International Journal of Intelligent Technologies and Applied Statistics
Vol.13, No.3 (2020) pp.229–256, DOI:10.6148/IJITAS.202009_13(3).0003
© Ainosco Press

Measuring the Expressive Language and Vocabulary of Latino English Learners Using Hand Transcribed Speech Data and Automated Scoring

Makoto Sano[1*], Doris Luft Baker[2], Marlen Collazo[3], Nancy Le[4] and Akihito Kamata[4]

[1]*Mack-the-Psych.com, Tokyo, Japan*
[2]*Department of Special Education, University of Texas at Austin, Austin, USA*
[3]*Department of World Languages and Literatures, Southern Methodist University, Dallas, USA*
[4]*Center on Research and Evaluation, Southern Methodist University, Dallas, USA*

Abstract

Purpose: Explore how different automated scoring (AS) models score reliably the expressive language and vocabulary knowledge in depth of young second grade Latino English learners.
Design/methodology/approach: Analyze a total of 13,471 English utterances from 217 Latino English learners with random forest, end-to-end memory networks, long short-term memory, and other AS models.
Findings: Random forest outperformed the other AS models as measured by the mean of quadratic weighted kappa (QWK = 0.70) followed by the end-to-end memory networks–long short-term memory (QWK = 0.69) across all tasks and data points. The QWK between humans was 0.90, while the human-machine agreement of three AS models and humans ranged from 0.66 to 0.70.
Practical implications: Examine closely misclassifications between human and machine scoring to better understand the specific words and structures the systems were not capturing.
Originality/value: Discuss findings in the context of developing efficient and reliable ways to analyze the natural speech of young English learners. This information could guide the vocabulary and language proficiency instruction in the early grades.

Keywords: Automated scoring; Human-machine reliability; Deep neural network models; Linguistic classification; Natural language processing; Vocabulary teaching and learning

* Corresponding author: makoto.sano@mack-the-psych.com

1. Introduction

With the increasing use of technology in school settings, the collection of natural language data provides a unique opportunity for researchers and teachers to understand better how students process information and use it to express their thoughts. Analyzing the speech of English learners (ELs) is particularly important given the process they have to go through to understand content and at the same time develop their academic vocabulary and language proficiency in English, their second language (L2). However, scoring transcribed student speech is time-consuming and not always reliable. Thus, the purpose of the current study is to explore the reliability of seven different automated scoring (AS) models and compare them with each other and with human scoring (i.e., scoring conducted by a research assistant following a specific rubric). We anticipate that outcomes from this study can inform future development of more precise AS models, particularly for ELs. ELs is a term used in the U.S. to refer to students who are learning English as a second language (ESL), and who require additional supports in order to understand and discuss content in classrooms where instruction is only provided in English (August & Shanahan, 2006).

1.1 Theoretical and empirical evidence to measure and score vocabulary knowledge

Vocabulary knowledge is a complex construct that cannot be understood solely in terms of the number of words known (Christ, 2011; Schoonen & Verhallen, 2008). Henriksen (1999) describes the process of learning words as network building while Perfetti (2007) refers to it as the lexical quality hypothesis in which the learner's level of knowledge of a word is connected to other words and ultimately to reading comprehension (i.e., as the learner gains more experience with a word, more grammatical classes and inflections are learned, and the meaning becomes incrementally more precise and less bound to context). According to Perfetti, high-quality representations or semantic networks in which elements of form and meaning are tightly connected to one another can be retrieved quickly while low-quality representation reduces the retrieval speed and the learner's ability to comprehend a passage.

In the case of ELs, vocabulary presents an additional challenge given that they need to understand (1) basic words in everyday life such as *rope*, *stairs*, *walk*; (2) academic vocabulary used in multiple different texts such as *admire*, *polite*, *survive*; and (3) content knowledge words such as *seasons*, *habitat*, *hibernation* (D. L. Baker, Basaraba, & Richards-Tutor, 2018; Hiebert, 2006). Therefore, vocabulary instruction has to occur at all different levels across grades (Gersten, Baker, & Lloyd, 2000; Shanahan & Beck, 2006), and we cannot wait until students have acquired basic word knowledge to introduce them to abstract words because of time constraints (Cena et al., 2013).

However, converging research indicates that ELs can perform as well as their English-only peers (EOs) on word-level skills such as decoding, word recognition, and spelling (S. K. Baker & Baker, 2008). Nonetheless, substantial evidence exists that ELs do not attain the same levels of performance on text-level skills such as reading comprehension and writing mainly due to their low vocabulary knowledge, listening comprehension, and syntactic skills (Geva & Farnia, 2012; Kieffer, 2010). The reason for the low comprehension could be explained in part because, cognitively, a lack of vocabulary reduces the mental processes students need to make sense of words, reducing their depth and breadth of understanding of content (Anderson & Freebody, 1981; Elleman, Lindo, Morphy, & Compton, 2009).

The evidence suggesting that ELs perform lower than non-ELs in vocabulary and reading comprehension assessments is reflected in the low academic performance of ELs on the National Assessment of Educational Progress (NAEP) (Kena et al., 2015). The data from these assessments indicate that ELs scored significantly below EOs on the fourth- and eighth-grade assessments of reading (by 38 and 45 points, respectively) and mathematics (by 25 and 41 points, respectively). This trend has not changed significantly for the last four years (U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress, 2019).

Potential malleable factors that can explain the low vocabulary knowledge of ELs, can be attributed to three main reasons. First, the increase in the vocabulary and language demands of the Common Core State Standards (CCSS) (Common Core State Standards Initiative, 2010), the Texas Essential Knowledge and Skills (TEKS) (Texas Education Agency, 2010) standards, and the Next Generation Science Standards (NGSS) (NGSS Lead States, 2013) require students to use more precise vocabulary across the content areas. Second, despite the demands from the CCSS, the TEKS, and the NGSS standards, teachers, in general, tend to spend less time teaching vocabulary during their literacy block or throughout the day than teaching comprehension or decoding (D. L. Baker & Kosty, 2012; Wanzek, 2014). Third, vocabulary tends to be taught whole group reducing the opportunities for ELs with low vocabulary to practice their vocabulary more frequently in small groups or using technology (D. L. Baker et al., 2018).

Therefore, developing an AS system that can provide teachers with timely information about student word knowledge would help teachers (1) make informed decisions on how to identify students with low vocabulary; (2) differentiate their vocabulary instruction within a Multi-Tier System of Support (MTSS); and (3) potentially increase the amount of time teachers spend on teaching vocabulary because they will be more aware of their student vocabulary needs.

1.2 Review of studies on AS models

Currently, there are several AS models that have been developed to automatically score spontaneous speech such as SpeechRater, E-Rater, and others. Each of these systems uses a slightly different model to score student essay production or student speech. Table 1 includes a summary of all the models. For example, Somasundaran and Chodorow (2014) developed a system to automatically score a vocabulary item task that required examinees (mainly non-native English speakers) to use two words in writing a sentence to describe a given picture. An operational scoring guide ranging from 0 to 3 points and statistical modeling (i.e., to determine the consistency of a response with respect to the picture) were used to evaluate 58,000 student responses both manually and by means of their AS model that used random forest (RF) to score student responses. Results indicated that the ratio of the number of responses correctly classified over the total number of responses by their AS system was 15 percentage points higher than the baseline that simply classifies all the responses into the majority class (76.23% vs. 61.00%). However, this system was still 10 percentage points lower than human scoring (86.00%). As a result, the quadratic weighted kappa (QWK) of human-machine agreement was 0.63, while the QWK of human-human agreement was 0.83.

