Using an Automated Content Scoring System for Spoken CALL Responses: The ETS submission for the Spoken CALL Challenge

Keelan Evanini[†], *Matthew Mulholland*[†], *Eugene Tsuprun*[†], *Yao Qian* [‡]

Educational Testing Service R&D
[†]660 Rosedale Rd., Princeton, NJ, USA
[‡]90 New Montgomery Street, Suite 1500, San Francisco, CA, USA
{kevanini,mmulholland,etsuprun,yqian}@ets.org

Abstract

In this study we investigate the performance of an automated content scoring system for accepting or rejecting learner responses in a spoken CALL application. Specifically, we employed a system based on word and character *n*-gram features in a support vector machine learning framework that was originally designed for scoring content in written texts and augmented its standard feature set with additional features from the following categories: prompt bias, text-to-text similarity to reference responses, and automatically detected grammatical errors. This system achieved a D score of 4.353 (compared to a baseline score of 1.694) on the test set consisting of Kaldi ASR output in the 2017 Spoken CALL Challenge. In this paper we also provide an analysis of the impact of the size and nature of the training data set (human transcriptions vs. ASR output) on the model's performance and present the results of feature ablation experiments to demonstrate which of the additional features are most helpful.

Index Terms: Spoken CALL Challenge, automated content scoring, bias features, text-to-text similarity features, grammar features

1. Introduction

Many studies have demonstrated the positive effects of targeted grammar feedback provided by language instructors to language learners in a classroom environment on the learners' ability to produce grammatically correct utterances [1, 2, 3, 4], and several types of feedback have been investigated, including prompts (repeating the question, providing metalinguistic feedback, etc.) and reformulation strategies (recasting the learner's response, providing explicit corrections, etc.). Recent improvements in automatic speech recognition (ASR), natural language processing (NLP), and spoken dialog system (SDS) infrastructure have enabled the creation of interactive, speech-based Computer-Assisted Language Learning (CALL) applications that attempt to automate the process of providing grammar feedback to language learners [5, 6, 7]. In order for these automated spoken CALL applications to provide valid grammar feedback to language learners, it is necessary for them to accurately detect erroneous responses. Since this field of research is relatively new, and since few shared resources exist for comparing various error detection methodologies on a common data set, [8] proposed a shared task for spoken CALL, in which spoken English responses provided by native speakers of German while using a CALL application would be released to the community along with annotations about the grammatical and semantic correctness of each response that can be used to train models for predicting whether the responses are erroneous or not.[1] This paper describes the system that ETS developed to participate in this shared task at SLaTE 2017.

To address the task of determining whether to accept or reject a spoken response produced in the context of a spoken CALL application, we explore the use of a pre-existing automated content scoring system developed at ETS. This system uses a machine learning approach based primarily on character and word *n*-gram features and has been applied to score content in a wide variety of tasks. Most research with this content scoring system has been conducted in the context of short answer tasks for the domains of elementary and secondary schools in areas such as science, English language arts, and math [9, 10]; however, it has also been successfully applied to longer written texts from other domains, such as a writing task from a standardized assessment for music teachers [11]. In addition, a recent study explored the use of this content scoring system for automated scoring of non-native spoken English responses provided in the context of a standardized assessment of English speaking proficiency [12]. That study is not directly comparable to the current study, though, since the nature of the spoken responses and the scores across the two studies are quite different. The spoken responses in the Spoken CALL Shared Task quite short (typically a single sentence) and were scored for grammar and meaning only, whereas the spoken responses in [12] consisted of spontaneous speech of approximately one minute in duration and were scored by human raters based on scoring rubrics that contained additional aspects of speaking proficiency, such as pronunciation and fluency.

This paper is organized as follows: in Section 2 we describe the features contained in the content scoring system that was used to train the prediction models as well as the additional types of features that were investigated; Section 3 presents two experiments that were conducted on the training set in order to determine the relative importance of the various additional features as well as the impact of the quality and quantity of training data on the model's performance; the official test results that were obtained by our submission to the Spoken CALL Shared Task are presented in Section 4 along with additional experiments on the test set that were conducted after the official submission deadline for the shared task; finally, Section 5 discusses the main findings and suggests steps for future research.

