Journal of Pan-Pacific Association of Applied Linguistics, 20(1), 55-73.

Automated Scoring of L2 Spoken English with Random Forests

Yuichiro Kobayashi[*]
Toyo University

Mariko Abe
Chuo University

Kobayashi, Y., & Abe, M. (2016). Automated scoring of L2 spoken English with random forests. *Journal of Pan-Pacific Association of Applied Linguistics*, *20*(1), 55-73.

The purpose of the present study is to assess second language (L2) spoken English using automated scoring techniques. Automated scoring aims to classify a large set of learners' oral performance data into a small number of discrete oral proficiency levels. In automated scoring, objectively measurable features such as the frequencies of lexical and grammatical items are generally used as "exploratory variables" to predict oral proficiency levels, any of which can be used as a "criterion variable" in this study. We have chosen the NICT JLE Corpus, a corpus of 1,281 Japanese EFL learners' speech productions coded into nine oral proficiency levels (Izumi, Uchimoto, & Isahara, 2004). The nine oral proficiency levels were used as the criterion variables and linguistic features analyzed in Biber (1988) as explanatory variables. We employed random forests (Breiman, 2001), a powerful method for text classification and feature extraction, to predict oral proficiency. As a result of random forests with the out-of-bag error estimate, 60.11% of the productions were correctly classified. Compared to the baseline accuracy of the simplest possible algorithm of always choosing the most frequent level (37.63%), our random forests model improved prediction by 22.48 points. The Pearson product-moment correlation coefficient with human scoring was 0.85. Predictors that showed a clear discrimination of oral proficiency levels were tokens, types, and the frequency of nouns in the order of strength.

Keywords: automated scoring, L2 spoken English, learner corpus, random forests

[*] First author: Yuichiro Kobayashi; second author: Mariko Abe.

[*] An earlier version of the present study was reported at the Second Asia Pacific Corpus Linguistics Conference held at Hong Kong Polytechnic University on March 7th, 2014. This work was supported by Grants-in-Aid for Scientific Research Grant Numbers 12J02669, 24520631, and 26770205.

1 Introduction

As a general trend of language testing, learners' performance is usually assessed by multiple human raters. This has an advantage in construct representation, but carries its disadvantages in halo effects, sequence effects, and central tendency effects. Further, apart from the expense, a long time period will be required for training proficient human raters. However, automated scoring systems can assess the performance of a large number of learners with higher speed and more adequate reliability.

Nowadays, discussions focus on whether automated scoring should be introduced to large-scale high-stakes tests such as the university entrance examination. For instance, the Korea Institute for Curriculum and Evaluation (KICE) has developed some automated scoring systems for assessing Korean English learners' speaking ability (Shin et al., 2013). Similarly, in Japan, the Ministry of Education, Culture, Sports, Science and Technology (MEXT) has explored the possibility of introducing an automated scoring program to the new university entrance examination slated to start in 2020. However, the accuracy of the automated assessment of learners' spontaneous speeches is heavily dependent on that of speech recognition systems (Xi, 2010). Thus, the current automated speech scoring systems focus mainly on the prosodic information of speech such as intonation and stress patterns, due to the mechanical limitation. In other words, they can hardly use the linguistic information for the assessment implying that they cannot fully cover the entire construct in the same manner as professional human raters can.

In the present study, we examine the effectiveness of an automated scoring system by investigating how well it can assess second language (L2) spoken English as compared to the use of human scoring. We also identify the linguistic features that can be used to assess learners' speaking performance.

2 Literature Review

2.1 Automated scoring

Automated scoring system is the ability of computer technology to evaluate and score written or spoken production (Shermis & Burstein, 2003). It can classify a large set of learners' productions into a small number of discrete proficiency levels. Since the 1960s, most automated scoring systems have been working on written productions. Major automated essay scoring systems include PEG (Page, 1994), e-rater (Burstein et al., 1998), IEA (Landauer, Laham, & Foltz, 2000), Betsy (Rudener & Liang, 2002), and IntelliMetric (Elliot, 2003). However, in recent years, some automated speech scoring systems, for example, Versant (Downey, Farhady, Present-Thomas, Suzuki, & Van Moere, 2008) and SpeechRater (Zechner, Higgins, & Williamson, 2009), have developed in response to the growing importance of oral communication, though they still have room for improvement with pertinence to the recognition of spontaneous accented speech. Zechner, Higgins, and Williamson (2009) show that the error rate of their speech recognizer was around 50%, and Bhat and Yoon (2015) also show that theirs was not lower than 27%. Owing to the lack of accurate speech recognition systems, it is difficult for the current automated speech scoring systems to accurately evaluate learners' spoken language from aspects other than intonation and stress patterns. Although there are several pilot studies focusing on linguistic aspects of L2 speech, such as vocabulary usage (Yoon, Bhat, & Zechner, 2012), syntactic complexity (Chen & Zechner, 2011), or topical content (Xie, Evanini, & Zechner, 2012), little is known about the whole set of linguistic features that can be used to fully assess a learner's communicative competence.