Table 1. Summary of AS models.

Model	Acronym	Characteristic	Use case in AS
Bayesian linear ridge regression	BLRR	Mainly used for supervised domain adaptation where a small number of labeled target data and relatively large number of labeled source data is available in the context of automated essay scoring (AES). The AES task is modeled as a regression problem and used linear regression within the context of Bayesian inference.	Phandi, Chai, & Ng (2015)
Convolutional neural network	CNN	One of the deep neural network models typically used to recognize objects in images. The network performs a series of convolutions and pooling operations for the feature detections.	Taghipour & Ng (2016)
End-to-end memory networks	MemN2N	One of the deep neural network models with attention mechanism over a possible large external memory and trained end-to-end. It requires significantly less supervision during training than the original Memory Networks (Weston, Chopra, & Borders, 2015).	Zhao, Zhang, Xiong, Botelho, & Heffernan (2017)

Table 1. Summary of AS models. (continued)

Model	Acronym	Characteristic	Use case in AS
k-nearest neighbor	*k*-NN	One of the regression technics where unscored response is predicted by the average of the scores of the *k* training essays nearest to the unscored response in the context of AES.	J. Chen, Fife, Bejar, & Rupp (2016)
Long short-term memory	LSTM	An extension for recurrent neural networks (RNN) which extends the memory to remember their inputs over a long period of time and capable of learning long-term dependencies.	Alikaniotis, Yannakoudakis, & Rei (2016); Taghipour & Ng (2016)
Multiple linear regression	MLR	The most common form of linear regression widely used for operational AES, for example, in E-Rater and SpeechRater by Educational Testing Service (ETS).	J. Chen et al. (2016); L. Chen et al. (2018); Yoon, Bhat, & Zechner (2012); Yoon et al. (2018)
Random forests	RF	A classifier uses a number of decision trees to predict the value of a dependent variable based on the independent variables. Each decision tree will make a prediction of the dependent variable through its training and the final prediction is conducted based on the votes across all the trees.	J. Chen et al. (2016); Somasundaran & Chodorow (2014)
Recurrent neural network	RNN	One of the deep neural network models which has internal memory and typically used for sequential data like time series, speech, and text.	Taghipour & Ng (2016)

Table 1. Summary of AS models. (continued)

Model	Acronym	Characteristic	Use case in AS
Support vector machine (support vector classification)	SVM	In the context of AS, the algorithm uses decision surfaces in the space to separate the responses optimally into different score categories as a classification task. The decision surfaces are chosen to maximize the average distance (margin) between the decision surface and the responses belonging to different score categories.	J. Chen et al. (2016); Sano, Baker, & Kamata (2018); Zechner & Bejar (2006)
Support vector regression	SVR	One of the regression technics based on SVM with maintaining all the main features that characterize the algorithm to maximize the margin. In this study, the algorithm used for regression task with support vector is identified as SVR while the algorithm for classification task is called SVM unless otherwise stated.	Alikaniotis et al. (2016)

On the other hand, J. Chen et al. (2016) evaluated the E-Rater essay scoring model that uses MLR to alternative scoring models, such as SVM, RF, and *k*-NN. Using data from four different writing tasks[1] in two large-scale college level assessments, J. Chen et al. assessed the performance of each E-Rater model using four statistics: (1) QWK, (2) percentage of exact agreement with human scores, (3) standardized mean difference (SMD) between human and rounded E-Rater scores, and (4) Pearson correlation between human and unbounded/raw E-Rater scores (theoretically range from -∞ to +∞ without truncation to be matched with the score scale of human ratings). Results revealed that the SVM model achieved the best results based on most of the evaluation metrics across four writing tasks. SVM-based scores and human scores were also related to examinees' scores on other sections of the test, thus providing increased validity for the SVM-based E-Rater scores. These

[1] Task A required examinees to critique an argument. Task B required examinees to articulate an opinion and support their opinions by using examples or relevant reasoning. Task C required test takers to read, listen, and then respond in writing by synthesizing the information that they had read with the information they had heard. Task D required test takers to articulate and support an opinion on a topic.

findings indicate that more complex models than MLR need to be developed to improve the performance of the E-Rater in scoring the quality of essay writing.

We found three other studies that examined AS models to measure the language proficiency of ELs (Crossley & McNamara, 2013; Lu, 2012; Yoon et al., 2012). Crossley and McNamara (2013) analyzed 244 transcribed spontaneous speech samples taken from the Test of English as a Foreign Language (TOEFL) Internet-Based Test (iBT) produced by 244 ESL learners. Using automated indices from three different tools (i.e., Coh-Metrix, the Computerized Propositional Idea Density Rater [CPIDR], and Linguistic Inquiry and Word Count [LIWC]), the authors focused on higher-level linguistics features such as speech delivery, language use, topic development, and their relation to attaining communicative competence. Moreover, Crossley and McNamara examined which linguistic features in speech were the most predictive of speaking proficiency.

While study results proved promising, the TOEFL implements a very limited rubric from which it defines speech proficiency (i.e., 4-point scale). Therefore, results may not be generalizable to speech produced by ELs that is more intricate and harder to assess within such a constricted scale. In that study no information about participants was included, nor the setting where the assessments took place, which could have had an effect on outcomes.

Yoon et al. (2012) measured 480 ELs vocabulary use in English from 2,880 spontaneous speech responses collected from the Academic English Screening Test (AEST). Lexical sophistication was measured by means of a vocabulary profile (VP) approach. This approach identified three classes of features: coverage-related, average word rank, and average word frequency to quantify the English usage of non-English speakers. Researchers concluded that these features had high correlations with human proficiency scores with average word frequency achieving the best correlation among the features. However, when using an AS model with other predictors of language proficiency, these features only showed marginal improvement in predicting human proficiency scores. These findings contribute to a limited existing body of literature on vocabulary usage essential to determining the language competence of ELs.

Recently, Yoon et al. (2018) conducted two experiments to develop an AS model that provides a holistic proficiency score using audio files and their transcriptions. The authors developed two sets of content features: (1) based on traditional content vector analysis, and (2) features based on word embedding. The first experiment included 8,700 ESL speakers. For this experiment, they randomly selected 438 questions (for inverse document frequency, idf)[2] that did not overlap with the 147 questions used for the scoring model training and evaluation set (SMTES). Researchers used the TOEFL iBT Speaking Test Rubrics to grade the responses, and

[2] The idf of each word which is obtained by calculating the total number of responses divided by the number of responses containing the word.

the two sets of content features developed by them. The initial correlation analyses between features and human scores were conducted using the SMTES. Findings indicated that performance of the content features based on word embedding was better than content vector analysis, and the best performing content feature was idf weighted word embedding.

Other researchers have used RNNs to examine the association between an essay and its assigned score without feature engineering (Taghipour & Ng, 2016). In addition, Taghipour and Ng (2016) also explored several machine learning models such as LSTM (Hochreiter & Schmidhuber, 1997) and CNN to determine how these models perform compared to other models. The Kaggle dataset from the Automated Student Assessment Prize (ASAP) competition[3] were analyzed, and QWK was used to evaluate the models. Findings indicated that LSTM performed significantly better than all other systems and outperformed baseline models such as SVR and BLRR (Phandi et al., 2015) by a large margin (i.e., 0.041 points in QWK) demonstrating statistically significant improvements ($p < 0.05$). Yet, the best model was one that combined CNN and LSTM, which outperformed the baseline (i.e., BLRR) by 0.056 points in QWK. In other words, results indicated that a system based on LSTM networks can achieve state-of-the-art performance in AES without requiring any feature engineering as it automatically learned the representations required for the task by extracting the necessary information from the scored essays.