2. Features

We explored the following four types of features in our models: features extracted using the automated content scoring system,

[1] Further details about the Spoken CALL Shared Task are available here: https://regulus.unige.ch/spokencallsharedtask/.

10.21437/SLaTE.2017-17

bias features based on the prompts and prompt categories, features based on the similarity between a speaker's utterance and the sample responses for each prompt, and features produced by a grammar checker. These features are described in more detail in the following sections.

2.1. Content Features

As described in Section 1, we used a pre-existing system for automated content scoring to extract four classes of features:

- Character *n*-grams for *n* = 2 to 5
- Token unigrams and bigrams
- Syntactic dependencies
- Length of response in characters

Each feature is represented in a sparse binary format. For example, the *n*-gram and syntactic dependency features have a value of 1 if present in a response and an implicit 0 if not. The length feature, however, is composed of a set of length bin features. Specifically, length is calculated for each response using the following formula (where "length" is the number of characters in the response): $log_2(1+length)$. Thus, there are features for different lengths and these features are represented in the same sparse format as the other features mentioned above. Syntactic dependencies were extracted using the ZPar dependency parser [13].

These features are referred to as "Content" features in this paper. Since the motivation for this study was to examine the performance of the automated content scoring system for the task of accepting / rejecting short responses provided in the context of a spoken CALL application, the Content features were included in all of the models that we experimented with, and the additional features described in the subsequent sections were added to augment the model's performance. Since it was not possible to train prompt-specific models for this task (see Section 2.2), generic models were trained to apply across all prompts with the expectation that the Content features could capture common patterns of linguistic errors that are shared across different prompts.

2.2. Prompt Bias

When the content scoring system is applied to score responses to test questions, a separate model is typically trained for each prompt [9]. While a few of the prompts in the Spoken CALL Task contain a reasonable number of responses for training robust models in the training data set—four prompts contain more than 100 responses with the largest number, 195, corresponding to the prompt *Sag: Ich habe keine Reservation* (*Say: I don't have a reservation*)—most of the 413 prompts do not contain enough responses per prompt to train models. In fact, 44 of the prompts only contained a single response in the training set, e.g., *Frag: 3 Tickets für Mamma Mia* (*Ask for: 3 tickets for Mamma Mia*) and *Sag: Ich möchte um 22 Uhr morgen abreisen* (*Say: I want to leave tomorrow at 10pm*). Furthermore, 11 out of the 264 prompts in the test set, e.g., *Frag: 2 Tickets für Montagabend* (*Ask for: 2 tickets for Monday evening*) and *Frag: pinkige Hosen* (*Ask for: pink trousers*) had no corresponding responses in the training set, so it would have been impossible to train prompt-specific models for these prompts.

Therefore, we adopted an approach in which a single model was trained using data from all of the prompts in the training set. In order to enable the model to be able to leverage information about prompt-specific grammar and vocabulary patterns that correspond to different scores, we explored the use of prompt bias features in the model. The prompt bias features consisted of a single binary feature per prompt per response represented in a sparse matrix. These features were inspired by the approach taken by [14] which used the content scoring system for Task 7 in the SemEval 2013 shared task.

In addition to the prompt bias features, we also explored the use of bias features based on categories of prompts that are similar to one another. The motivation for this approach was the observation that many of the prompts are expected to elicit responses that are similar to each other in form and differ only based on a small number of key words. These similarities among expected resposnes for certain prompts could be leveraged by the bias features to reduce data sparsity; in addition, the prompt category bias features could help address the fact mentioned above that some responses in the test set were drawn from prompts that were not contained in the training set. An example of one of the prompt categories that was designed for this study can be seen in the many prompts in the data set that are related to the communicative task of buying clothes, specifically, making a request to buy a particular item of clothes. All of the sample responses listed in the reference grammar for these prompts are identical except for the specific item of clothes that was specified in the prompt. For example, Table 1 shows some of the 161 sample responses for two prompts in this category: *Frag: blaue Sandalen* (*Ask for: blue sandals*) and *Frag: braune Stiefel* (*Ask for: brown boots*); in the table, the content words that are specific to each prompt and are expected to differ are underlined.

