Detecting Structural Events for Assessing Non-Native Speech

Lei Chen
Educational Testing Service
Princeton NJ USA
LChen@ets.org

Su-Youn Yoon
Educational Testing Service
Princeton NJ USA
SYoon@ets.org

Abstract

Structural events, (i.e., the structure of clauses and disfluencies) in spontaneous speech, are important components of human speaking and have been used to measure language development. However, they have not been actively used in automated speech assessment research. Given the recent substantial progress on automated structural event detection on spontaneous speech, we investigated the detection of clause boundaries and interruption points of edit disfluencies on transcriptions of non-native speech data and extracted features from the detected events for speech assessment. Compared to features computed on human-annotated events, the features computed on machine-generated events show promising correlations to holistic scores that reflect speaking proficiency levels.

1 Introduction

Spontaneous speech utterances are organized in a structured way and generated dynamically with optional disfluencies. In second language acquisition (SLA) research, information related to the structure of utterances and profile of disfluencies has been widely used to monitor speakers' language development processes (Iwashita, 2006). However, structural events in human conversations have not been actively used in the automated speech assessment research. For example, most research that used Automatic Speech Recognition (ASR) technology to automatically score speaking proficiency (Neumeyer et al., 2000; Zechner et al., 2007) focused on word-level cues for fluency and accuracy.

In the last decade, a large amount of research (Gotoh and Renals, 2000; Shriberg et al., 2000; Liu, 2004; Ostendorf et al., 2008) has been conducted on structural event detection (i.e., sentence and disfluency structure). This research has resulted in better models for structural event detection. The detected structural events have been found to help many of the following natural language processing (NLP) tasks: speech parsing, information retrieval, machine translation, and extractive speech summarization (Ostendorf et al., 2008).

Because structural event information: (1) is important for understanding/processing speech, (2) has been successfully used in monitoring language development, which will be summarized in Section 2, (3) has received limited attention in automated speech assessment, and (4) has been actively investigated in the speech research domain in the past decade, it is worthwhile investigating the utility of using structural event detection on automated speech assessment. Because of the fairly low word accuracy currently achieved when recognizing spontaneous non-native speech of mixed proficiency levels and native language backgrounds, this study will focus on the transcribed words rather than speech recognition outputs.

This paper is organized as follows: Section 2 reviews previous research; Section 3 reports on the data used in the paper, including the collection, scoring, transcription, and annotation processes; Section 4 discusses the methods we utilized for structural event detection; Section 5 describes the experiments of structural event detection; Section 6 described the features derived from the event sequence

Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications, pages 38–45,
Portland, Oregon, 24 June 2011.

for assessing speech and evaluation results on these features; Section 7 discusses the findings of our research and plans for future directions.

2 Previous Research

In the SLA and child language development research fields, language development is measured according to fluency, accuracy, and complexity (Iwashita, 2006). Structural events are used to derive the features measuring syntactic complexity. For example, typical metrics for measuring syntactic complexity include: length of production units (e.g., T-units[1], clauses, verb phrases, and sentences), amount of embedding, subordination and coordination, range of structural types, and structural sophistication. Iwashita (2006) investigated several measures of syntactic complexity on data generated by learners of Japanese. The author reported that some measurements (e.g., T-unit length, the number of clauses per T-unit, and the number of independent clauses per T-unit) were good at predicting learners' proficiency levels.

In addition, speech disfluencies are used to measure language development. For example, Lennon (1990) used a dozen features related to speed, pauses, and several disfluency markers, such as filled pauses per T-unit, to measure the improvement of English proficiency for four German-speaking women during a six-month study in England. He found a significant change in filled pauses per T-unit during the study process.

These two types of features derived from structural events were combined in other previous studies. For example, Mizera (2006) used fluency factors related to speed, voiced smoothness (frequency of repetitions or self-corrections), pauses, syntactic complexity (mean length of T-units), and accuracy, to measure speaking proficiency on 20 non-native English speakers. In this experiment, disfluency-related factors, such as the total number of voiced disfluencies, correlated strongly with the fluency score ($r = -0.45$); however, the syntactic complexity factor only showed a moderate correlation ($r = 0.310$).

There have been previous efforts in using NLP technology to automatically calculate syntactic complexity metrics on learners' writing data. For example, Lu (2009) and Sagae et al. (2005) used parsing to get structural information on written texts; however, such efforts have not been undertaken in assessing speech data.

