers);
2) errors by the ASR system;
3) sentence/clause boundary detection errors;
4) parser errors; and
5) rule extraction errors.

In future work we will need to address each error source to obtain a higher overall system performance.

Classifier	Accuracy	Precision	Recall	F score
Clause boundary	0.954	0.721	0.748	0.734
Sentence boundary	0.975	0.811	0.755	0.782

Table 4. Performance of clause and sentence boundary detectors.

7 Conclusion and Future Work

In this paper, we investigated associations between speakers' syntactic complexity features and their speaking proficiency scores provided by human raters. By exploring empirical evidence from non-native and native speakers' data sets of spontaneous speech test responses, we identified 17 features related to clause types and parse trees as effective predictors of human speaking scores. The features were implemented based on Lu's L2 Syntactic Complexity Analyzer toolkit (Lu, 2011) to be automatically extracted from human or ASR transcripts. Three multiple regression models were built from non-native speech training data with different parameter setup and were tested against five testing sets with different preprocessing steps. The best model used the complete set of 17 features and exhibited a correlation with human scores of r=0.49 on human transcripts with boundary annotations.

When using automated classifiers to predict clause or sentence boundaries, correlations with human scores are around r=0.2. Our experiments indicate that by enhancing the accuracy of the two main automated preprocessing components, namely ASR and automatic sentence and clause boundary detectors, scoring model performance will increase substantially, as well. Furthermore, this result demonstrates clearly that syntactic complexity features can be devised that are able to predict human speaking proficiency scores.

Since this is a preliminary study, there is ample space to improve all major stages in the feature extraction process. The errors listed in the previous section are potential working directions for preprocessing enhancements prior to machine learning. Among the five types of errors, we can work on improving the accuracy of the speech recognizer, sentence and clause boundary detectors, parser, and feature extraction rules; as for the grammatical errors produced by test takers, we are envisioning to automatically identify and correct such errors. We will further experiment with syntactic complexity measures to balance construct richness and model simplicity. Furthermore, we can also experiment with additional types of machine learning models and tune parameters to derive scoring models with better performance.

Acknowledgements

The authors wish to thank Lei Chen and Su-Youn Yoon for their help with the sentence and clause boundary classifiers. We also would like to thank our colleagues Jill Burstein, Keelan Evanini, Yoko Futagi, Derrick Higgins, Nitin Madnani, and Joel Tetreault, as well as the four anonymous ACL reviewers for their valuable and helpful feedback and comments on our paper.

References

Bachman, L.F. (1990). Fundamental considerations in language testing. Oxford: Oxford University Press.

Bernstein, J. (1999). PhonePass testing: Structure and construct. Menlo Park, CA: Ordinate Corporation.

Bernstein, J., DeJong, J., Pisoni, D. & Townshend, B. (2000). Two experiments in automatic scoring of spoken language proficiency. Proceedings of In-STILL 2000, Dundee, Scotland.

Bernstein, J., Cheng, J., & Suzuki, M. (2010). Fluency and structural complexity as predictors of L2 oral proficiency. Proceedings of Interspeech 2010, Tokyo, Japan, September.

Chen, L., Tetreault, J. & Xi, X. (2010). Towards using structural events to assess non-native speech. NAACL-HLT 2010. 5th Workshop on Innovative Use of NLP for Building Educational Applications, Los Angeles, CA, June.

Condouris, K., Meyer, E. & Tagger-Flusberg, H. (2003). The relationship between standardized measures of language and measures of spontaneous speech in children with autism. American Journal of Speech-Language Pathology, 12(3), 349-358.

Cooper, T.C. (1976). Measuring written syntactic patterns of second language learners of German. The Journal of Educational Research, 69(5), 176-183.

Cucchiarini, C., Strik, H. & Boves, L. (1997). Automatic evaluation of Dutch pronunciation by using speech recognition technology. IEEE Automatic Speech Recognition and Understanding Workshop, Santa Barbara, CA.

Cucchiarini, C., Strik, H. & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. Journal of the Acoustical Society of America, 107, 989-999.

Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R. & Butzberger, J. (2000a). The SRI EduSpeak system: Recognition and pronunciation scoring for language learning. Proceedings of InSTiLL-2000 (Intelligent Speech Technology in Language Learning), Dundee, Scotland.

Franco, H., Neumeyer, L., Digalakis, V. & Ronen, O. (2000b). Combination of machine scores for automatic grading of pronunciation quality. Speech Communication, 30, 121-130.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I.H. (2009). The WEKA Data Mining Software: An Update. SIGKDD Explorations, 11(1).

Halleck, G.B. (1995). Assessing oral proficiency: A comparison of holistic and objective measures. The Modern Language Journal, 79(2), 223-234.

Henry, K. (1996). Early L2 writing development: A study of autobiographical essays by university-level students on Russian. The Modern Language Journal, 80(3), 309-326.

Ho-Peng, L. (1983). Using T-unit measures to assess writing proficiency of university ESL students. RELC Journal, 14(2), 35-43.

Hunt, K. (1965). Grammatical structures written at three grade levels. NCTE Research report No.3. Champaign, IL: NCTE.

Iwashita, N. (2006). Syntactic complexity measures and their relations to oral proficiency in Japanese as a foreign language. Language Assessment Quarterly, 3(20), 151-169.

Kameen, P.T. (1979). Syntactic skill and ESL writing quality. In C. Yorio, K. Perkins, & J. Schachter (Eds.), On TESOL '79: The learner in focus (pp.343-364). Washington, D.C.: TESOL.

Klein, D. & Manning, C.D. (2003). Fast exact inference with a factored model for a natural language parsing. In S.Becker, S. Thrun & K. Obermayer (Eds.), Advances in Neural Information Processing Systems 15 (pp.3-10). Cambridge, MA: MIT Press.

Larsen-Freeman, D. (1978). An ESL index of development. Teachers of English to Speakers of Other Languages Quarterly, 12(4), 439-448.

Levy, R. & Andrew, G. (2006). Tregex and Tsurgeon: Tools for querying and manipulating tree data structures. Proceedings of the Fifth International Conference on Language Resources and Evaluation.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. International Journal of Corpus Linguistics, 15(4), 474-496.

Lu, X. (2011). L2 Syntactic Complexity Analyzer. Retrieved from http://www.personal.psu.edu/xxl13/downloads/l2sca.html

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

Perkins, K. (1980). Using objective methods of attained writing proficiency to discriminate among holistic evaluations. Teachers of English to Speakers of Other Languages Quarterly, 14(1), 61-69.

Roll, M., Frid, J. & Horne, M. (2007). Measuring syntactic complexity in spontaneous spoken Swedish. Language and Speech, 50(2), 227-245.

Sampson, G. (1997). Depth in English grammar. Journal of Linguistics, 33, 131-151.

Wolfe-Quintero, K., Inagaki, S. & Kim, H. Y. (1998). Second language development in writing: Measures of fluency, accuracy, & complexity. Honolulu, HI: University of Hawaii Press.

Xi, X., & Mollaun, P. (2006). Investigating the utility of analytic scoring for the TOEFL® Academic Speaking Test (TAST). TOEFL iBT Research Report No. TOEFLiBT-01.

Zechner, K., Higgins, D. & Xi, X. (2007). SpeechRater(SM): A construct-driven approach to score spontaneous non-native speech. Proceedings of the 2007 Workshop of the International Speech Communication Association (ISCA) Special Interest Group on Speech and Language Technology in Education (SLaTE), Farmington, PA, October.

Zechner, K., Higgins, D., Xi, X, & Williamson, D.M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. Speech Communication, 51 (10), October.