In another study, Alikaniotis et al. (2016) introduced LSTM as a bi-directional, deep, neural network that can represent both contextual and usage information by learning the extent to which specific words contribute to a score of student answers (i.e., Score-Specific Word Embeddings, SSWEs). The sample for the analysis was also extracted from the Kaggle dataset that includes 12,000 essays written by students in Grades 7 through 10 consisting of eight distinct sets produced by eight different prompts each with individual marking criteria and score range. Findings suggested that the SSWEs + LSTM model scored the essays in the most human-like way, outperforming SVR. Moreover, this model did not require the inclusion of prior knowledge about the language in the test such as grammar. The model applying SSWEs also has the advantage of potentially reflecting correct/incorrect spelling to the essay scoring.

Another model in the use of neural networks is MemN2N (Sukhbaatar, Szlam, Weston, & Fergus, 2015; Zhao et al., 2017). MemN2N is a neural network with attention mechanism over a possible large external memory and trained end-to-end network. It requires significantly less supervision during training than the original Memory Networks (Weston et al., 2015). MemN2N has been primarily applied to question and answer tasks where a set of statements are stored in memory after performing word embedding. The query is also embedded to compute the match between the query and each memory stored. Zhao et al.'s (2017) model is an extended

[3] For more information on ASAP, see the website https://www.kaggle.com/c/asap-aes/.

type of MemN2N. Instead of using one input layer with word embedding, they applied separated input layers for each score class (e.g., score 0 to 3). The proposed model outperformed existing models including LSTM and LSTM + CNN for the data set from the ASAP competition.

Our study applies one input layer version of MemN2N to incorporate the match between a student response and correct answer examples for the scoring since the number of responses for the highest score class in sentence task is very limited. When we applied MemN2N to convert the output into a scoring result, LSTM was imposed (hereinafter we call this model MemN2N–LSTM) before passing the concatenated memory o and the input embedding u to the final weight matrix W (see Figure 1). It is suggested by Weston et al. (2015) that the response (i.e., scoring) component could be RNN (or LSTM) conditioned on the output of o and u as illustrated in Figure 1.

In summary, although currently reliable AS models exist, most of them tend to be used in the upper grades and college to evaluate the language proficiency or essays of students in middle school to college. In addition, the systems do not focus on students learning vocabulary specifically, and few studies compared more than

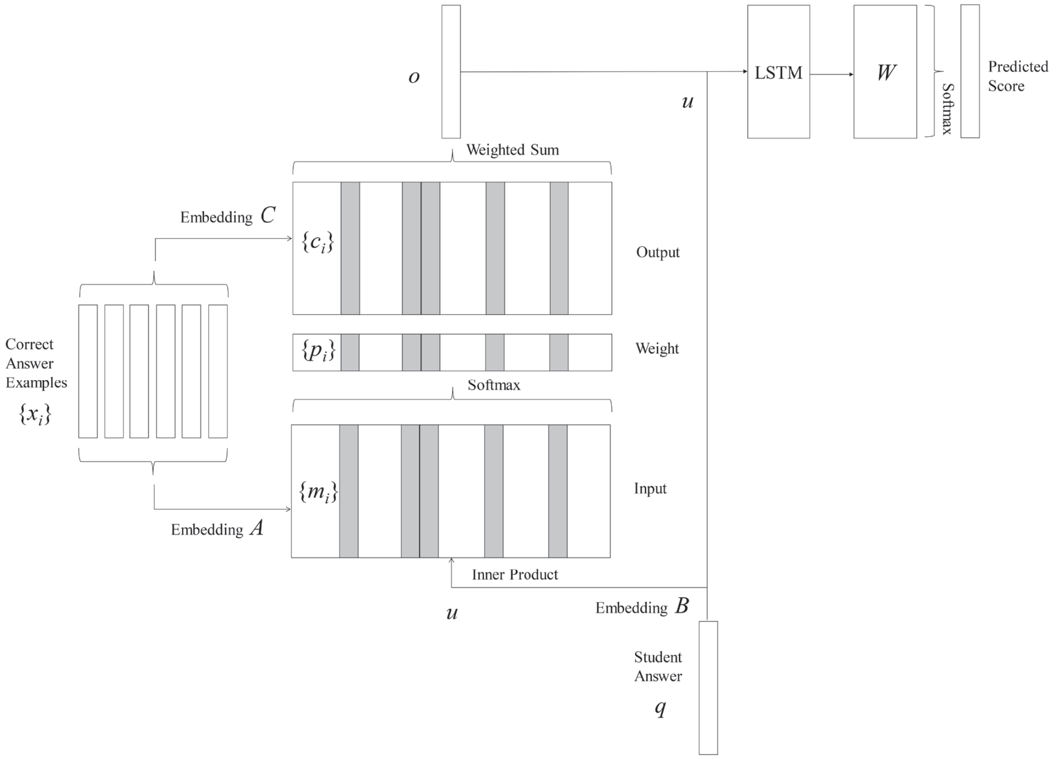

Figure 1. MemN2N–LSTM path.

two or three models using a word embedding approach. Thus, this exploratory study extends current research on AS models by (1) comparing the accuracy of seven different scoring models using speech data of ELs in second grade, and (2) comparing the scoring between humans and AS models. Specifically, in this study we will attempt to answer the following three research questions:

(1) Which AS model appears to have the highest human-machine agreement as measured by QWK?

(2) What is the inter-rater agreement between human scorers, and between human scorers and AS models?

(3) Which student responses appeared to have the maximum extend of disagreement between humans and AS models?

2. Method

The data collected for this study came from a larger vocabulary project designed to develop the in-depth vocabulary knowledge in science and social studies of second grade 486 Latino students using an intelligent tutoring system. As part of the project, we transcribed and analyzed student responses to questions related to their knowledge of academic words that appear in common science and social studies topics in second grade curricula and are suggested to be necessary to understand content in the CCSS and the NGSS.

Participants were 217 Latino second grade students (i.e., 116 girls, and 200 students eligible for free and reduced lunch, a measure of poverty). In addition, 75 of the 217 students attended two English-only charter schools, and 19 attended a charter school with a two-way trilingual (English, Spanish, and Chinese) program where Spanish-speaking and English-speaking students were taught half the day in English and half the day in Spanish with Chinese instruction for 30 minutes every day. Out of the 217 students, 109 attended four public schools that had a one-way bilingual program. In this program, every student in the class speaks Spanish as their native language and they receive, in general, mathematics and social science instruction in English, and reading, writing, and science instruction in Spanish.

A total of 13,471 English utterances from participants related to academic vocabulary knowledge at pre-test and post-test were jointly analyzed. Students were asked to define a target word and use the prompt word in a sentence. Thus, each question evoked two answers by each student as the products of definition and sentence comprehension tasks. The total number of student responses were 4,169 at pre-test definition task and 4,137 at pre-test sentence task for 23 words, and 2,609 at post-test definition task and 2,556 at post-test sentence task for 15 words.

2.1 Vocabulary measure

To compare the accuracy of different AS models and human scoring, we used a vocabulary measure, depth of knowledge (DOK), which has been validated previously in several studies (see D. L. Baker et al., 2020; S. K. Baker et al., 2013). The measure is individually administered and examiners ask students to define a word and use it in a sentence. Students receive a score for a definition (0–2) and a score for using the word in a sentence (0–3) following the rubric below.