Chen et al. (2010) annotated structural events (such as clause structure and disfluencies) on English language learners' speech transcriptions and extracted features based on the structural event profile. They found that the features derived from structural event profile show promising correlation to human holistic scores. Berstein et al. (2010) also computed the features related to sentence lengths and the counts of syntactic entities. They found the extracted features were highly correlated to holistic scores measuring test-takers' language proficiency in both English and Spanish.

In the speech research domain, a large amount of research has been conducted to detect structural events in speech transcriptions and recognized words using lexical and prosodic cues. Using a language model (LM) trained on words combined with the events of interest is a popular technique for using textual information for structural event detection. For example, Heeman and Allen (1999) developed a LM including part of speech (POS) tags, discourse markers (e.g., *right, anyway*), speech repairs, and intonational phrases. In this way, structural information (e.g., speech repairs), could be predicted using a traditional speech recognition approach.

Prosodic information has been widely used to further improve textual models. For example, a simple prosodic feature, pause duration between words, was used in Gotoh and Renals (2000) to detect sentence boundaries. It was found that the pause duration model alone was better than using an LM alone, and the combination of the two models further improved the performance.

More advanced prosody models were used in other research on sentence boundary and speech repair detections (Shriberg et al., 2000; Shriberg and Stolcke, 2004). A general framework was built combining textual and prosodic cues to detect various kinds of structural events in speech, including sentence boundaries, disfluencies, topic boundaries, dialog acts, emotion, etc. Shriberg and Stolcke (2004) extracted prosodic features such as pause, phone du-

[1] A T-unit is defined as essentially a main clause plus any other clauses which are dependent upon it (Hunt, 1970).

ration, rhyme duration, and F_0 features. Using all of these features, a decision tree was built to detect possible structural events. An LM augmented with structural event tokens was also used to detect structural events based on textual cues. Finally, a Hidden Markov Model (HMM) was used to combine estimations from the textual model (an augmented LM with structural events) and prosodic model (decision-tree based on prosodic features).

Research on structural event detection has been strongly affected by the DARPA EARS program (EARS, 2002). As in Shriberg et al. (2000), the structural event detection (e.g., sentence units (SUs) and speech repairs) investigated in EARS was a classification task utilizing both prosodic and textual knowledge sources. New approaches for combining the two knowledge sources, including maximum entropy (MaxEnt) and conditional random fields (CRFs), were studied to address the weaknesses of the generative HMM approach (Liu et al., 2004). Liu et al. (2005) concluded that "adding textual information, building a more robust prosodic model, using conditional modeling approaches (Maxent and CRF), and system combination all yield performance gains."

3 Non-native Structural Event Corpus

Non-native speech data were collected from the TOEFL Practice Test Online (TPO) (ETS, 2006). In each TPO test, test-takers were required to respond to six speaking test items, in which they were required to provide information or opinions on familiar topics, based on their personal experience or background knowledge. For example, the test-takers were asked to describe their opinions about living on or off campus.

A total of 1066 responses were collected from examinees. Then, a group of experienced human raters scored these items based on the scoring rubrics designed for scoring the TPO test. For each item, two human raters independently assigned 4-point holistic scores for test-takers' English proficiency levels.

The speaking content was transcribed by a professional transcribing agency. On the transcriptions, structural event annotations were added, including (1) locations of clause boundaries, (2) types of clauses (e.g., noun clauses, adjective clauses, adverb clauses, etc.), and (3) disfluencies.

Disfluencies can further be sub-classified into several groups: silent pauses, filled pauses (e.g., *uh* and *um*), false starts, repetitions, and repairs. The repetitions and repairs were denoted as "*edit disfluency*", which were comprised of a *reparandum*, an optional *editing term*, and a *correction*. The *reparandum* is the part of an utterance that a speaker wants to repeat or change, while the *correction* contains the speaker's correction. The *editing term* can be a filled pause (e.g., *um*) or an explicit expression (e.g., *sorry*). The interruption point (IP), occurring at the end of the reparandum, is where the fluent speech is interrupted to prepare for the correction.

For the research reported in this paper, we focus on two structural events: the locations of clause-ending boundaries (CBs) and interruption points (IPs) of edit disfluencies. Note that if several clauses (in different layers of a clause hierarchy) end at the same word boundary, these clause boundaries were collapsed into one CB event.