Automated scoring has been criticized for the "brute-empirical approach" (quoted by Shermis, Burstein, & Bursky, 2013, p. 2) which assesses learners' production by using a limited number of quantitative information such as word frequencies and sentence length (e.g., Ericsson, 2006; McGee, 2006). However, as Page (2003) already stated more than 10 years ago, no obvious differences exist between the performance of computers and human performance for scoring results. He examined the correlations among five judges, four human raters and one computer system, and found that the results of five judges correlated with each other at 0.5 approximately.

The comparison of automated and human scoring has long been discussed in the field of language testing, and it is generally believed that automated scoring is strong in conscientiousness and consistency, but weak in construct representation (Bejar, Williamson, & Mislevy, 2006; Williamson, 2013). However, human raters do not necessarily ensure that the construct is appropriately represented in their ratings (Attali, 2013). Furthermore, even professional human raters do not have a comprehensive understanding of their own rating process. On the other hand, automated scoring has an advantage in estimating the weighting of each criterion, which is used in the human assessment. Therefore, the visibility of automated scoring processes allows us to deeply understand the nature and characteristics of human scoring processes. The understanding of the strengths and limitations of human scoring helps to improve automated scoring systems with regard to the quality and relevance of measurement.

2.2 Learner corpus studies

Learner corpora are essential to the calibration of automated scoring engines as well as the identification of linguistic characteristics that can clearly discriminate different proficiency levels (Higgins, Ramineni, & Zechner, 2015). Learner corpora elucidate the specific learning abilities or the lack thereof at a specific developmental stage. For instance, Meunier and Littre (2013) tracked French learners' progress in the acquisition of the English tense and aspect system by analyzing the Longitudinal Database of Learner English (LONGDALE), and showed that tense and aspect errors decrease over time. Thewissen (2013) also traced the accuracy developmental patterns by examining more than 40 error types found in the International Corpus of Learner English (ICLE), and reported that a difference exists in the error patterns between the B1 and B2 Common European Framework of Reference (CEFR) levels. Furthermore, Hawkins and Filipović (2012) described both positive features (correct properties) and negative features (errors) for the different levels of the CEFR using the Cambridge Learner Corpus (CLC). They demonstrated that certain sentence structures such as *that*-clauses and infinitives can distinguish each CEFR level, and that several lexico-grammatical errors significantly decrease as the level goes up. Although these descriptive studies provided a number of implications for language teaching and testing, few attempts have been made to verify the effectiveness of linguistic features automatically extracted from learner corpora in language testing.

3 Purpose

The purpose of the present study is to automatically assess second language spoken English using corpus linguistic and machine learning techniques, and to identify linguistic features relevant for predicting oral proficiency. The following research questions were investigated:

1) How accurately can the automated scoring system predict learners' oral proficiency level assessed by professional human raters?
2) Which linguistic features can be useful for predicting the oral proficiency level of learners?

By pursuing these research questions, this study aims to bridge a gap between automated scoring applications seeking more appropriate outcome measures and learner corpus studies describing learners' developmental patterns. Moreover, it sheds a light on the process of human scoring by uncovering the relationship between actual learners' production and their test scores.

4 Procedure

4.1 NICT JLE corpus

The corpus used in this study is the NICT JLE (National Institute of Information and Communications Technology Japanese Learner English)

Corpus (Izumi, Uchimoto, & Isahara, 2004). It is a learner corpus consisting of Japanese learners of English. It comprises of the transcription of 325 hours of interview sessions, conducted with 1,281 test takers. The 15-minute oral proficiency test is called the Standard Speaking Test (SST), and its assessment criteria conform to those of the American Council on the Teaching of Foreign Language Oral Proficiency Interview (ACTFL OPI). However, this test is designed for assessing the oral proficiency levels of Japanese EFL learners. The test is interactive and uses a one-on-one approach; moreover, test-takers have no planning time and they are not permitted to use any references. It consists of five stages: (a) warm-up questions, (b) a single picture description, (c) a role-play task, (d) sequential picture storytelling, and (e) wind-down questions. After the interview, two certificated raters assess the learner's speaking proficiency in accordance with the SST Manual (ACTFL-ALC Press, 1996). If the two raters find that their ratings do not agree, the Master Rater provides the third and final rating. Table 1 summarizes the number and percentage of learners and the tokens of each oral proficiency level. As indicated, 66.20% of the learners were assessed as levels 4, 5, and 6, and 37.63% of them were classified into the most common level (level 4).

Table 1. Distribution of Oral Proficiency Level of Japanese Learners in the NICT JLE Corpus

Level	Learners	Tokens
1	3 (0.23%)	428 (0.04%)
2	35 (2.73%)	7,701 (0.81%)
3	222 (17.33%)	95,169 (9.98%)
4	482 (37.63%)	308,177 (32.31%)
5	236 (18.42%)	203,759 (21.36%)
6	130 (10.15%)	130,492 (13.68%)
7	77 (6.01%)	85,309 (8.94%)
8	56 (4.37%)	68,470 (7.18%)
9	40 (3.12%)	54,341 (5.70%)
Total	1,281 (100.00%)	953,846 (100.00%)

4.2 Variables

Automated scoring can be considered as an application of machine learning (Larkey & Croft, 2003). In machine learning, objectively measurable features such as the frequencies of lexical and grammatical items in the learners' performance are used as "explanatory variables" for predicting scores chosen as "criterion variables." We used nine levels of the SST as the criterion variables in the automated scoring. By using a wide range of linguistic features, such as parts-of-speech, grammar, and discourse features, as explanatory variables, learners' speech can be evaluated from various angles.