2.1.1 Definition comprehension task

To score the DOK vocabulary measure we scored a definition as 0 when there was an incorrect use of the word or use of the word without details or context, or the student used the word itself to define the word. A score of 1 was given when the definition was in general correct with 2 or more grammatically correct-errors, contextually missing information or no details were provided. A score of 2 was given when students used the word or root in a generally grammatically correct sentence with some context or at least one detail provided.

2.1.2 Sentence comprehension task

Students received a 0 in the use of the word in a sentence when there was no response (NR), a response was incomprehensible, the sentence did not include the target word, or the response was very basic. Students received a score of 1 when the word or root was generally correct, with 2 or more errors or there was context or details missing. A score of 2 represented a sentence that was generally grammatically correct with some content, or detail provided. A score of 3 represented a sentence that used the word in a grammatically correct complex sentence in an academic context with at least two or more clauses.

2.2 Validity and reliability of the DOK

S. K. Baker et al. (2013) used the paper-and-pencil version of the DOK measure to assess the effects of a read aloud intervention in first grade on student oral reading fluency, vocabulary, and reading comprehension. Correlations between DOK post-test and Gates McGinitie were 0.41 and 0.35 with the Stanford Achievement Test 10th Edition (SAT-10) (Harcourt Educational Measurement, 2005) reading comprehension subtest. Inter-rater reliability was above 95% agreement between two testers. The directions of the DOK measure have also been translated into Spanish and used in three additional randomized control trials. Inter-rater reliability has been above 90% in the three studies. Internal consistency reliability (Cronbach's α) ranged from 0.84 to 0.89 in the larger study that included the words of the current study (D. L. Baker et al., 2018).

In summary, results from these studies suggest that the paper and pencil DOK measure is reliable, and it can be used to differentiate the vocabulary knowledge within students (i.e., from pre-test to post-test) and between students (i.e., students receiving different type of vocabulary instruction). Thus, the creation of an AS model using this measure can be an enhanced approach to better assess and support the vocabulary development of ELs, particularly of Latino ELs with different cultural and social backgrounds (Goldenberg, Reese, & Rezaei, 2011), and with different levels of English and Spanish language proficiency.

2.3 AS models using captured linguistic features

The AS system based on psycho-linguistic measures of assessment content (PLIMAC) (Sano, 2015, 2016) performed natural language processing (NLP) (Nadkarni, Ohno-Machado, & Chapman, 2011) by capturing the linguistic features from each human-transcribed response and the given correct answer examples. The captured linguistic features were used to classify the response to score 0 to 2 or 0 to 3 by supervised classification methods of tree-based regression (TBR), linear SVM (classification), radial basis function (RBF) kernel SVM (non-linear SVM; hereinafter SVM-rbf), RF, and feedforward neural networks (1 to 5 hidden layers; hereinafter called NN-HL1 to NN-HL5). These supervised classification methods automatically captured complex relationships between human scoring results and the linguistic features of the responses by attempting to replicate the human scoring results.

The NLP module of the AS model counts the numbers of matched lemma[4] words, synonyms,[5] hypernyms,[6] or hyponyms[7] in the student answers with the question word or the correct answer examples. The synonyms, hypernyms, and hyponyms were retrieved from WordNet (Miller, 1995). The NLP module also captured the word frequencies, retrieved from the Open American National Corpus (Reppen, Ide, & Suderman, 2005), as well as Pointwise Mutual Information (PMI) values of adjacent word pairs or triplets in the student answers. PMI is the log ratio of the probability of adjacent word co-occurrence and the product of the probabilities of each word occurrence. PMI of a word pair (bigram AB, A represents the first word and B the second word) is defined as:

$$PMI_{bigram} = \log_2 \frac{p(AB)}{p(A) \times p(B)}. \tag{1}$$

PMI of a word triple (trigram ABC, C represents the third word) is defined as:

[4] Transformed words into their dictionary base forms in order to generalize the comparison analysis. For example, “produced” is normalized as “produce” (Chong, Specia & Mitkov, 2010).

[5] A word whose meaning is nearly the same as another word. For example, “circumstance” and “status” are the synonyms of “condition.”

[6] A word whose meaning includes the meaning of a more specific word. For example, “appear” and “begin” are the hypernyms of “erupt.”

[7] A word whose meaning is included in the meaning of another more general word. For example, “mathematician” and “psychologist” are hyponyms of “scientist.”

$$PMI_{trigram} = \log_2 \frac{p(ABC)}{p(A) \times p(B) \times p(C)}. \quad (2)$$

For example, the PMI value of a word pair *something you* is 0.0124 and the PMI value of a word pair *burst out* is 7.599, indicating the word pair *burst out* is much more tightly connected than the word pair *something you*. As another example, the PMI value of a word triple *when something you* is 1.391 while the PMI value of a word triple *put pressure on* is 12.0196. A summary of the linguistics features that were captured from the transcribed responses to build the classification rules and to score an individual answer are available upon request to the first author.

2.4 AS models not using captured linguistic features

In this study we also included approaches such as LSTM and MemN2N–LSTM that did not require linguistic feature engineering. In these approaches, instead of capturing linguistic features from each student answer and the correct answer examples, the system splits student answers into each word (i.e., a token) and embeds the words to the matrix for self-learning the relationship between correct answer examples and scored student responses. LSTM is capable to remember their inputs over a long period of time and capable of learning long-term dependencies. In this study, a word embedding matrix was generated from the text strings concatenating the correct answer examples and following each student answer.

MemN2N is a neural network with attention mechanism over a possible large external memory and trained end-to-end. In this study, the embedded correct answer examples were stored in memory and each student answer was separately embedded to weigh the response to matched answers.

2.5 AS models used to score student responses

We used the human-transcribed student responses coupled with the human scoring results (0 to 2, 0 to 3) to train the models. Three fourth of the randomly sampled data from each data set was used for classification modeling. The remainder of one fourth was used for the score prediction and evaluation.

For the SVM classification, the linear SVC (SVM classification) model of scikit-learn (Pedregosa et al., 2011), a machine learning toolkit in Python, was used.[8] We

[8] Other than penalty parameter C of the linear SVC model, the default parameter settings were applied. In order to find the best model-scored accuracy with C parameter, a grid search was performed by ranging the C parameter from 10^{-4} to 10^{10}.

also used scikit-learn in Python for RBF kernel SVM[9] and RF.[10] For NN-HL1 to NN-HL5,[11] LSTM, and MemN2N–LSTM[12] the TensorFlow framework (Abadi et al., 2015) was used. We used classification and regression tree (CART) algorithm implemented in PLIMAC (Sano, 2015, 2016) for the TBR classification. The paradigm of pruning the tree was also applied to TBR to avoid overfitting the data. The pruning of the tree feature works by checking these overfitted pairs of nodes (i.e., potentially classified two groups of responses) that have a common parent and verifies if the pruning (merging) the nodes would increase the deviance[13] just within a certain threshold. In this study, the threshold value of the acceptable level was set to 5.0[14] for all the TBRs. The final AS scores by TBR were assigned to each response as the mean score of all questions belonging to the node after rounding off to the nearest integer.