Frag: blaue Sandalen (*Ask for: blue sandals*)	*Frag: braune Stiefel* (*Ask for: brown boots*)
blue sandals	brown boots
blue sandals please	brown boots please
can i buy blue sandals	can i buy brown boots
can i buy blue sandals please	can i buy brown boots please
...	...
yes could you please give me blue sandals	yes could you please give me brown boots
...	...
yes i'd like blue sandals please	yes i'd like brown boots please

Table 1: *Sample responses for two prompts in the* `buying_clothes` *category*

In total, 49 different prompt categories were defined manually by the authors for the 487 prompts listed in the reference grammar provided by the task organizers. A few of these prompt categories contained a large number of responses; Table 2 lists the 5 most frequent along with the number of prompts associated with each and a few examples.

On the other hand, some of the prompts were not closely related to any of the other prompts, such as *Frag: Welcher Bus fährt dorthin?* (*Ask: Which bus goes there?*) and *Sag: Mir gefällt die Farbe nicht* (*Say: I don't like the color*). In these cases, the prompts were not grouped together with any other prompts and the prompt category label was identical to the prompt label; 22 prompts were treated in this manner for

Prompt Category	N	Samples
`buying_tickets`	106	*Frag: Tickets für Sonntagabend* (*Ask for: tickets for Sunday evening*), *Frag: ein Ticket für heute Abend* (*Ask for: one ticket for this evening*), ...
`departure`	55	*Sag: Ich möchte um 9 Uhr morgen abreisen* (*Say: I want to leave tomorrow at 9am*), *Sag: Ich möchte am Sonntagabend gehen* (*Say: I would like to go Sunday night*), ...
`buying_clothes`	51	*Frag: blaue Sandalen* (*Ask for: blue sandals*), *Frag: braune Stiefel* (*Ask for: brown boots*), ...
`ordering_food`	48	*Frag: Apfelkuchen* (*Ask for: the apple pie*), *Frag: ein Wasser ohne Kohlensäure* (*Ask for: a glass of still water*), ...
`directions`	24	*Frag: Wo ist der Coiffeur?* (*Ask: Where is the hairdresser?*), *Frag: Wo ist die Tate Modern?* (*Ask: Where is the Tate Modern?*), ...

Table 2: *Five most frequent prompt categories*

the prompt category labels. It should be pointed out that this definition of "prompt category" was based strictly on linguistic (vocabulary and grammar) similarities between expected responses across prompts and is not intended to correspond to the lesson groups that were defined for the spoken CALL system that the task data were drawn from [15].

2.3. Similarity to Reference

We also investigated several text-to-text similarity features to measure the similarity between an utterance to the sample responses for each prompt provided by the task organizers in the file `referenceGrammar.xml`. These features are motivated by the fact that utterances that are not contained in the grammar but that are similar to the sample responses are more likely to be correct. This set of features is intended to address two potential sources of false rejects by the system, namely correct responses that are not listed in the grammar and correct responses that are in the grammar but that are misrecognized by the ASR system, thus causing the input to the scoring module to be out-of-grammar. The following four text-to-text similarity features were explored in this study:

- Minimum Word Error Rate (WER) between the utterance and each reference response for the prompt
- Maximum WER between the utterance and each reference response for the prompt
- Mean WER across all reference responses for the prompt
- BLEU score between the utterance and the reference responses for the prompt

The WER is obtained by first calculating the edit distance, i.e., the minimum distance obtained by applying dynamic programming to align two sequences of words with possible insertions, deletions, and substitutions [16], between the speaker's utterance and one of the responses in the reference grammar and then dividing the distance by the length of the sample response. We calculated the WER for each utterance/sample-response pair for a corresponding prompt and then calculated the minimum, mean, and maximum WER values across all of the sample responses for a prompt and used these three values as features in the model.