4.2.1 Criterion variables

In the present study, the learner corpus data employed were assessed by multiple certificated raters. The recorded spoken interviews were rated based on the following holistic assessment: (a) global tasks and functions, (b) social context and content area, (c) fluency and the accuracy of grammar, vocabulary, pronunciation, and (d) text types and quality of speech.

Nine oral proficiency levels of the SST were used as the criterion variables. Comparison with the oral proficiency levels of the ACTFL OPI shows that the SST emphasizes novice (levels 1 to 3) and intermediate (levels 4 to 6) learners of English (see Table 2). The difference between these two proficiency scales can be caused by the oral proficiency distribution of Japanese learners, because many Japanese learners are categorized into the SST levels 4, 5 or 6 (see Table 1), which correspond to the "intermediate low" and "intermediate mid" in the ACTFL OPI.

Table 2. A comparison of Oral Proficiency Levels of the ACTFL OPI and the SST (Based on Lewis, 1999, p. 28)

ACTFL OPI	SST
Superior	9 (Advanced)
Advanced High	
Advanced Mid	
Advanced Low	
Intermediate High	8 (Intermediate High)
Intermediate Mid	7 (Intermediate Mid-plus)
	6 (Intermediate Mid)
Intermediate Low	5 (Intermediate Low-plus)
	4 (Intermediate Low)
Novice High	3 (Novice High)
Novice Mid	2 (Novice Mid)
Novice Low	1 (Novice Low)

4.2.2 Explanatory variables

As already mentioned above, we still do not have a full understanding of the human rating system, though automated scoring and human rating are expected to measure the same constructs. For example, Attali (2013) suggests the possibility of using two different types of feature sets, a wide range of 67 linguistic features, which were used in Biber (1988), and linguistic measurement, such as fluency, lexical and syntactic complexity, and accuracy. The linguistic features of Biber (1988) are broadly used in the field of corpus linguistics to examine variation in natural language texts (e.g., Conrad & Biber, 2001; Frignal, 2013; Sardinha & Pinto, 2014). Applying this trend to learner corpus research, our previous studies have already succeeded in discriminating learners' first language (Abe, Kobayashi, & Narita, 2013), describing the overall patterns of variation across different oral proficiency levels (Abe, 2014), and identifying the frequently used linguistic items in learners' performance on different writing tasks (Kobayashi & Abe, 2014). In the present study, 60 linguistic features were selected from the original list of 67 linguistic features in Biber (1988). Seven features, namely (a) demonstratives, (b) gerunds, (c) present participial clauses, (d) past participial clauses, (e) present participial WHIZ deletion relatives, (f) sentence relatives, and (g) subordinator-that deletion, were not included in this study because differences were found in the software used to annotate part-of-speech tags. However, we added three more features (i.e., tokens, types, and the mean length of sentences) as the measures of fluency. Linguistic features used in this study are listed in Table 3.

Table 3. Linguistic Features Analyzed in the Present Study

A. Tense and aspect markers
1. past tense, 2. perfect aspect, 3. present tense
B. Place and time adverbials
4. place adverbials, 5. time adverbials
C. Pronouns and pro-verbs
6. first person pronouns, 7. second person pronouns, 8. third person pronouns (excluding *it*), 9. pronoun *it*, 10. demonstrative pronouns, 11. indefinite pronouns, 12. pro-verb *do*
D. Questions
13. direct WH-questions
E. Nominal forms
14. nominalizations, 15. other total nouns
F. Passives
16. agentless passives
17. *by*-passives
G. Stative forms
18. *be* as main verb, 19. existential *there*
H. Subordination
H1. Complementation
20. *that* verb complements, 21. *that* adjective complements, 22. WH-clauses, 23. Infinitives
H2. Participial forms
24. past participial postnominal (reduced relative) clauses
H3. Relatives
25. *that* relatives in subject position, 26. *that* relatives in object position, 27. WH relatives in subject position, 28. WH relatives in object position, 29. WH relatives with fronted preposition
H4. Adverbial clauses
30. causative adverbial subordinators, 31. concessive adverbial subordinators, 32. conditional adverbial subordinators, 33. other adverbial subordinators
I. Prepositional phrases, adjectives, and adverbs
34. total prepositional phrases, 35. attributive adjectives, 36. predicative adjectives, 37. total adverbs (except conjuncts, hedges, emphatics, discourse particles, downtoners, amplifiers)
J. Lexical classes
38. type/token ratio, 39. word length, 40. conjuncts, 41. downtoners, 42. hedges, 43. amplifiers, 44. emphatics, 45. discourse particles
K. Modals
46. possibility modals, 47. necessity modals, 48. predictive modals