2.6 Data analysis procedure

To analyze the developed classification rules and their results, the AS results were compared to each other and to human scoring results (as the benchmark scores). To obtain the benchmark values of scoring performance, we conducted an inter-rater reliability study by calculating QWK across three human raters for two words (*erupt* and *scientist*). Given that all three raters were assigned to all the two-word responses, three possible pairs out of three rating results were extracted from the human rating results. QWK refers to the summary statistics of AS and human scoring agreement taking into account the agreement expected by chance. The weighting indicates the seriousness of the extent of disagreement as shown below.

[9] Other than parameter C and gamma of the RBF kernel SVC model, the default parameter settings were applied. In order to find the best model-scored accuracy with C and gamma parameter, a grid search was performed by ranging the C parameter from 10^{-2} to 10^{4} and gamma parameter from 10^{-5} to 10^{0}.

[10] Through the all NN-HL models, the number of hidden units were chosen explanatorily ranging from 25 to 200 and the number of training epochs was fixed as 1,000 in the use of the maximum likelihood method with TensorFlow's AdamOptimizer function to minimize the loss function.

[11] A grid search of the parameter space was conducted for the best fit parameters in the use of scikit-learn's GridSearchCV function. The parameters range as their 'n_estimators': [100], 'max_features': [1, 'auto', None], 'max_depth': [1, 5, 10, None], 'min_samples_leaf': [1, 2, 4].

[12] For both of LSTM, and MemN2N–LSTM, the number of training epochs was 20 with the batch size of 100 in the use of the maximum likelihood method with TensorFlow's AdamOptimizer function (the learning rate was 0.001) to minimize the loss function.

[13] The sum of squared differences between an answer score and the mean score of all questions belonging to a single node.

[14] The threshold value was chosen exploratorily through the iterative executions of pruning the tree and the evaluations of the classification performance indices.

$$k = 1 - \frac{\sum_{i,j} w_{i,j} o_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}, \tag{3}$$

where O, w, and E are the matrices of observed scores, weights, and expected scores respectively. Matrix $O_{i,j}$ corresponds to the number of responses that receive a score i by AS (or the first rater) and a score j by human rate (or the second rater). The weight entries are

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^2}, \tag{4}$$

where N is the number of possible ratings. Matrix E is calculated by taking the outer product between the score vectors of machine and human (or the two raters), which are then normalized to have the same sum as O above.

3. Results

We present results by research question as indicated below.

3.1 Which AS model appears to have the highest human-machine agreement as measured by QWK?

Table 2 shows the results of the QWK among all the AS models for definitions and sentences. Results were analyzed with and without a score of 0 indicating that students did not know the word (DK) or NR. As indicated in Table 3, the human-machine agreements were reduced when the DK and NR responses were excluded.

Findings also indicate that across all tasks and data points, RF showed the best performance except in definition task. For definition task with DK and NR, MemN2N–LSTM outperformed (QWK = 0.80) the other AS models. However, the mean of QWK for RF across all tasks and data points was 0.70 followed by MemN2N–LSTM (QWK = 0.69), SVM-rbf (QWK = 0.68 for both), and LSTM (QWK = 0.68). These findings are remarkable given that there is no feature engineering performed by LSTM and MemN2N–LSTM (i.e., the algorithm itself identifies the features and classification rules without any input from other characteristics of the word). For NN-HL1 to NN-HL5, linear SVM and TBR, their mean of QWKs across all tasks and data points were 0.47 to 0.64, suggesting that simple models using captured linguistic features only might not be as precise as more complex models.

Table 2. Human-machine agreements as measured by QWK.

	QWK					
	Definition		Sentence			
AS model	w/ DK NR	w/o DK NR	w/ DK NR	w/o DK NR	Mean	Note
No. of responses	6,778	2,162	6,693	2,299	4,483	
NN-HL1	0.57	0.30	0.64	0.39	0.47	Neural network (1 hidden layer)
NN-HL2	0.75	0.59	0.75	0.47	0.64	Neural network (2 hidden layers)
NN-HL3	0.73	0.53	0.73	0.44	0.61	Neural network (3 hidden layers)
NN-HL5	0.63	0.42	0.65	0.33	0.51	Neural network (5 hidden layers)
RF	0.78^{b}	0.65^{a}	0.79^{a}	0.58^{a}	0.70^{a}	Random forest
SVM	0.70	0.53	0.73	0.45	0.60	Linear SVM (classification)
SVM-rbf	0.77	0.63	0.76	0.55^{b}	0.68	RBF kernel SVM
TBR	0.76	0.61	0.74	0.46	0.64	Tree-based regression
LSTM	0.78^{b}	0.62	0.79^{a}	0.54	0.68	LSTM
MemN2N–LSTM	0.80^{a}	0.64^{b}	0.78^{b}	0.55^{b}	0.69^{b}	End-to-end memory networks + LSTM

Note: ^{a}best performance. ^{b}second best performance.

3.2 What is the inter-rater agreement between human scorers, and between human scorers and AS models?

To answer this question, we first calculated the extent of agreement between two scorers, and between one scorer and three AS models (i.e., RF, LSTM, and MemN2N–LSTM). We selected the RF model because it showed the highest agreement between models as indicated in Table 3. We also used LSTM and the combined MemN2N–LSTM models because they are currently more widely used (Alikaniotis et al., 2016; Taghipour & Ng, 2016; Zhao et al., 2017).

For this question we only used the definitions and sentences for two words: *scientist* and *erupt*. Both words were part of the larger study, and they were taught in-depth by the virtual tutoring system. The definitions for *scientist* and *erupt*, and examples of sentences are included in Table 4. Table 3 includes the extent of agreement between scorers. The mean agreement of scorers was 0.90, while the human-machine agreement of three AS model and humans ranged from 0.66 to 0.70. The reason why the number of responses on the top and bottom in Table 3 vary is because the data for human-human agreements is not the complete set but the

Table 3. Human-human agreements as measured by QWK.

	QWK					
	Definition		Sentence			
AS model	w/ DK NR	w/o DK NR	w/ DK NR	w/o DK NR	Mean	Note
No. of responses	404	282	404	282	345	
Rater1-rater2	0.97	0.96	0.89	0.84	0.91	No score 3 rating observed
Rater1-rater3	0.95	0.93	0.86	0.79	0.88	
Rater2-rater3	0.93	0.90	0.91	0.87	0.90	
Human-human mean	0.95	0.93	0.89	0.83	0.90	
No. of responses	696	390	696	393	544	No. of score 3 sentence task response eliminated
RF	0.78	0.71	0.73	0.57	0.70	12 for w/ and 12 for w/o DK NR
LSTM	0.74	0.63	0.76	0.53	0.66	13 for w/ and 12 for w/o DK NR
MemN2N–LSTM	0.77	0.66	0.74	0.57	0.69	14 for w/ and 12 for w/o DK NR
Human-machine mean	0.76	0.67	0.74	0.55	0.68	

Table 4. Definition and examples of sentences for the words *erupt* and *scientist.*

Word	Definition	Sentence
Erupt	It means to explode or to burst out with force.	• This volcano close to Mexico City sometimes erupts. • A person is shaking a soda can. When he opens the can, the soda erupts. • The heat caused the dry grass to erupt into flames. • My parents erupt into cheers when I score a goal. • When you hear a really funny joke, you might erupt into laughter.
Scientist	A person who studies how the world works through observations and experiments.	• Ellen Ochoa studies how technology can be used in space. She is a scientist. • Jane Goodall studied how chimpanzees live in the jungle. She was a scientist. • Louis Pasteur studied how heating up juice and milk makes them safe to drink. He was a scientist. • Mario Molina studies how pollution affects the environment. He is a scientist.

subset of the responses from the larger study. Note that on the bottom of Table 3, the responses scored 3 by human or machine are eliminated from the data because we could not find a score of 3 in the data scored by three humans for the words *scientist* and *erupt*.

