Information Engineering Express
International Institute of Applied Informatics
2015, Vol.1, No.3, 29 – 38

The Relationship of English Foreign Language Learner Proficiency and an Entropy Based Measure

Brendan Flanagan [*], Sachio Hirokawa [†]

Abstract

It is important for education systems to analyze and provide an appropriate level of feedback to meet the needs of learners. Predicting a learner's proficiency level can be used to inform learner's about their progress, and can also aid other parts of the characteristic analysis and feedback process, such as: focused analysis on learner proficiency subgroups. In this paper, we propose a measure based on the frequency of words in the sentences produced by learners during speaking exams to predict the learner's language proficiency. The proposed measure is compared to the learner's vocabulary size by correlation analysis. The results suggest that there is a stronger correlation between the proposed measure and the proficiency of the learner than the learner's vocabulary size.

Keywords: Learner Proficiency, Proficiency Prediction, Speaking Errors, Entropy.

1 Introduction

Foreign language learners at different levels of proficiency are faced with different needs and problems. It is important to provide appropriate support and feedback that matches these needs. In a traditional classroom environment a teacher would estimate the progress and proficiency of the learner and provide suitable support. However, as language learning increases due to globalization and the use of the Internet as a multi-national multi-lingual platform, the demand for language teaching outpaces the supply and availability of such services. The prediction of a learner's language proficiency level could be used to provide automated feedback so the learner may understand his or her own progress.

In previous research, we have investigated the automatic prediction of foreign language writing errors on a corpus collected from a language learning SNS [1]. However, as the proficiency level of learners on these SNS is often broad, which makes it difficult to predict errors, as a machine classifier has to deal with a wide range of writing complexity, which increases the chance of false positive classification. This problem serves as our main motivation in the study, as opposed to automatically determining the official score of a proficiency test. We hope to use the prediction of a learner's foreign language characteristics to provide tailored tools that can focus

[*] Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka, Japan
[†] Research Institute for Information Technology, Kyushu University, Fukuoka, Japan

on the particular errors, support, and feedback needs of learners at specific language proficiency levels. Therefore, the proficiency prediction method should not rely on other features, such as error prediction, that could be adversely affected by the proficiency of the learner.

In this paper, we propose a measure based on the entropy of word occurrences in the sentences of learner discourse during a speaking exam. The rationale behind this is that the discourse of an elementary learner can be thought of as having limited available words selections due to restricted vocabulary, and less variation in word use due to only knowing a small amount of grammar patterns in the target language when compared with intermediate or advance learners. The measure is then compared to the vocabulary size and also the entropy of word occurrences at the learner level, to examine which has a stronger correlation with the learner's language proficiency.

2 Related Work

The origins of automated language scoring can be traced back to Page in 1968 [2], who proposed that it was feasible to score essays using a computer. This has spawned a number of different goals and approaches ranging from the automatic scoring of written essays and language tests, to oral discourse assessment. Previous research on the prediction of foreign language proficiency has focused on a number of different approaches, including: errors, fluency of discourse, grammatical, lexical, and syntactical complexity.

Supnithi et al. [3], analyzed the vocabulary, grammatical accuracy and fluency features of learners in speaking exams. These features were then used to train Support Vector Machine and Maximum Entropy machine-learning algorithms to automatically predict the proficiency level of the learner. Vocabulary features, such as: bi-grams, words expressed by both the examiner and learner, words only expressed by the learner, and words from a list of twelve different levels of proficiency. A maximum prediction accuracy of 65.57% was achieved using an SVM classifier. In the present paper, we analyzed the same corpus, and propose a different measure that could be used to simplify the prediction of learner proficiency.

There has also been research into commercial proficiency scoring systems to give learners quick feedback for exams. Chen et al. [4], created a corpus based on the TOEFL Practice Test Online annotated with structural events, such as: clause boundaries and disfluency. They then extracted features based on words and structural events, and it was found that disfluency had a higher correlation with human scorers than syntactic complexity features. A combination of these features was used to further improve the disfluency correlation. Chen and Zechner [5], examined syntactic complexity features that are related to the oral proficiency of language learners with the goal of creating automatic scoring models that correlate well with human scorers. Three multiple regression models were built with the best model made from 17 syntactic features were extracted and had a significant correlation of 0.49 with human scorers. Higgins et al. [6], created a system called SpeechRater^{SM} for the internet-delivered TOEFL oral test, which processes responses in three stages: filtering, scoring, and aggregation. In the scoring stage, features such as: fluency, pronunciation, vocabulary diversity, and grammar were examined to estimate the proficiency score. The features for vocabulary diversity were based on unique word counts that were normalized by total word duration and speech duration. The results found that there is a correlation of 0.7 between the scores generated by the system and human scorers. Zechner et al. [7], analyzed 1,400 speaking tests using automatic speech recognition and feature extraction for fluency, pronunciation, prosody, and grammatical accuracy. Different linear regression models were built for each of the 21 speaking items in the test and

were used to predict the proficiency level of the learner. Their system achieved a correlation of 0.73 with human rater scores. In this paper, we analyze a corpus of transcribed speaking tests without extracting features concerning the production of utterances, and propose a measure for the prediction of learner proficiency.