L. Specialized verb classes
49. public verbs, 50. private verbs, 51. suasive verbs, 52. *seem* and *appear*
M. Reduced forms and dispreferred structures
53. contractions, 54. stranded prepositions, 55. split infinitives, 56. split auxiliaries
N. Coordination
57. phrasal coordination, 58. independent clause coordination (clause initial *and*)
O. Negation
59. synthetic negation, 60. analytic negation
P. Fluency
61. tokens, 62. Types, 63. mean length of sentences

For predicting learners' oral proficiency levels, we annotated the NICT JLE Corpus with the TreeTagger (Schmid, 1994) and counted the frequencies of 63 linguistic features with the Perl program. The program was originally developed by Murakami (2009) and modified by us for more accurate processing of L2 performance data. Then, the frequency matrix of 63 linguistic features was employed for the machine learning method of random forests (Breiman, 2001).

4.3 Random forests

Random forests is defined as an ensemble learning method that operates by constructing a large collection of decision trees. The decision tree technique is one of the most intuitive and popular machine learning techniques, which visualizes a sequence of data classification in the form of a flowchart-like diagram (Breiman, Friedman, Olshen, & Stone, 1984). Figure 1 shows a simple example of the correspondence relation between a scatter plot and the decision tree figure. It is based on the frequencies of two types of words X and Y in three types of texts represented as circle, square, and triangle. In this example, texts represented as circle and those as square and triangle can be distinguished by the frequencies of word Y (Split line 1). Then, squares and triangles can be classified through the frequencies of word X (Split line 2). By using the decision tree technique, the optimal discriminate thresholds (i.e., Split line 1 and Split line 2) can be mathematically identified, and each threshold can be visualized as branches in a decision tree figure (i.e., Branching 1 and Branching 2).

In the random forests model, the ensemble of decision trees (the forest) is generated using the ensemble learning technique to obtain better predictive performance than what can be possibly obtained from any of the constituent decision tree models (see Figure 2). The bagging ensemble learning algorithm (Breiman, 1994) is widely used for combining multiple decision tree models. It generates a number of datasets by using bootstrap sampling technique, and then constructs multiple classification models (i.e., Model 1, Model 2, and Model *n*) that are based on each bootstrap sample.

Following these steps, all results of each model are combined using the majority vote in order to make a final prediction.

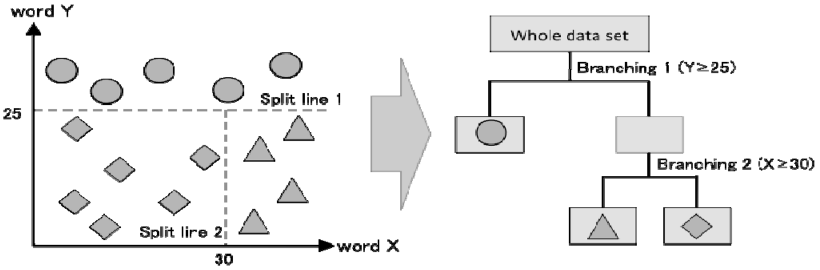

Figure 1. Image of the decision tree technique

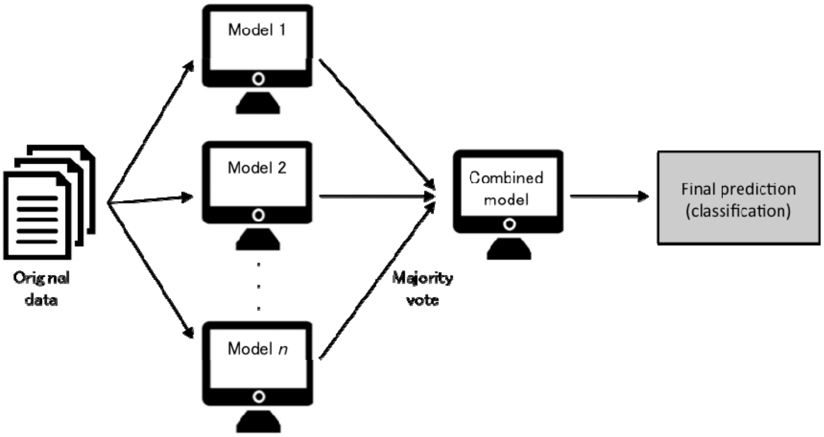

Figure 2. Image of the ensemble learning technique

By combining the decision tree and the bagging ensemble learning techniques, the random forests method generally achieves higher levels of accuracy than other machine learning techniques, such as *k*-nearest neighbors or support vector machine (Jin & Murakami, 2007). It can also handle thousands of explanatory variables statistically in an efficient manner, and provide reliable estimates of what variables are important for the classification (Tabata, 2012). Furthermore, it can highly estimate the accuracy of the classification through the out-of-bag (OOB) error estimation technique in a precise way, using approximately one-third of the data for checking the validity of the classification model constructed using the rest of the data (Breiman & Cutler, n.d.). Because of these advantages, it is regarded as a powerful method for text classification and feature extraction in the field of data mining.