3.3 Which student responses appeared to have the maximum extend of disagreement between humans and AS models?

To respond to Question 3, we compared the scoring of student definitions and sentences for the words *erupt* and *scientist* to illustrate some of the challenges of developing an AS system to score student utterances. To answer this question, we used the data without the DK and NR responses. We explain potential reasons why three AS models (RF, LSTM, and MemN2N–LSTM) might have scored student utterances differently.

3.3.1 Misclassification of responses for the word *erupt* in the definition task

As it shows on the top-left side of Table 5, RF has the maximum number of misclassification of the responses (7) at the maximum extent of disagreement (scored 2 by human, but scored 0 by AS) across three AS models. RF is an AS model that uses captured linguistic features that depend heavily on a linguistic feature "count match with correct answer" for the scoring of definition task, which is likely the primary cause of the misclassification. The correlation between the AS scores (RF) and the feature values of "count match with correct answer" is 0.47 as the highest one among all linguistic features. For example, a typical response that was misclassified by RF is "Like an explosion." Given that this response has no matched

Table 5. Number of misclassifications of the responses at the maximum extent of disagreement.

	Word						
	Erupt			Scientist			
Human/AS scores	2/0	0/2	Word total	2/0	0/2	Word total	Total
Panel A. Definition							
RF	7	1	8	1	2	3	11
LSTM	4	1	5	0	9	9	14
MemN2N–LSTM	5	4	9	1	3	4	13
Total	16	6	22	2	14	16	38
Panel B. Sentence							
RF	8	3	11	5	5	10	21
LSTM	3	7	10	3	7	10	20
MemN2N–LSTM	4	4	8	5	5	10	18
Total	15	14	29	13	17	30	59

word with the correct answer, the response received the lowest score by RF as its linguistic feature value of “count match with correct answer” was 0.

On the other hand, LSTM scored the response as 1 and MemN2N–LSTM scored as 2. LSTM and MemN2N–LSTM do not use the captured linguistic features and do not perform word matching-based similarity measuring as RF does. Instead, the networks apply a word embedding technique to learn and quantify the similarity of the words by applying empirically captured probability distributions of the word occurrence.

MemN2N–LSTM had the maximum number of misclassification (4) of the responses at the maximum extent of disagreement (scored 0 by human but scored 2 by AS) across three AS models for the word *erupt*. Even though MemN2N–LSTM does not use the captured linguistic features directly, it has an attention mechanism weighting the response words matched with correct answers. This could have been the primary cause of the misclassification. For example, one of the typical responses misclassified by MemN2N–LSTM is “To talk out when someone else says something.” Since this response has a matched word *out* with the correct answer example, the response received the highest score by MemN2N–LSTM. A plausible reason for the misclassification by MemN2N–LSTM is because there is currently only one correct definition provided for the word *erupt*, which is not enough to maximize the capability of attention mechanism that weights the response words by matching the response to several correct plausible answers. RF and LSTM, on the other hand, scored the response as 0 just like the human scorers.

3.3.2 Misclassification of responses for the word *scientist* in the definition task

On the top-right side of Table 5, LSTM shows the maximum number of misclassifications (9) of the responses at the maximum extent of disagreement (scored 0 by humans, but scored 2 by AS) across three AS models for the word *scientist*. LSTM had a large number of misclassifications. For example, LSTM misclassified the sentence “He gets a fruit topic or animal.” This response has no matched word with the correct answer example, suggesting that the student utterance quantified by LSTM is not appropriate. On the other hand, RF and MemN2N–LSTM scored the response as 0, the same as a human scorer.

3.3.3 Misclassification of responses for the word *erupt* in the sentence task

Similar to the definition task, RF has the maximum number of misclassification (8) of the responses at the maximum extent of disagreement (scored 2 by human but scored 0 by the AS model) on the bottom-left side of Table 5 across three AS models for the word *erupt*. The primary cause for the misclassification might be the dependence of RF on a linguistic feature PMI trigram max for scoring sentences. An example of misclassification was the sentence: “When you shake a bottle it will erupt.” The response has the PMI trigram max value of 0 (i.e., there is no match with the trigram in the reference answers). LSTM scored the response as 1 with one-point disagreement and MemN2N–LSTM scored as 2 with no disagreement with the human score, suggesting that LSTM and MemN2N–LSTM have a different strategy to quantify the probability of adjacent word co-occurrence as the recurrence of the words is indexed by memory lookups to the word sequence in MemN2N (Sukhbaatar et al., 2015) and indexed by the sequence itself in LSTM.

LSTM has the maximum number (7) of misclassification of the responses at the maximum extent of disagreement (scored 0 by humans, but scored 2 by the AS model). LSTM has the capability to maintain information in memory for longer periods than RNN. Given this characteristic, the correlations between the AS scores (LSTM) and the feature values of word count were the highest as 0.52 among the all linguistic features even though LSTM does not use the captured linguistic features directly. One of the typical responses misclassified by LSTM is “Erupt is when someone is telling you about another person and you do not want to hear from them.” Since this response is the second-longest one across all the responses to the two words as its word count is 19, it is likely to be the primary cause of the highest score biased by LSTM. On the other hand, RF and MemN2N–LSTM scored the response as 0 exactly like the human scorers.

3.3.4 Misclassification of responses for the question word *scientist* in the sentence task

Similar to the word *erupt*, LSTM had the maximum number of misclassifications (7) of the responses at the maximum extent of disagreement (scored 0 by human, but scored 2 by AS). One of the typical responses misclassified by LSTM was “Scientist can make all sorts of different things.” This is another example of LSTM’s higher score bias with longer word count as the response got the highest score. On the other hand, the RF model scored the response as 0 with no disagreement and the MemN2N–LSTM model scored it as 1 with one-point disagreement with the human score.

Another example of misclassifications at the maximum extent of disagreement (scored 2 by human but scored 0 by LSTM and MemN2N–LSTM) was the sentence “Albert Einstein refused to wear socks; Albert Einstein is a scientist.” Albert Einstein cannot be found in the correct answer examples. However, if the name Albert Einstein were in the answer examples, the classification accuracy would have improved.

4. Discussion

The purpose of this study was to explore the reliability of seven different AS models to score student responses on a vocabulary measure administered in English and in Spanish, and previously scored by data collectors with high reliability in five different studies. In this study, a total of 13,471 English utterances related to academic vocabulary knowledge were analyzed. Second grade Latino students were asked to define a target word and use the prompt word in a sentence. The AS system based on PLIMAC (Sano, 2015, 2016) performed NLP by capturing the linguistic features from each human-transcribed response and the given correct answer examples. In addition, we also used LSTM and MemN2N–LSTM models that do not require linguistic feature engineering. To analyze the developed classification rules and their results, the AS results were compared to each other and to human scoring results by calculating QWK. Findings from this study suggest that (1) the most accurate AS models for our data appear to be RF and MemN2N–LSTM, (2) extending the answer examples might increase the correlations between human scoring and machine scoring, and (3) AS models have the potential to score EL utterances even for students in the lower grades. To our knowledge, this is the first AS study that has been conducted with students in the lower elementary grades. It is also the first study that compared the agreement of several different systems to each other, and to human scoring. We discuss our outcomes in the context of other studies.