Two persons annotated the corpus separately and their annotation quality was monitored by using several Kappa computations. For CBs, κ ranges from 0.85 to 0.90; for IPs, κ ranges from 0.63 to 0.83. Generally, a κ greater than 0.8 indicates a good between-rater agreement and κ in the range of 0.6 to 0.8 indicates acceptable agreement (Landis and Koch, 1977). Therefore, we believe that our human annotations are sufficiently reliable to be used in the following experiments.

4 Methods of Structural Event Detection

4.1 Features for structural event detection

In previous research (Gotoh and Renals, 2000; Shriberg et al., 2000; Liu, 2004), prosodic cues were found to be helpful, however, such findings on native speech data may not work well with non-native speech data. Anderson-Hsieh and Venkatagiri (1994) compared the pause frequencies of three groups of speakers (native, high-scoring, and low-scoring non-native speakers). They found that pause frequency was higher for groups of speakers with lower speaking skills. For native speakers, a long pause after a word-ending boundary is an important cue for signaling the existence of a sentence or clause boundary. However, the fact that there are

more frequent pauses in non-native speech obscures this relationship.

On our non-native speech corpus, we conducted a pilot study on a widely-used prosodic feature, the pause duration[2] after a word, for its predictive ability to detect clause boundaries. If the duration of the pause after a word boundary is longer than 0.15 second, we call it a *long pause*. We measured the likelihood of being a CB event on the words followed by a *long pause*. For each score level, the likelihoods are: 15% for a score of 1, 22% for a score of 2, 28% for a score of 3, and 35% for a score of 4. Clearly, for low-proficiency speakers (i.e., speakers with a score of 1), long pauses in their utterances are not tightly linked to CBs. Therefore, more research is needed to utilize prosodic cues on non-native speech; in this paper, we focus on lexical features.

4.2 Statistical models

Based on lexical features, the structural event detection task can be generalized as follows:

$$\hat{E} = arg\max_{E} P(E|W)$$

Given that E denotes the between-word event sequence and W denotes the corresponding lexical cues, the goal is to find the event sequence that has the greatest probability, given the observed features.

Recently, conditional modeling approaches were successfully used in sentence units (SUs) and speech repairs detection (Liu, 2004). Hence, we use the Maximum Entropy (MaxEnt) (Berger et al., 1996) and Conditional Random Fields (CRFs) (Lafferty et al., 2001) approaches to build statistical models for structural event detection.

5 Structural Event Detection Experiment

5.1 Setup

In our experiment, the whole corpus described in Section 3 was split into a training set (*train*), a development test set (*dev*), and testing set (*test*), without speaker overlap between any pair of sets. Table 1 summarizes the numbers of items and words, as well as structural events of each dataset.

[2]Pause durations were obtained by running forced alignment using speech and transcriptions on a tri-phone HMM speech recognizer

	train	*dev*	*test*
# item	664	101	301
# word	71523	10509	33754
# CB	6121	918	2852
# IP	1767	267	1112

Table 1: The number of items, words, and structural events of the three sets in the TPO corpus

On average, each item contains about 108.6 words, 9.3 CBs, and 3.0 IPs. 9% of the word boundaries are associated with a CB event and 3% of the word boundaries are associated with an IP event. Clearly, these CB and IP events are sparse and such a skewed distribution of structural events increases the difficulty of structural event detection.

5.2 Models

The following two conditional models were built to detect CB and IP events:

- **MaxEnt**: Given w_i as the word token at position i, the word n-gram features include: $\langle w_i \rangle$, $\langle w_{i-1}, w_i \rangle$, $\langle w_i, w_{i+1} \rangle$, $\langle w_{i-2}, w_{i-1}, w_i \rangle$, $\langle w_i, w_{i+1}, w_{i+2} \rangle$, and $\langle w_{i-1}, w_i, w_{i+1} \rangle$. Given t_i as the POS tag[3] at position i, the POS n-gram features include: $\langle t_i \rangle$, $\langle t_{i-1}, t_i \rangle$, $\langle t_i, t_{i+1} \rangle$, $\langle t_{i-2}, t_{i-1}, t_i \rangle$, $\langle t_i, t_{i+1}, t_{i+2} \rangle$, and $\langle t_{i-1}, t_i, t_{i+1} \rangle$.