5 Results and Discussion

5.1 Accuracy of automated scoring

Random forests was conducted to examine the level of accuracy with which the automated scoring would assess the L2 spoken English in this study. The results of scoring are shown in Table 4. In this table, columns of the matrix represent the SST levels predicted by random forests while rows represent the actual levels of learners. Thus, the column "Accuracy" shows the agreement rates between predicted levels and actual SST levels.

Table 4. The Result of the Prediction of Proficiency Levels of L2 Spoken English

Level	1	2	3	4	5	6	7	8	9	Accuracy
1	**0**	3	0	0	0	0	0	0	0	0.00%
2	0	**21**	14	0	0	0	0	0	0	60.00%
3	0	2	**145**	74	1	0	0	0	0	65.32%
4	0	0	33	**407**	40	2	0	0	0	84.44%
5	0	0	0	102	**119**	14	1	0	0	50.42%
6	0	0	0	19	61	**39**	5	5	1	30.00%
7	0	0	0	1	21	24	**21**	9	1	27.27%
8	0	0	0	0	12	14	22	**4**	4	7.14%
9	0	0	0	0	1	5	14	6	**14**	35.00%
Total accuracy rate										60.11%

Notes. The results of random forests were validated with OOB error estimation technique.

As shown in Table 4, the accuracy rate of random forests was 60.11% in total. This accuracy rate is 22.48 higher than the baseline accuracy rate of the simplest possible algorithm of always choosing the most frequent category, which is the proportion of level 4 learners in the present study (37.63%). The Pearson product-moment correlation coefficient with human scoring was 0.85. The correlation score of our random forests model compares favorably with those of major existing automated essay scoring systems such as e-rater (0.75~0.86) and IntelliMetric (0.72~0.84), summarized in Keith (2003).

The results also indicate the effectiveness of the linguistic features used in Biber (1988) for the assessment of L2 spoken English. As has already been mentioned, the bottle-neck of the current automated speech scoring systems is the low accuracy of L2 speech recognition, and consequently the technical limitation has prevented researchers from the comparisons of useful explanatory variables for automated speech scoring systems. However, the present study succeeded in revealing an effective set of variables by investigating the transcribed speeches annotated with learners' oral proficiency levels.

5.2 Variable importance

As a next step, we identified effective linguistic features for predicting the oral proficiency of learners by checking the variable importance, which was calculated using the Gini index measure. Figure 3 shows the variable importance of the top 10 linguistic features that have a strong relationship with the test scores assessed by professional human raters.

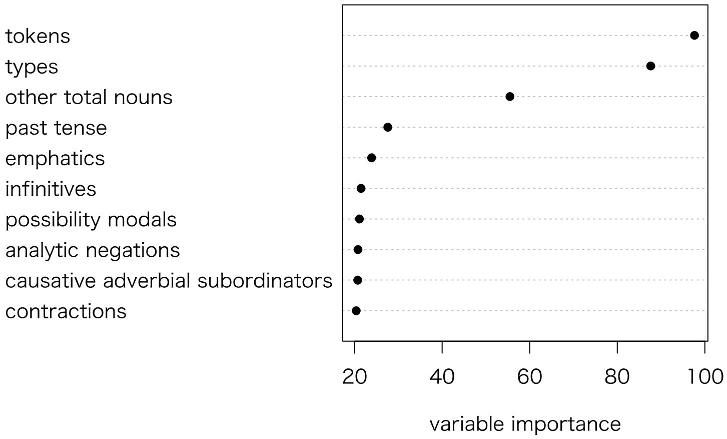

Figure 3. Variable importance of the top 10 linguistic features

The variable importance scores demonstrated that the category of tokens is the most important predictor, followed by types and nouns. As is clear from the scores shown in Figure 3, these three linguistic features are far more effective than other linguistic features for the prediction of learners' oral proficiency. Therefore, the relationship of these linguistic features and oral proficiency levels should be examined in detail.

The box plots in Figures 4 and 5 show the variations of the tokens and types across nine different proficiency levels. They indicate that these linguistic features are positively correlated to the proficiency levels, and that human raters of SST may have put emphasis on these variables when scoring learners' performance, given the variable importance scores shown in Figure 3. Automated scoring is often criticized for the "brute-empirical approach," which places emphasis on the formal aspects of learners' performance. However, as the results of the present study indicate, human raters also measure learners' performance from the standpoint of quantitative aspects, such as the number and types of words, similar to the process of machine scoring.