Crossley et al. [8], examined the importance of different lexical features that could be analyzed to create a model of learner proficiency. Human raters based on standardized lexical criteria evaluated a corpus of 240 foreign language writings. It was reported that lexical diversity, word hypernymy values and content word frequency accounted for around 44% of the variance in the lexical proficiency evaluations. In further research, Crossley and McNamara [9] further developed their model of predicting learner proficiency by incorporating features relating to cohesion and linguistic sophistication. They argue that learners with high proficiency don't necessarily produce writing with more cohesion, but instead use less frequent and familiar words to increase lexical diversity.

Other research has analyzed learner corpora to extract features that identify characteristics of certain proficiency levels. Yoon et al. [10], investigated the distribution of syntactical patterns in the form of parts of speech (POS). A large learner corpus that had been classified into four different levels of proficiency was parsed to extract POS tags, which were then indexed to create vector space models. The cosine similarity of the test vectors and corpus vectors were then compared. The proficiency prediction was based on the proficiency of the most similar corpus vector. Abe [11], examined the extraction of 58 different linguistic features by frequency, correspondence, and cluster analysis across different oral proficiency groups. It was found that there are patterns of feature frequencies that rise, fall or are flat across proficiency levels. It was suggested that these features could be used to determine how learner languages change across different levels of proficiency. In the present paper, we propose that an entropy based measure has a stronger correlation to proficiency than analysis by simple word frequency.

3 Data

The data analyzed in this paper is based on a collection of recorded oral proficiency interview exams conducted as a part of the ACTFL English Standard Speaking Test (SST) [12]. This corpus is commonly known as the National Institute of Information and Communications Technology Japanese Learner English (NICT-JLE) Corpus and is made up of transcripts from 15 minute speaking exams. There are nine different proficiency levels in the SST exam, with level 1-3 representing elementary proficiency, level 4-8 as intermediate, and level 9 representing learners who have advanced proficiency. Professional examiners determined the SST proficiency level grade for each exam. This provides a reliable insight into the proficiency level of the learner, as opposed to other corpora that rely on learner experience, such as: length of study [13]. The corpus is split into two main sets of tagged data: learner original, and learner error tagged transcriptions. Error tagged transcripts of the same learner were also included in the learner original dataset, and we removed duplicates across the two datasets. A total of 1114 original learner transcripts were analyzed to build an index upon which a special purpose search engine was constructed using GETA‡. The transcripts are marked up with a custom tag set that includes non-lexical tags associated with discourse events such as: long pauses, non-verbal

‡ http://geta.ex.nii.ac.jp/

sounds, etc. The transcripts also contain the dialog spoken by the interviewer in the exam. The information provided by these tags was not used for analysis in this paper. The transcripts were preprocessed to remove non-lexical information and dialog by the interviewer. Each of the learners utterances were indexed as individual documents within the search engine, and tagged with the SST proficiency level as provided in the header of the transcripts.

4 Correlation between Proficiency and Learner transcript characteristics

4.1 Baseline: Vocabulary Size

In this paper, the vocabulary size of a learner's exam transcript will be analyzed as a baseline for comparison with our proposed method. The vocabulary size is calculated as the number of distinct words contained in a single learner's transcript and does not take into account the word occurrence frequency. The formula in Equation 1 was used to calculate the vocabulary size for each learner.

$$V(u_i) = \sum_{j \in W} \#\{w_j \mid tf(w_j, u_i) > 0\} \quad (1)$$

Where u_i represents learner i, W is the set of all words contained within the corpus, and $tf(w_j, u_i)$ is the occurrence frequency of the word w_j in the exam transcript of learner u_i.

4.2 An Entropy like measure of Language Learner Transcripts

In 1948, Shannon [14] introduced the theory of information entropy to determine the expected amount of information contained in an event. In this paper, we propose that a measure based on the entropy of learner transcripts can be used in the analysis of learner proficiency. We propose that the information entropy formula in Equation 2 can be used to calculate the information in the transcript of a learner's exam.

$$E(u_i) = -\sum_{j \in W} P(w_j, u_i)\, log\, P(w_j, u_i) \quad (2)$$

Where $E(u_i)$ is the information entropy of u_i which represents the i^{th} learner, W is the set of all words, and w_j represents a word contained within the corpus. In Shannon's theory, $P(w_j, u_i)$ is the probability of occurrence of the word w_j occurring in the exam transcript of the learner u_i.

$$P(w_j, u_i) = \frac{tf(w_j, u_i)}{\sum_{k \in U} tf(w_k, u_i)} \quad (3)$$

The formula in Equation 3 would usually be used to calculate this probability, where $tf(w_j, u_i)$ represents the occurrence frequency of the word w_k in the transcript of learner u_i.

We propose an alternate formula as seen in Equation 4 for the calculation of this term. It is based on the frequency of sentences in a learner's transcript in which a word occurs.

$$P(w_j, u_i) = \frac{df(w_j, u_i)}{|U|} \tag{4}$$

Where $df(w_j, u_i)$ is the number of sentences in the transcript of learner u_i which contain the word w_j, and U represents the set of all learner transcripts in the corpus.