  For IP detection, in addition to the n-gram features described above, another four features that capture syntactic pattern of disfluencies are utilized:

  - **filled pause adjacency**: This feature has a binary value showing whether a filled pause such as *uh* or *um* was adjacent to the current word (w_i).
  - **word repetition**: This feature has a binary value showing whether the current word (w_i) was repeated in the following 5 words or not.

[3]POS tags were obtained by tagging words using a MaxEnt POS tagger, which was implemented in the OpenNLP toolkit and trained on the Switchboard (SWBD) corpus. This POS tagger was trained on about $528K$ word/tag pairs and achieved an tagging accuracy of 96.3% on a test set of $379K$ words.

- **similarity**: This feature has a continuous value which measures the similarity between the reparandum and correction. Assuming that w_i was the end of the reparandum, the start point and the end point of the reparandum and correction were estimated, and the string edit distance between the reparandum and correction was calculated. The start point and the end point of the reparandum and correction were estimated as follows; if w_i appeared in the following 5 words, the second occurrence was defined as the end of the correction. Otherwise, w_{i+5} was defined as the end of correction. Secondly, N, the length of the correction was calculated, and w_{i-N+1} was defined as the start point of the reparandum. During the calculation of the string edit distance, a word fragment was considered to be the same as a word whose initial character sequences matched it.
- **length of correction**: This feature counts the number of words in the correction.

The first two features are similar to the features used in (Liu, 2004) while the last two features provide important keys in distinguishing edit disfluencies from fluent speech. Since the correction is composed of word sequences that are similar to the reparandum, these two features are higher than zero when the target word is a part of the edit disfluency. In addition, these two numeric features were discretized by using an equal-distance binning approach.

Using n-gram features for CB detection and all these lexical features for IP detection, we used the Maxent toolkit designed by Zhang (2005) to build MaxEnt models. The L-BFGS parameter estimation method is used, with the Gaussian-prior smoothing technique to avoid over-fitting. The Gaussian prior is estimated on the *dev* set.

- **CRF**: All features which were described in building MaxEnt models were used in the CRF model. We used the Java-based NLP package Mallet (McCallum, 2005) to build CRF models. Similar to MaxEnt models, Gaussian-prior smoothing was used with the priors estimated on the *dev* set.

These models were trained using the *train* set. Besides Gaussian priors, other parameters in the model training (i.e., the training iteration number as well as the cutting-point for event decisions) were estimated using the *dev* set. Finally, the trained models were evaluated on the *test* set.

5.3 Evaluation of event detection

Since structural event detection was treated as a classification task in this paper, four standard evaluation metrics were used:

$$\begin{aligned} accuracy &= \frac{TP+TN}{TP+FP+TN+FN} \\ precision &= \frac{TP}{TP+FP} \\ recall &= \frac{TP}{TP+FN} \\ F1 &= 2 \times \frac{recall \times precision}{recall + precision} \end{aligned}$$

where, TP and FP denote the number of true positives and false positives, and TN and FN denote the number of true negatives and false negatives. A structural event (a CB or IP boundary) is treated as a positive class. In our experiment, since we treated precision and recall as equally important, the $F1$ measurement was used.

For each model, if the estimated probability, $P(E_i|W)$, is larger than a threshold, the corresponding word boundary will be estimated to be a positive class. The threshold was chosen when a maximal $F1$ score was achieved on the *dev* set.

A model that always predicts the majority class (a no-event in this study) was treated as a baseline model. For CB detection, this type of baseline model resulted in an accuracy of 91.6%; for IP detection, this type of baseline model resulted in an accuracy of 96.7%.

5.4 Results of structural event detection

Table 2 summarizes the performance of the two models on the CB and IP detection tasks.

For CB detection, two conditional models are superior to the baseline CB detection (with an accuracy of 91.6%); they achieved relatively high $F1$ scores

	Acc.	Pre.	Rec.	$F1$
		CB		
MaxEnt	94.5	66.1	71.8	0.689
CRF	96.1	82.3	68.6	0.749
		IP		
MaxEnt	98.1	61.8	55.2	0.583
CRF	98.4	76.9	48.0	0.591

Table 2: Experimental results of the CB and IP detection measurement using accuracy (Acc.), precision (Pre.), recall (Rec.) and $F1$ measurement ($F1$) on the TPO data

ranging from 0.689 to 0.749. Between the two models, the CRF model achieved the higher $F1$ score at 0.749, The lower F-score of the MaxEnt model may be caused by the fact that the MaxEnt model does not use event history information in its decoding process.