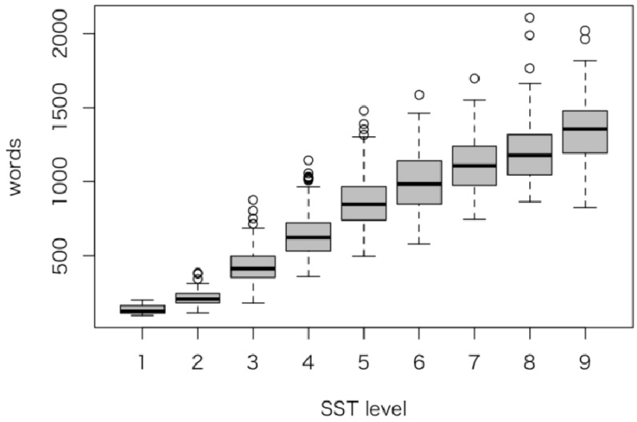

Figure 4. Box plots of the tokens

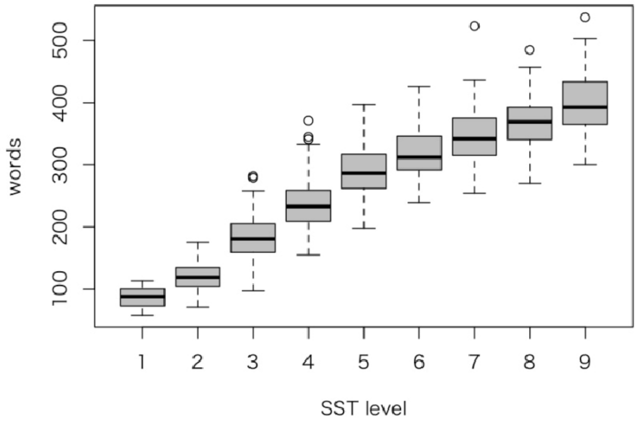

Figure 5. Box plots of the types

The third important variable category is that of the nouns. Interestingly, the frequency of nouns is negatively correlated to the proficiency levels. The vertical axis of the box plot in Figure 6 represents the relative frequency per 100 words. The box plot of level 1 learners is located around 60, which suggests that around 60% of spoken production consists of nouns.

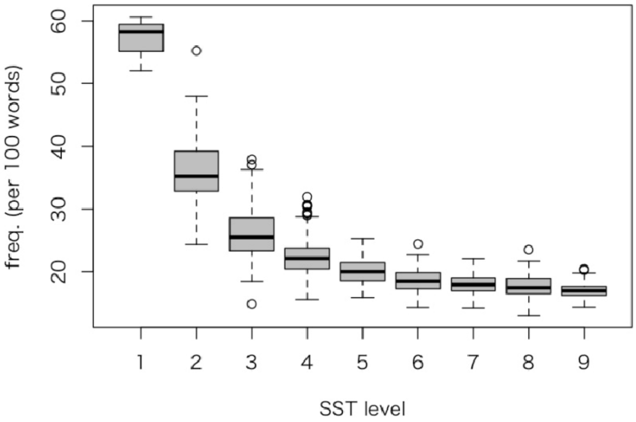

Figure 6. Box plots of other total nouns

Novice learners frequently use stand-alone nouns in their utterances, as shown in the following examples.

> **Interviewer:** You come by train or bus?
> **Test-taker:** Er, train.
> **Interviewer:** OK. Train. OK, in *XXX04,* do you live with your family?
> **Test-taker:** Yes.
> **Interviewer:** OK. Please tell me about your family.
> **Test-taker:** Er, wife and children *ka*.

The lack of sentence structure in novice learners' speech can be qualitatively assessed by the ratio of nouns in their utterance, which offers suggestions for developing an automated L2 speech scoring system.

Other linguistic features listed in Figure 3 are also linguistically and pedagogically significant. The fourth important variable is the past tense category. As Ishikawa (2008) pointed out, lower-level learners mainly speak in the present tense, whereas higher-level learners speak more naturally using a combination of present and past tense. It is known that the frequency of past tense can be a developmental index in the written performance of English learners as well as in their spoken performance (Kobayashi, 2013). The fifth and seventh important variables are emphatics and possibility modals respectively, which are positively correlated to the proficiency levels. These linguistic features that contribute to the indirectness and directness of utterances represent a major "rhetorical gap" that language learners have to cross before they can gain membership in a discourse community (Hyland, 1995). The sixth important variable is the category of infinitives, which are likely to be related to the development of verb phrase structure. As the verb phrase structure in the learners' performance develops, the frequency of infinitives increases. Thus, it can serve as a developmental index that can distinguish the performance of novice learners from those of intermediate and advanced learners (Kimura, Tanaka, & Tomiura, 2005). Finally, the remaining three variables ranked in the top 10 in Figure 3, (a) analytic negations, (b) causative adverbial subordinators, and (c) contractions, are characteristic of spoken language (Biber, 1988; Biber, Johansson, Leech, Conrad, & Finegan, 1999). As the proficiency level rises, learners gradually acquire the natural use of spoken language.

From the viewpoint of language teaching and learning, automated scoring systems should provide not only test scores but also pedagogical feedback for learners. Most of the current automated speech feedback applications have focused on learners' pronunciation (e.g., Franco et al,, 2010; Pelton, 2012; Stanley, Hacioglu, & Pellom, 2011), and there are few applications which can offer lexical, grammatical, and discourse feedback. In order to address this problem, various types of speech characteristics that distinguish highly advanced learners from less-advanced learners should be captured through learner corpus analyses (Higgins, Ramineni, & Zechner, 2015). In this respect, the results of variable importance in the present study can be quite useful for developing better scoring and feedback systems. Moreover, it contributes to our understanding of human scoring processes from an evidentiary perspective, since automated scoring algorithms can estimate the weightings of linguistic features used in the human assessment.