4.1 Closest human-machine agreement

Across all tasks and data points, RF showed the best performance followed by MemN2N–LSTM, SVM-rbf, and LSTM. These findings are remarkable given that LSTM and MemN2N–LSTM do not require any feature engineering. Thus, the fact that the RF model that relies on linguistic features appeared to be more reliable is notable. Nonetheless, the advantage of LSTM and MemN2N–LSTM is its word embedding mechanism which represents tight connections of the vocabularies (Perfetti, 2007) by the semantic networks. Thus, if the systems are trained well, they should outperform other models using captured linguistic features. Given that this was not our case, below we provide two suggestions that might increase the accuracy of all the models, including memory based models.

4.1.1 Expanding the human-machine agreement

To reduce misclassifications and increase the accuracy of the systems, a proposed suggestion is to incorporate more answer examples in both definition and sentence tasks by incorporating publicly available language resources with pre-trained embeddings. Additional examples whether drawn from a publicly available corpus of responses, or from a larger pool of responses by students, could increase the accuracy of the systems independently of whether the AS model focus on capturing linguistic features such as RF, or whether they use memory (e.g., LSTM and MemN2N–LSTM) to match student answers to previous responses and examples. Alikaniotis et al. (2016), for example, used pre-trained embeddings of word2vec (Mikolov, Chen, Corrado, & Dean, 2013) as a technique to learn word embeddings using neural networks on a corpus of words created by Google News to improve their AS model with positive results. We anticipate obtaining similar results with a larger pool of answers.

4.1.2 Examining more closely the discrepancies between human scoring and AS models

Results from our inter-rater reliability study across three human raters for two words (*erupt* and *scientist*) indicated that the mean extent of agreement as examined by QWK for humans was 0.90, substantially higher than the human-machine agreement of the three AS models (i.e., RF, LSTM, and MemN2N–LSTM) that ranged from 0.66 to 0.70. In reviewing student responses that appeared to have the maximum extend of disagreement between humans and AS models to garner a better understanding of the nature of the misclassifications by the different models, our findings indicated that RF had the highest number of misclassification responses (i.e., responses were scored with a 2 by a human scorer, but scored with a 0 by RF). A potential reason for the misclassification could be that RF heavily depends on a linguistic feature count match with correct answer for the scoring of definition task and on a linguistic feature PMI trigram max for the scoring of sentence task. Thus, these linguistic features are more sensitive to unexpected responses. For example, in the response of definition task: Something that blows lava out, RF might not have recognized the semantic use of erupt in this sentence, and therefore scored the response as 0. On the other hand, LSTM and MemN2N–LSTM scored the response with a 2 just like human scorers, perhaps because it was able to identify the different uses of erupt from previous answers.

4.2 Limitations

A major challenge in this study was to find ways to maximize the use of state-of-the-art AS models with short answers provided by second grade students. A potential solution could be to use outside linguistic resources and pre-trained models to develop more accurate and reliable AS models. However, we also recognize that it is impossible to incorporate every word into an AS model, and particularly proper names such as Albert Einstein and Mt. Saint Helen. Nonetheless, these names could be widely recognized and associated with the word *scientist* and *erupt*, at least in the U.S. Therefore, incorporating some common proper names into a corpus of potential responses could also increase the reliability of the AS models. We also acknowledge that increasing the number of responses students provided to more words, could also help train the systems, particularly for models that rely more on relating current responses to previous ones such as in the LSTM and MemN2N–LSTM models.

4.3 Future research and conclusions

In the future we intend to use two separate models, one for definition and one for sentence tasks. In addition, we plan to expand the answer examples through either outside linguistic resources or pre-trained models in order for the AS models we select to identify more precisely accurate responses. Nonetheless, this exploratory study presents a unique opportunity to better understand how AS models perform compared to human scoring. As more natural language data are collected by researchers through technology, there is a need to find more efficient and accurate ways to process and score these language data, particularly in the context of vocabulary instruction. Moreover, in MTSS of instructional support, AS of student speech is a promising technique to help teachers immediately obtain results that will help them differentiate their vocabulary and content instruction. As the number of ELs increase in classrooms worldwide, the need to capture their level of understanding of content through their transcribed speech is imperative to make instructional decisions that will lead to an increase in their academic performance.

Acknowledgment

This research was supported by grant no. R305A140471, titled "English Learner Vocabulary Acquisition (ELVA): Promoting the Vocabulary and Language Proficiency of Spanish Speaking ELs in Second Grade," funded by the U.S. Department of Education (USDOE), Institute of Education Sciences (IES) to Southern Methodist University. The views expressed within this article are those of the authors and do not represent those of IES or the USDOE.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., . . . Corrado, G. S. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems*. Retrieved from https://research.google/pubs/pub45166

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 715–725). Berlin, Germany: Association for Computational Linguistics. doi:10.18653/v1/P16-1068

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In Guthrie J. T. (Ed.), *Comprehension and teaching: Research reviews* (pp. 77–117). Newark, DE: International Reading Association.

August, D., & Shanahan, T. (2006). *Developing literacy in second language learners: Report of the national literacy panel on language-minority children and youth*. Mahwah, NJ: Lawrence Erlbaum.

Baker, D. L., Basaraba, D. L., & Richards-Tutor, C. (2018). An introduction to second language acquisition: The role of academic English across content areas. In D. L. Baker, C. Richards-Tutor, & D. Basaraba (Eds.), *Second language acquisition: Methods, perspectives, and challenges* (pp. 1–16). New York, NY: Nova Science Publishers.

Baker, D. L., & Kosty, D. (2012). Effect of learning opportunities on the reading performance in Spanish of first grade English learners. In D. L. Baker (Chair), *The relation between observations of student-teacher interactions and student performance*. Symposium conducted at the Pacific Coast Research Conference, San Diego, CA.

Baker, D. L., Ma, H., Polanco, P., Conry, J., Kamata, A., Al Otaiba, S., . . . Cole, R. (2020). Development and promise of a vocabulary intelligent tutoring system for second-grade Latinx English learners. *Journal of Research on Technology in Education*. doi:10.1080/15391523.2020.1762519.

Baker, S. K., & Baker, D. L. (2008). English learners and response to intervention: Improving quality of instruction in general and special education. In E. L. Grigorenko (Ed.), *Educating individuals with disabilities: IDEA 2004 and beyond* (pp. 249–273). New York, NY: Springer.

Baker, S. K., Santoro, L. E., Chard, D. J., Fien, H., Park, Y., & Otterstedt, J. (2013). An evaluation of an explicit read aloud intervention taught in whole-classroom formats in first grade. *Elementary School Journal*, *113*, 331–358. doi:10.1086/668503

Cena, J. S., Baker, D. L., Kame'enui, E. J., Baker, S. K., Park, Y., & Smolkowski, K. (2013). The impact of a systematic and explicit vocabulary intervention in Spanish with Spanish-speaking English learners in first grade. *Reading and Writing: An Interdisciplinary Journal*, *6*, 1289–1316. doi:10.1007/s11145-012-9419-y

Chen, J., Fife, J. H., Bejar, I. I., & Rupp, A. A. (2016). *Building E-Rater® scoring models using machine learning methods* (ETS Research Report No. RR-16-04). Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12094

Chen, L., Zechner, K., Yoon, S.-Y., Evanini, K., Wang, X., Loukina, A., . . . Gyawali, B. (2018). *Automated scoring of nonnative speech using the SpeechRaterSM v. 5.0 Engine* (ETS Research Report No. RR-16-04). Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12198

Chong, M., Specia, L., & Mitkov, R. (2010) *Using natural language processing for automatic detection of plagiarism*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.458.9440&rep=rep1&type=pdf

Christ, T. (2011). Moving past "right" or "wrong" toward a continuum of young children's semantic knowledge. *Journal of Literacy Research*, *43*, 130–158. doi:10.1177/1086296X11403267

Common Core State Standards Initiative. (2010). *Common core standards for English language arts & literacy in history/social studies, science, and technical subjects*. Retrieved from http://www.corestandards.org/assets/CCSSI_ELAStandards.pdf

Crossley, S. A., & McNamara, D. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology*, *17*, 171–192.