However, these two models achieved lower performance on the task of detecting IPs for editing disfluencies. F-scores became about 0.58 to 0.59 for IP detections. The degraded performance may be caused by the extremely low IP distribution (only 3%) in our data. Between the two modeling approaches, consistent with the result shown for CB detection, the CRF model achieved a higher $F1$ score (0.591).

6 Using Detected Structural Events for Speech Assessment

6.1 Features assessing proficiency

Many previous SLA studies used the length of production units and frequency of disfluencies as metrics to measure language development (Iwashita, 2006; Lennon, 1990; Mizera, 2006). Our automated structural event detection provides the locations of CBs and IPs, which can be used to compute these features for use in speech assessment.

Using N_w to represent the total number of words in the spoken response (without pruning the reparandums and edit terms in the edit disfluencies), N_C as the total number of CBs, and N_{IP} as the total number of IPs detected on transcriptions of speech streams, the following features (i.e, *mean length of clause* (MLC), *interruption points per clause* (IPC), and *interruption points per word* (IPW)) were derived:

$$\begin{aligned} MLC &= N_w/N_C \\ IPC &= N_{IP}/N_C \\ IPW &= N_{IP}/N_w \end{aligned}$$

The IPW can be treated as the IPC normalized by the MLC. The reason for this normalization is that disfluency behavior is influenced by various factors, such as speakers' proficiency levels as well as the difficulty of utterances' structure. For example, Roll et al. (2007) found that the complexity of expression, computed based on the language's parsing-tree structure, influenced the frequency of disfluencies in their experiment on Swedish responses. Therefore, the fact that IPW is the IPC normalized by MLC (a feature related to complexity of utterances' structure) helps to reduce the impact of utterances' structure and to highlight contributions from the speaker's proficiency.

6.2 Results of measuring the derived features

On the *test* set, we produced CB and IP event sequences estimated by the MaxEnt and CRF models, respectively. These machine-generated events were evaluated by comparison with human annotations, which were denoted as REF.

The proposed features described in Section 6.1 were computed on the word/event sequence of each item. In addition, given the fact that each item only covers approximately one-minute of speech and the content is quite limited, we also extracted features on the test-taker level by combining the detected events of all of the items spoken by each test-taker. Then, according to the score handling protocol used in TPO, the human-holistic scores from the first human rater were used as item scores to compute Pearson correlation coefficients (rs) with the features. For the test-taker level evaluation, we used the average score for each test-taker from all of his/her item scores.

Table 3 reports on the evaluation results of the features derived from the structural event estimations. Compared to rs computed on the speaker level using multiple (as many as 6) items, rs computed on the item level are generally lower. This is because words and events are limited in this one-minute long response. Among the three features, the

Model	r_{MLC}	r_{IPC}	r_{IPW}
Per item			
REF	0.003	−0.369	−0.402
MaxEnt	−0.012	−0.329	−0.343
CRF	−0.042	−0.328	−0.335
Per speaker			
REF	0.066	−0.453	−0.516
MaxEnt	0.055	−0.396	−0.417
CRF	0.043	−0.355	−0.366

Table 3: Correlation coefficients (rs) between the features derived from structural events with human scores on the item and speaker levels

MLC shows the lowest r to human holistic scores. In contrast, the two features derived from interruption points show promising rs to human holistic scores. Between them, the IPW always shows a higher r than the IPC. Compared to the features extracted on human annotations, the features derived from structural events automatically estimated by the two NLP models show a lower but sufficiently high r. The features derived from the MaxEnt model's estimations on the test-taker level show a greater r than the features derived from the CRF model estimations.

7 Discussion

Three features measuring syntactic complexity and disfluency profile of speaking, MLC, IPC, and IPW, were extracted on the structural event sequences estimated by the developed models. Compared to the features extracted from the human-annotated structural events, the features derived from machine-generated event sequences show promisingly close correlations.

Applying automated structural event detection to spontaneous speech brings many benefits for automatic speech assessment. First, obtaining information beyond the word level, such as the structure of clauses and disfluencies, can expand and improve the construct[4] coverage of speech features. Second, knowing the structure of utterances helps to facilitate the application of more NLP processing methods (e.g., collocation detection that requires information about sentence boundaries), to speech content. In this study, using only simple word and POS based n-gram features, CBs can be detected relatively well (with an $F1$ score of approximately 0.70). More lexical features reflecting repair properties were found to help improve IP detection performance. In addition, IP-based features derived from machine-generated event sequences show promising correlation with human holistic scores. Results in detection of clause boundaries and interruption points support the approach of utilizing automated structural event detection on speech assessment.