6 Conclusions and Future Work

We predicted the oral proficiency levels of 1,281 L2 learners of English with the use of random forests. The criterion variables were the SST levels scored by human raters and the explanatory variables consisted of the 63 linguistic features. As a result, the total accuracy rate of random forests was 60.11%, and the correlation coefficient between the predicted scores and human scoring was 0.85. Predictors that can clearly discriminate oral proficiency levels were tokens, types, and nouns. The results of this study can be applied for creating assessments that are more appropriate for scaling the oral performances of EFL learners.

In our future study, other explanatory variables, such as articles, which are considered to be the most difficult features to learn for Japanese EFL learners, n-grams, and errors of spoken production should be investigated in order to cover the broader aspects of learners' performance. In particular, error analysis can be useful for automated speech scoring (Kobayashi, 2014) as well as for grammatical feedback to learners (Gamon, Chodorow, Leacock, & Tetreault, 2013). Additionally, other criterion variables, such as the TOEFL score or the TOEIC score, should be examined in order to compare the speaking performance of learners with their listening, reading, and writing abilities.

References

Abe, M. (2014). Frequency change patterns across proficiency levels in Japanese EFL learner speech. *Journal of Applied Language Studies, 8*(3), 85-96.

Abe, M., Kobayashi, Y., & Narita, M. (2013). Using multivariate statistical techniques to analyze the writing of East Asian learners of English. *Learner Corpus Studies in Asia and the World, 1,* 55-65.

ACTFL-ALC Press (1996). *Standard speaking test manual.* Tokyo: ALC Press.

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 181-198). New York: Routledge.

Bejar, I. I., Williamson, D. M., & Mislevy, R. J. (2006). Human scoring. In D.M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 49-81). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bhat, S., & Yoon, S. (2015). Automatic assessment of syntactic complexity for spontaneous speech scoring. *Speech Communication, 67,* 42-57.

Biber, D. (1988). *Variation across speech and writing.* Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* Harlow: Longman.

Breiman, L. (1994). Bagging predictors. *Machine Learning, 24*(2), 123-140.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-23.

Breiman, L., & Cutler, A. (n.d.). Random forests. Retrieved from http://www.stat.berkeley.edu/~breiman/RandomForests/

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees.* Boca Raton. FL: Chapman and Hall.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1998). Automated scoring using hybrid feature identification technique. *Proceedings of the 17th International Conference of Computational Linguistics, 1*, 206-210.

Chen, M., & Zechner, K. (2011). Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics,* 722-731.

Conrad, S., & Biber, D. (Eds.) (2001). *Variation in English: Multi-dimensional studies.* Harlow: Longman.

Downey, R., Farhady, H., Present-Thomas, R., Suzuki, M., & Van Moere, A. (2008). Evaluation of the usefulness of the Versant for English Test: A response. *Language Assessment Quarterly, 5*(2), 160-167.

Elliot, S. (2003). IntelliMetric: From here to validity. In M. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71-86). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ericsson, P. F. (2006). The meaning of meaning: Is a paragraph more than an equation? In P. F. Ericsson & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 28-37). Logan, UT: Utah State University Press.

Franco, H., Bratt, H., Rossier, R, Gadde, V. R., Shriberg, E., Abrash, V., & Precoda, K. (2010). EduSpeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing, 27*(3), 401-418.

Frignal, E. (2013). Twenty-five years of Biber's multi-dimensional analysis: Introduction to the special issue and an interview with Douglas Biber. *Corpora, 8*(2), 137-152.

Gamon, M., Chodorow, M., Leacock, C., & Tetreault, J. (2013). Grammatical error detection in automatic essay scoring and feedback. In M. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 251-266). New York: Routledge.

Hawkins, J. A., & Filipović, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the Common European Framework.* Cambridge: Cambridge University Press.

Higgins, D., Ramineni, C., & Zechner, K. (2015). Learner corpora and automated scoring. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 587-604). Cambridge: Cambridge University Press.

Hyland, K. (1995). The author in the text: Hedging scientific writing. *Hong Kong Papers in Linguistics and Language Teaching, 18,* 33-42.

Ishikawa, S. (2008). Proficiency and speech vocabulary: A study on the NICT-JLE Corpus. In G. Weir & T. Ozasa (Eds.), *Studies in language and text analysis* (pp. 11-20). Glasgow: University of Strathclyde Publishing.

Izumi, E., Uchimoto, K., & Isahara, H. (2004). *A speaking corpus of 1200 Japanese learners of English.* Tokyo: ALC Press.

Jin, M., & Murakami, M. (2007). Authorship identification using random forests. *Proceedings of Institute of Statistical Mathematics, 55*(2), 255-268.