Elleman, A. M., Lindo, E. J., Morphy, P., & Compton, D. L. (2009). The impact of vocabulary instruction on passage-level comprehension of school-age children: A meta-analysis. *Journal of Research on Educational Effectiveness*, *2*, 1–44. doi:10.1080/19345740802539200

Gersten, R., Baker, S., & Lloyd, J. W. (2000). Designing high-quality research in special education: Group experimental design. *The Journal of Special Education*, *34*, 2–18. doi:10.1177/002246690003400101

Geva, E., & Farnia, F. (2012). Developmental changes in the nature of language proficiency and reading fluency paint a more complex view of reading comprehension in ELL and EL1. *Reading and Writing: An Interdisciplinary Journal*, *25*, 1819–1845. doi:10.1007/s11145-011-9333-8

Goldenberg, C., Reese, L., & Rezaei, A. (2011). Contexts for language and literacy development among dual-language learners. In A. Yücesan Durgunoğlu & C. Goldenberg (Eds.), *Language and literacy development in bilingual settings* (pp. 3–28). New York, NY: Guilford Press.

Harcourt Educational Measurement. (2005). *Stanford Achievement Test (SAT-10 edition)*. San Antonio, TX: Author.

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, *21*, 303–317. doi:10.1017/S0272263199002089

Hiebert, E. H. (2006, April). *A principled vocabulary curriculum*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780. doi:10.1162/neco.1997.9.8.1735

Kena, G., Musu-Gillette, L., Robinson, J., Wang, X., Rathbun, A., Zhang, J., & Dunlop Velez, E. (2015). *The condition of education 2015* (NCES 2015-144). Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubs2015/2015144.pdf

Kieffer, M. J. (2010). Socioeconomic status, English proficiency, and late-emerging reading difficulties. *Educational Researcher*, *39*, 484–486. doi:10.3102/0013189X10378400

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*, 190–208. doi:10.1111/j.1540-

4781.2011.01232.x

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. Retrieved from https://arxiv.org/abs/1301.3781

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*, 39–41. doi:10.1145/219717.219748

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association*, *18*, 544–551. doi:10.1136/amiajnl-2011-000464

NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states*. Washington, DC: The National Academies Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*, 357–383. doi:10.1080/10888430701530730

Phandi, P., Chai, K. M. A., & Ng, H. T. (2015). Flexible domain adaptation for automated essay scoring using correlated linear regression. In L. Màrquez, C. Callison-Burch, & J. Su (Eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 431–439). Stroudsburg, PA: Association for Computational Linguistics. doi:10.18653/v1/D15-1049

Reppen, R., Ide, N., & Suderman, K. (2005). *American National Corpus (ANC) second* release (LDC2005T35) [Data set]. Philadelphia, PA: Linguistic Data Consortium.

Sano, M. (2015). *Automated capturing of psycho-linguistic features in reading assessment text*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Sano, M. (2016). *Improvements in automated capturing of psycho-linguistic features*. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Sano, M., Baker, D. L., & Kamata, A. (2018, September). *テキスト化された英語口頭試験結果の自動採点：SVMと回帰木アルゴリズムの比較* [*Automatic scoring of textualized English oral exam results: Comparison of SVM and regression tree algorithm*]. Paper presented at the 16th meeting of Japanese Associaion for Research on Testing, Tokyo, Japan.

Schoonen, R., & Verhallen, M. (2008). The assessment of deep word knowledge in young first and second language learners. *Language Testing*, *25*, 211–236. doi:10.1177/0265532207086782

Shanahan, T., & Beck, I. (2006). Effective literacy teaching for English-language learners. In D. L. August & T. Shanahan (Eds.), *Developing literacy in a second language: Report of the National Literacy Panel* (pp. 415–488). Mahwah, NJ: Lawrence Erlbaum.

Somasundaran, S., & Chodorow, M. (2014). Automated measures of specific vocabulary knowledge from constructed responses ("use these words to write a sentence based on this picture"). In J. Tetreault, J. Burstein, & C. Leacock (Eds.), *Proceedings of the 9th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 1–11). Stroudsburg, PA: Association for Computational Linguistics. doi:10.3115/v1/W14-1801

Sukhbaatar, S., Szlam, A., Weston, J., & Fergus, R. (2015). End-to-end memory networks. In C. Cortes, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Proceedings of the 28th International Conference on Neural Information Processing Systems* (pp. 2440–2448). Cambridge, MA: MIT Press. doi:10.5555/2969442.2969512

Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In J. Su, K. Duh, & X. Carreras (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language* (pp. 1882–1891). Stroudsburg, PA: Association for Computational Linguistics. doi:10.18653/v1/D16-1193

Texas Education Agency. (2010). *Texas essential knowledge and skills*. Retrieved from https://tea.texas.gov/Academics/Curriculum_Standards/TEKS_Texas_Essential_Knowledge_and_Skills

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress. (2019). *Various years, 1990–2019 mathematics and reading assessments*. Retrieved from https://www.nationsreportcard.gov/mathematics/supportive_files/2019_infographic.pdf

Wanzek, J. (2014). Building word knowledge: Opportunities for direct vocabulary instruction in general education for students with reading difficulties. *Reading and Writing Quarterly*, *30*, 139–164.

Weston, J., Chopra, S., & Bordes, A. (2015, May). *Memory networks*. Paper presented at the 3rd International Conference on Learning Representations, San Diego, CA.

Yoon, S.-Y., Bhat, S., & Zechner, K. (2012). Vocabulary profile as a measure of vocabulary sophistication. In Association for Computational Linguistics (Ed.), *Proceedings of the 7th Workshop on Building Educational Applications Using*

NLP (pp. 180–189). Stroudsburg, PA: Association for Computational Linguistics. doi:10.5555/2390384.2390406

Yoon, S.-Y., Loukina, A., Lee, C. M., Mulholland, M., Wang, X., & Choi, I. (2018). Word-embedding based content features for automated oral proficiency scoring. In L. E. Anke, D. Gromann, & T. Declerck (Eds.), *Proceedings of the 3rd Workshop on Semantic Deep Learning* (pp. 12–22). Stroudsburg, PA: Association for Computational Linguistics.

Zechner, K., & Bejar, I. (2006). Towards automatic scoring of non-native spontaneous speech. In R. C. Moore, J. Bilmes, J. Chu-Carroll, & M. Sanderson (Eds.), *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 216–223). New York, NY: Association for Computational Linguistics.

Zhao, S., Zhang, Y., Xiong, X., Botelho, A., & Heffernan, N. (2017). A memory-augmented neural model for automated grading. In C. Urrea, J. Reich, & C. Thille (Eds.), *Proceedings of the 4th Annual Association for Computing Machinery Conference on Learning at Scale* (pp. 189–192). New York, NY: Association for Computing Machinery. doi:10.1145/3051457.3053982