We plan to continue our research in the following three directions. First, we will investigate integrating prosodic cues to further improve the structural event detection performance on non-native speech. Second, we will investigate estimating structural events directly on speech recognition results. Third, other aspects of syntactic complexity, such as the embedding of clauses, will be studied to provide a broader set of features for speech assessment.

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

References

J. Anderson-Hsieh and H. Venkatagiri. 1994. Syllable duration and pausing in the speech of chinese ESL speakers. *TESOL Quarterly*, pages 807–812.

A. Berger, S. Pietra, and V. Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22:39–72.

J. Berstein, J. Cheng, and M. Suzuki. 2010. Fluency and Structural Complexity as Predictors of L2 Oral Proficiency. In *Proc. of InterSpeech*.

L. Chen, J. Tetreault, and X. Xi. 2010. Towards using structural events to assess non-native speech. In *Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, page 74.

EARS. 2002. DARPA EARS Program. `http://projects.ldc.upenn.edu/EARS/`.

ETS. 2006. TOEFL Practice Online Test (TPO).

Y. Gotoh and S. Renals. 2000. Sentence boundary detection in broadcast speech transcript. In *Proceedings of the International Speech Communication Association (ISCA) Workshop: Automatic Speech Recognition: Challenges for the new Millennium ASR-2000*.

P. Heeman and J. Allen. 1999. Speech repairs, intonational phrased and discourse markers: Modeling speakers' utterances in spoken dialogue. *Computational Linguistics*.

K. W. Hunt. 1970. Syntactic maturity in school children and adults. In *Monographs of the Society for Re-*

search in Child Development. University of Chicago Press, Chicago, IL.

N. Iwashita. 2006. Syntactic complexity measures and their relation to oral proficiency in Japanese as a foreign language. *Language Assessment Quarterly: An International Journal*, 3(2):151–169.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random field: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning (ICML)*.

J. R Landis and G. G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

P. Lennon. 1990. Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40(3):387–417.

Y. Liu, A. Stolcke, E. Shriberg, and M. Harper. 2004. Comparing and combining generative and posterior probability models: Some advances in sentence boundary detection in speech. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*.

Y. Liu, E. Shriberg, A. Stolcke, B. Peskin, J. Ang, Hillard D., M. Ostendorf, M. Tomalin, P. Woodland, and M. Harper. 2005. Structural Metadata Research in the EARS Program. In *Proceedings of the International Conference of Acoustics, Speech, and Signal Processing (ICASSP)*.

Y. Liu. 2004. *Structural Event Detection for Rich Transcription of Speech*. Ph.D. thesis, Purdue University.

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

A. McCallum. 2005. Mallet: A machine learning toolkit for language. `http://mallet.cs.umass.edu`.

G. J. Mizera. 2006. *Working memory and L2 oral fluency*. Ph.D. thesis, University of Pittsburgh.

L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub. 2000. Automatic Scoring of Pronunciation Quality. *Speech Communication*, 30:83–93.

M. Ostendorf, B. Favre, R. Grishman, D. Hakkani-Tur, M. Harper, D. Hillard, J. Hirschberg, Heng Ji, J.G. Kahn, Yang Liu, S. Maskey, E. Matusov, H. Ney, A. Rosenberg, E. Shriberg, Wen Wang, and C. Woofers. 2008. Speech segmentation and spoken document processing. *Signal Processing Magazine, IEEE*, 25(3):59–69, May.

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

K. Sagae, A. Lavie, and B. MacWhinney. 2005. Automatic measurement of syntactic development in child language. In *Proc. of ACL*, volume 100.

E. Shriberg and A. Stolcke. 2004. Direct modeling of prosody: An overview of applications in automatic speech processing. In *Proceedings of the International Conference on Speech Prosody*.

E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1-2):127–154.

K. Zechner, D. Higgins, and Xiaoming Xi. 2007. SpeechRater: A Construct-Driven Approach to Scoring Spontaneous Non-Native Speech. In *Proc. SLaTE*.

L. Zhang. 2005. Maximum Entropy Modeling Toolkit for Python and C++. `http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html`.