We calculated an additional feature based on the similarity between the test response and the reference grammar by obtaining the BLEU score between the utterance and the set of sample responses for the prompt. BLEU uses a modified precision metric and is typically used to evaluate the quality of a machine translation against several human reference translations [17].

2.4. Grammar

We used the *language-check* Python wrapper[2] for *LanguageTool*, an open-source English proofreading package, to check whether an utterance contains any grammatical errors based on predefined rules; the number of errors detected in each response was then used as a feature. For example, one of the system's rules (Phrase Repetition Rule) looks for duplicated phrases; an example response from the training set that would be flagged by this rule is *i would like a* <u>ticket to ticket to</u> *pay with dollars* (ID #5974). Another example rule (Baseform Rule) checks whether the base form of a verb is used after certain modal verbs; an example response that would be flagged by this rule is *i will* <u>*boots*</u> (ID #5958).[3]

Several rules in *LanguageTool* were not appropriate for use with ASR output (for instance, a rule that checks whether a sentence ends with a period) or for spoken responses in general (for example, a rule that checks whether an input is a sentence fragment). A total of 5 rules were excluded for these reasons from the feature calculation pipeline, leaving a total of 200 grammar rules; of these, 39 rules were matched by one or more responses in the training set. Overall, only 3.8% (200 out of 5,222) of the utterances in the training set were associated with one or more grammar errors for this feature.

3. Experiments on Training Set

This section describes experiments that were conducted on the training set in order to determine the relative impact of the different types of features on the model's performance as well as to investigate the influence of the two different sources of input to the model: transcriptions and ASR output. Since the WER of the Kaldi ASR output provided by the task organizers (0.147) was substantially lower than the WER of the Nuance ASR output (0.319), the Kaldi ASR output was used for all of these experiments. The models described in this section (as well as the models that were used to produce results on the test set) use the machine learning approach that is typically used for the automated content scoring system, namely support vector regression with hyperparameter grid search optimization. In order to train models with different combinations of features from the different feature sets described in Section 2, the features from

[2] `https://pypi.python.org/pypi/language-check`

[3] Note that the speaker intended *will* to be the main verb of the utterance (as in *want*) and *boots* to be a plural noun, not a 3rd-person singular verb. Since *LanguageTool*'s rules are based on pattern matching and do not make use of POS tagging, system errors of this nature are to be expected. Additionally, it should be pointed out that, even though the system labeled the error incorrectly, the utterance does, in fact, contain an error (i.e., use of the incorrect verb), and would therefore be correctly rejected by the application of this rule.

the different sets were combined into a single feature vector for each response which was then used as input to learn the support vector regression parameters for each model.

3.1. Feature Evaluation

We conducted a feature ablation experiment by training several models on the training set to determine the relative contributions of the different types of features for predicting erroneous utterances when added to the features from the content scoring system (this experiment was conducted after the test results were submitted for the shared task, so its findings could not be incorporated into the official submission). We trained a series of models that all included the base set of Content features and added one additional feature type at a time (we also trained a model that used only the base Content features). This resulted in a total of eight models for this experiment: one for each of the three WER features, one for the BLEU feature, one for the grammar feature, one for each of the two bias features, and one for the base Content feature set by itself. The results, in terms of D score, which was the chosen evaluation metric for the Spoken CALL Shared Task and is defined as the ratio of the relative correct reject rate to the relative false reject rate [8][4], obtained with these models (using the output of the Kaldi ASR system and 10-fold cross-validation on the training set) are reported in Table 3.

Features	D
Content + Prompt Category Bias	9.114
Content + BLEU	9.281
Content + Prompt Bias	9.556
Content	9.771
Content + Grammar	9.782
Content + Max. WER	9.944
Content + Mean WER	10.366
Content + Min. WER	11.498

Table 3: *Results obtained using the Kaldi ASR output and 10-fold cross-validation on the training set for eight different models based adding individual features one by one to the Content features*

After this experiment was completed, a second round of feature ablation was conducted in which each feature was added into the base Content feature set starting with the best-perfoming features until eventually all features were included; the results of this experiment are presented in Table 4. As the table shows, the model based on the Content features with the addition of the Minimum WER and Mean WER features performed best with a D score of 11.504; the performance of the model then declined with the addition of subsequent features. In addition to the D score, additional performance metrics were calculated on the best performing system on the training data set in order to more fully understand its strengths and weaknesses; these are as follows: 3,719 correct accept (71.2%), 242 plain false accept (4.6%), 235 gross false accept (4.5%), 865 correct reject (16.6%), 161 false reject (3.1%), 87.8% accuracy, 88.6% precision, and 95.9% recall.