Keith, T. Z. (2003). Validity of automated essay scoring systems. In M. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-*

disciplinary perspective (pp. 147-167). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Kimura, M., Tanaka, S., & Tomiura, Y. (2005). Tracing Japanese EFL learners' development in productive vocabulary: The NICT JLE Corpus-based analysis for seeking lexical characteristics. *The Proceedings of the NICT JLE Corpus Symposium,* 54-71.

Kobayashi, Y. (2013). A comparison of spoken and written learner corpora: analyzing developmental patterns of vocabulary used by Japanese EFL learners. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Twenty years of learner corpus research: Looking back, moving ahead* (pp. 277-287). Louvain-la-Neuve: Presses universitaires de Louvain.

Kobayashi, Y. (2014). Computer-aided error analysis of L2 spoken English: A data mining approach. *Proceedings of the Conference on Language and Technology 2014,* 127-134.

Kobayashi, Y., & Abe, M. (2014). A machine learning approach to the effects of writing task prompts. *Learner Corpus Studies in Asia and the World, 2,* 163-175.

Landauer, T. K., Laham, D., & Foltz, P. W. (2000). The Intelligent Essay Assessor. *System, 15(5),* 27-31.

Larkey, L. S., & Croft, W. B. (2003). A text categorization approach to automated essay grading. In M. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 55-70). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lewis, M. (1999). *Oral proficiency self-study course for beginners.* Tokyo: ALC Press.

McGee, T. (2006). Taking a spin on the Intelligent Essay Assessor. In P. F. Ericsson & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 79-92). Logan, UT: Utah State University Press.

Meunier, F., & Littre, D. (2013). Tracking learners' progress: Adopting a dual 'corpus cum experimental data' approach. *The Modern Language Journal, 97*(S1), 61-76.

Murakami, A. (2009). *A corpus-based study of English textbooks in Japan and Asian countries: Multidimensional approach* (Unpublished master's thesis). Tokyo: Tokyo University of Foreign Studies.

Page, E. B. (1994). New computer grading of student prose, using modern concepts and software. *Journal of Experimental Education, 62*(2), 127-142.

Page, E. B. (2003). Project Essay Grade: PEG. In M. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43-54). Hillsdale: Lawrence Erlbaum Associates.

Pelton, T. (2012). Mining pronunciation data for consonant cluster problems. O. Engwall (Ed.), *Proceedings of International Symposium on Automatic Detection of Errors in Pronunciation Training* (pp.31-36).

Stockholm, Sweden: KTH, Computer Science and Communication Department of Speech, Music and Hearing.

Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment, 1*(2), 3-21.

Sardinha, T. B., & Pinto, M. V. (Ed.) (2014). *Multi-dimensional analysis, 25 years on: A tribute to Douglas Biber.* Amsterdam: John Benjamins.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing, 12*(4), 44-49.

Shermis, M. D., & Burstein, J. C. (Ed.) (2003). *Automated essay scoring: A cross-disciplinary perspective.* New York: Routledge.

Shermis, M. D., Burstein, J. C., & Bursky, S. A. (2013). Introduction to automated essay evaluation. In M. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 1-15). New York: Routledge.

Shin, D., Min, H., Park, S., Jung, C. K., Joo, H., & Kim, M. (2013, July). *Validation research for developing and applying the automated scoring program for the speaking section of the NEAT.* Paper presented at the 35th Annual Language Testing Research Colloquium.

Stanley, T., Hacioglu, K., & Pellom, B. (2011, August). Statistical machine translation framework for modeling phonological errors in computer assisted pronunciation training system. *Proceedings of International Speech Communication Association Special Interest Group Workshop on Speech and Language Technology in Education,* 125-128.

Tabata, T. (2012). 'Key' words and stylistic 'signatures': Textometry and random forests. In T. Tabata (Ed.), *Mining textual patterns* (pp. 45-64). Tokyo: Institute of Statistical Mathematics.

Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal, 97*(S1), 77-101.

Williamson, D. M. (2013). Developing warrants for automated scoring of essays. In M. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 153-180). New York: Routledge.

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing, 27*(3), 291-300.

Xie, S., Evanini, K., Zechner, K., & Williamson, D. M. (2012). Exploring content features for automated speech scoring. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 103-111.

Yoon, S., Bhat, S., & Zechner, K. (2012). Vocabulary profile as a measure of vocabulary sophistication. *Proceedings of the 7th Workshop on*

Innovative Use of NLP for Building Educational Applications, 180-189.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication, 51*(10), 883-895.

Yuichiro Kobayashi
Faculty of Sociology, Toyo University
5-28-20, Hakusan, Bunkyo-ku, Tokyo
112-8606, Japan
Tel: 03-3945-8660
E-mail: kobayashi077@toyo.jp

Mariko Abe
Faculty of Science and Engineering, Chuo University
1-13-27, Kasuga, Bunkyo-ku, Tokyo
112-8551, Japan
Tel: 03-3817-1958
E-mail: abe.127@g.chuo-u.ac.jp

Received: March 21, 2016
Revised: June 19, 2016
Accepted: June 30.2016