Features	D
Content	9.771
+ Min. WER	11.498
+ Mean WER	**11.504**
+ Max. WER	11.462
+ Grammar	11.397
+ Prompt Bias	10.850
+ BLEU	9.232
+ Prompt Category Bias	10.458

Table 4: *Results obtained using the Kaldi ASR output and 10-fold cross-validation on the training set for eight different models based on the step-wise addition of each feature set to the Content features*

3.2. Quantity and Quality of Training Data

Since the training data set in the Spoken CALL shared task contains both transcriptions and ASR output for the spoken responses, we experimented with different configurations of these two sources of input to the model. Specifically, we conducted 10-fold cross-validation experiments on the training data set using only the Content features to determine what type of input to the model is most effective for scoring the ASR output; the results of these experiments are presented in Table 5.[5]

Training Data	Testing Data	D
transcriptions	Kaldi ASR	5.727
Kaldi ASR	Kaldi ASR	8.838
Kaldi ASR + transcriptions	Kaldi ASR	11.126
transcriptions	transcriptions	52.424

Table 5: *10-fold cross-validation results using different configurations of training and testing data for models consisting of the Content features*

As the table shows, a model trained on the transcriptions performs worse when evaluated on the ASR output than a model trained on the ASR output. This is expected, due to the mismatch between the characteristics of the vocabulary and grammar contained in the responses in the training and testing sets. However, Table 5 also shows that the performance improves when the ASR output and the transcriptions are combined together in the training set (this was done by including separate entries in the model training set for the ASR output and transcription, thus doubling the size of the training set from 5,222 to 10,444). This seems to suggest that, despite the mismatch, the addition of the transcriptions to the model can provide some additional information that is helpful for the model's prediction. However, it could also be the case that the combined model simply performs better because it is based on twice the amount of training data.

[4] For the Spoken CALL Shared Task, a weighting factor was applied in the calculation of the D score to penalize *gross false accepts* (where the system accepts a response that was annotated as incorrect in meaning) three times as heavily as *plain false accepts* (where the system accepts a response that was annotated as incorrect in grammar but correct in meaning)

[5] Note that the value of 8.838 reported for the "Kaldi ASR" training/"Kaldi ASR" test row in Table 5 differs from the value reported in Table 4 when using the base Content features on the Kaldi ASR due to the use of an earlier version of the data provided by the organizers and a different ordering of the data (with respect to the cross-validation folds).

To investigate which of these explanations is correct, we first conducted an experiment in which we randomly selected five different halves of the combined training data set to make it equal in size to the original training set, i.e., 5,222 responses consisting of 2,611 responses represented by their ASR output and 2,611 responses represented by their transcription. Then, five different cross-validation experiments with shared folds were conducted (regardless of whether a response was randomly assigned to the ASR or transcription category for a given experiment, the ASR output was always used for the responses in the testing set for each fold). The average D score across these five cross-validation experiments was 8.4 (with a range from 8.1 to 8.9), which is similar to the performance of the model trained using the ASR output on the training set consisting of 5,222 responses. This result likely indicates that the performance improvement observed by combining the ASR and transcriptions in the training set is due primarily to the increased size of the training set, not because the transcriptions themselves are beneficial when testing on ASR output.

Finally, Table 5 shows that a model trained on the transcriptions performs quite well when tested on the transcriptions (D = 52.4); this result represents an upper bound on performance that can be expected if a spoken CALL system had completely accurate ASR.

4. Test Results

Based on the results presented in Section 3.2 about the differences between system performance when the models were trained on ASR output or a combination of ASR output and transcriptions, we decided that one of the official test submissions should be based on ASR output only and another should be based on a combination of ASR output and transcriptions; all of the features described in Section 2 were included in these two submissions. For the third submission, we used models trained only on ASR output with the subset of features consisting of a combination of the Content features and bias features (including both the prompt bias and prompt category bias features), since initial results we obtained suggested that the bias feature set performed well in comparison to the other additional feature sets (these initial results were obtained prior to the controlled feature ablation study that was presented in Section 3.1). Table 6 presents the official test results in terms of the D score for these three submissions (labeled MMM, NNN, and OOO by the task organizers); for all of these systems, the models were trained using all responses from the training set and tested on the Kaldi ASR output provided by the task organizers.

Submission	Features	Model Responses	D
MMM	All	ASR output	4.353
OOO	All	ASR output and transcriptions	4.273
NNN	Content and bias	ASR output	3.188
		Baseline Kaldi	1.694

Table 6: *Evaluation results for three official submissions on the test set*

As Table 6 shows, the system with all features substantially outperformed the system trained using the subset of Content and bias features; furthermore, the system trained on ASR output only slightly outperformed the system trained on a combination of ASR output and transcriptions. All three submissions outperform the Kaldi baseline provided by the shared task organizers which checks to see whether the utterance is contained in the reference grammar.

After submitting the official results for the shared task, controlled experiments were conducted to determine the relative contribution of each of the additional features to the Content features on the test set, similar to the experiments conducted on the training set. As shown in Section 3.1, a system trained using the Content features plus two features based on similarity to the sample responses in the reference grammar (Minimum WER and Mean WER) performed the best on the training set. Similar evaluations were also conducted on the test set to determine whether a different combination of features could improve on the official results shown in Table 6; these results are shown in Table 7.

Features	D
Content + Prompt Category Bias	2.801
Content + BLEU	4.214
Content + Prompt Bias	3.248
Content	4.207
Content + Grammar	3.674
Content + Max. WER	4.525
Content + Mean WER	4.358
Content + Min. WER	4.204

Table 7: *Results obtained using the Kaldi ASR output on the test set for eight different models based adding individual features one by one to the Content features*

As shown in Table 7, the features based on comparing the test response to the sample responses in the reference grammar (WER and BLEU features) outperform the grammar and bias features; this is similar to the results shown for the training data in Table 3, except that the BLEU features are more effective on the test set. Finally, we also trained models with feature sets added one-by-one in the same order as was done for the training set in Table 4; these results are presented in Table 8.

Features	D
Content	4.207
+ Min. WER	4.204
+ Mean WER	4.278
+ Max. WER	4.395
+ Grammar	4.119
+ Prompt Bias	4.066
+ BLEU	**4.565**
+ Prompt Category Bias	4.353

Table 8: *Results obtained using the Kaldi ASR output on the test set for eight different models based on the step-wise addition of each feature set to the Content features*

As shown in Table 8, the best performance on the test set was produced by a model that included all of the features except for the Prompt Category Bias feature, with a D score of 4.565. Additional evaluation metrics for this model are as follows: 649 correct accept (65.2%), 51 plain false accept (5.1%), 59 gross false accept (5.9%), 170 correct reject (17.1%), 67 false reject

(6.7%), 82.2% accuracy, 85.5% precision, and 90.6% recall.

5. Discussion and Conclusion

In this paper we described the system that was used by ETS to participate in the Spoken CALL Shared Task at SLaTE 2017. The system is based on a pre-existing automated content scoring system that consists primarily of character and *n*-gram features. This system substantially outperforms the baseline on both the training and the test data sets, thus indicating the usefulness of this system for the task of accepting / rejecting responses in a spoken CALL application.

In addition, we explored the use of three additional types of features (bias, text-to-text similarity to sample responses, and grammar) in combination with the features from the content scoring system. Since these experiments showed that the WER-based features that compared the spoken response to the sample responses in the reference grammar were most effective for this task, future research could benefit from investigating additional matching approaches that could potentially be more robust to ASR errors than WER. [18] explores several of these, such as regular expressions, BLEU, and LM score, in the context of scoring the content of short responses provided in the context of an assessment of English speaking proficiency for teachers of English.

The difference in performance between the cross-validation results on the training set reported in Section 3 and the results on the test set reported in Section 4 is striking. Further research will be necessary to investigate why the performance is so much lower on the test set. As mentioned in Section 2.2, some of the responses in the test set were drawn from prompts that were not seen in the training set, and it is possible that the model did not generalize well to these responses to unseen prompts. However, these responses only represent a small percentage of the test set, and therefore cannot be the sole explanation for the substantial performance difference. Further error analysis should be conducted to determine whether there are other characteristics of the test set that differ from the training set that could explain the observed difference. As shown in Sections 3 and 4, the accuracy of the best system on the test set was 82.2% compared to an accuracy of 87.8% for the best system on the training set; this difference of 5.6% in accuracy corresponded to a difference of 6.939 in D score. Since the proportion of false rejects on the test set (6.7%) is much higher than on the training set (3.1%), it appears that the D score may be more sensitive to this type of error than the standard performance evaluation metrics.

6. References

[1] R. Lyster, "Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classrooms," *Language Learning*, vol. 48, no. 2, pp. 183–218, 1998.

[2] F. A. Morris, "Negotiation moves and recasts in relation to error types and learner repair in the foreign language classroom," *Foreign Language Annals*, vol. 35, no. 4, pp. 395–404, 2002.

[3] E. Kartchava and A. Ammar, "The noticeability and effectiveness of corrective feedback in relation to target type," *Language Teaching Research*, vol. 18, no. 4, pp. 428–452, 2014.

[4] D. Brown, "The type and linguistic foci of oral corrective feedback in the L2 classroom," *Lantuage Teaching Research*, vol. 20, no. 4, pp. 436–458, 2016.

[5] H. Morton and M. A. Jack, "Scenario-based spoken interaction with virtual agents," *Computer Assisted Language Learning*, vol. 18, no. 3, pp. 171–191, 2005.

[6] W. Johnson and A. Valente, "Tactical language and culture training systems: using AI to teach foreign languages and cultures," *AI Magazine*, vol. 30, no. 2, pp. 72–83, 2009.

[7] K. Lee, S.-O. Kweon, S. Lee, H. Noh, and G. G. Lee, "POSTECH immersive English study (POMY): Dialog-based language learning game," *IEICE Transactions on Information and Systems*, vol. 97, no. 7, pp. 1830–1841, 2014.

[8] C. Baur, J. Gerlach, M. Rayner, M. Russell, and H. Strik, "A shared task for spoken CALL?" in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

[9] M. Heilman and N. Madnani, "The impact of training data on automated short answer scoring performance," in *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications*, 2015, pp. 81–85.

[10] O. L. Liu, J. A. Rios, M. Heilman, and M. C. Linn, "Validation of automated scoring of science assessments," *Journal of Research in Science Teaching*, vol. 53, pp. 215–233, 2016.

[11] N. Madnani, A. Cahill, and B. Riordan, "Automatically scoring tests of proficiency in music instruction," in *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 2016, pp. 217–222.

[12] A. Loukina and A. Cahill, "Automated scoring across different modalities," in *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, San Diego, California*, 2016, pp. 130–135.

[13] Y. Zhang and S. Clark, "Syntactic processing using the generalized perceptron and beam search," *Computational Linguistics*, vol. 37, no. 1, pp. 105–151, 2011.

[14] M. Heilman and N. Madnani, "ETS: Domain adaptation and stacking for short answer scoring," in *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 2013, pp. 275–279.

[15] C. Baur, "The potential of interactive speech-enabled CALL in the Swiss education system: A large-scale experiment on the basis of English CALL-SLT," Ph.D. dissertation, University of Geneva, 2015.

[16] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[17] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[18] K. Zechner and X. Wang, "Automated content scoring of spoken responses in an assessment for teachers of English," in *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, 2013, pp. 73–81.