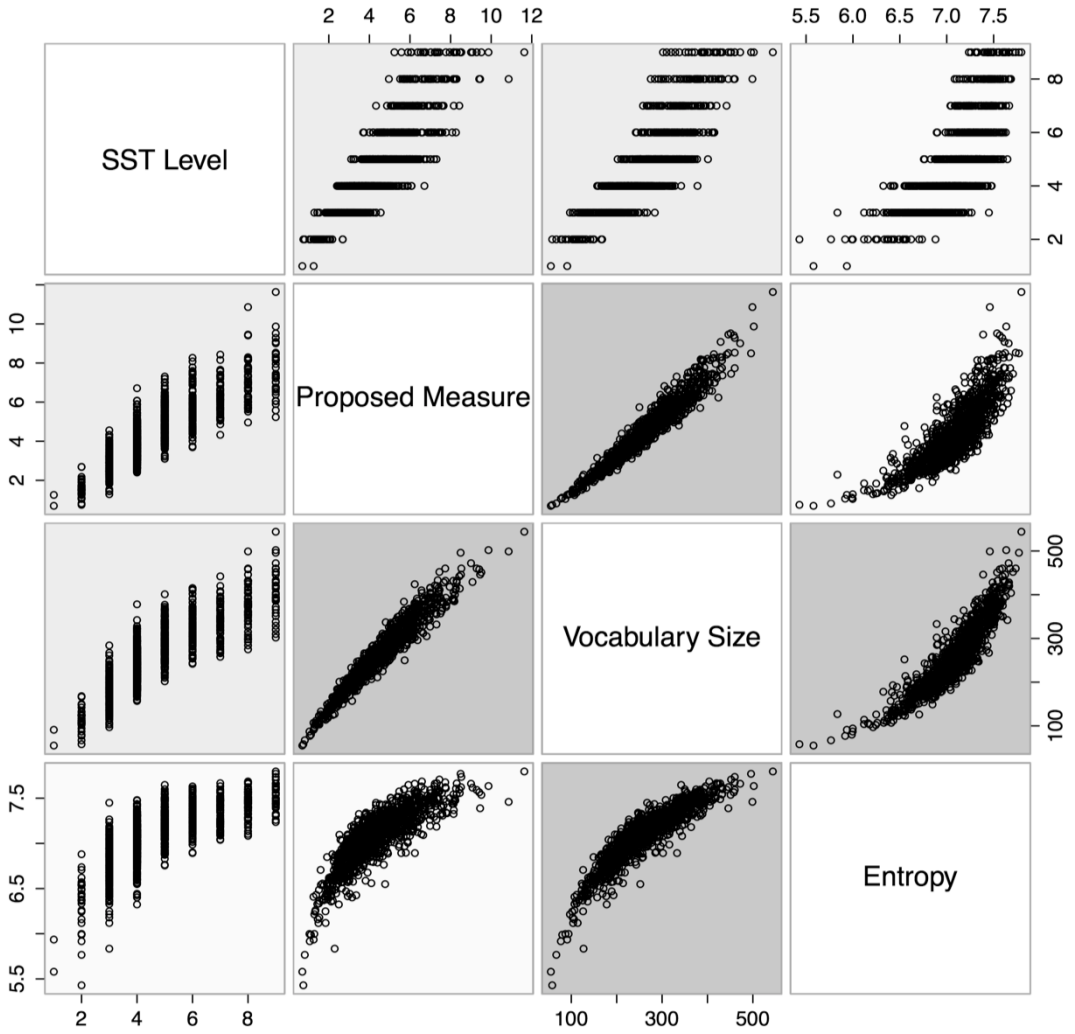

Figure 1: Correlation scatter plot matrix of proficiency level, proposed measure, vocabulary size, and entropy of the discourse of each learner.

Scatter plots of all three measures versus the learner proficiency (SST level) are shown in a correlation scatter plot matrix in Figure 1. The variables of the matrix are ordered so that strong correlations are closer to each other on the principle diagonal axis. The scatter plot for the relation between the proficiency and entropy suggests that it is broad when compare to the other measures. Compared to entropy, the relation between the proposed measure and SST learner proficiency is narrow, suggesting that it is a better fit to predicting proficiency. The vocabulary size of a learner's transcript increases steadily as proficiency rises until around SST level 6, at

which point the vocabulary increases at a diminished rate. This would suggest that vocabulary size is a strong determiner of proficiency from elementary to intermediate levels. However, at higher proficiency level the use of similar size vocabularies might have an affect on the perceived proficiency level scored.

We examined the differences in word usage for learners with SST levels from 4 to 9 by analyzing the corpus using a part of speech parser, TreeTagger [15], to divide the vocabulary into subsets. The vocabulary size and proposed measure was calculated for each of these subsets. These were then analyzed to determine the strength of the correlation between the POS subsets and SST proficiency. A correlation scatter plots matrix of learner proficiency versus the vocabulary size of the top 9 POS subsets are shown in Figure 2. It should be noted that the granularity is course because the total count of some POS subsets is small, and therefore it increases the possibility that multiple results occur in the same position in the graph.

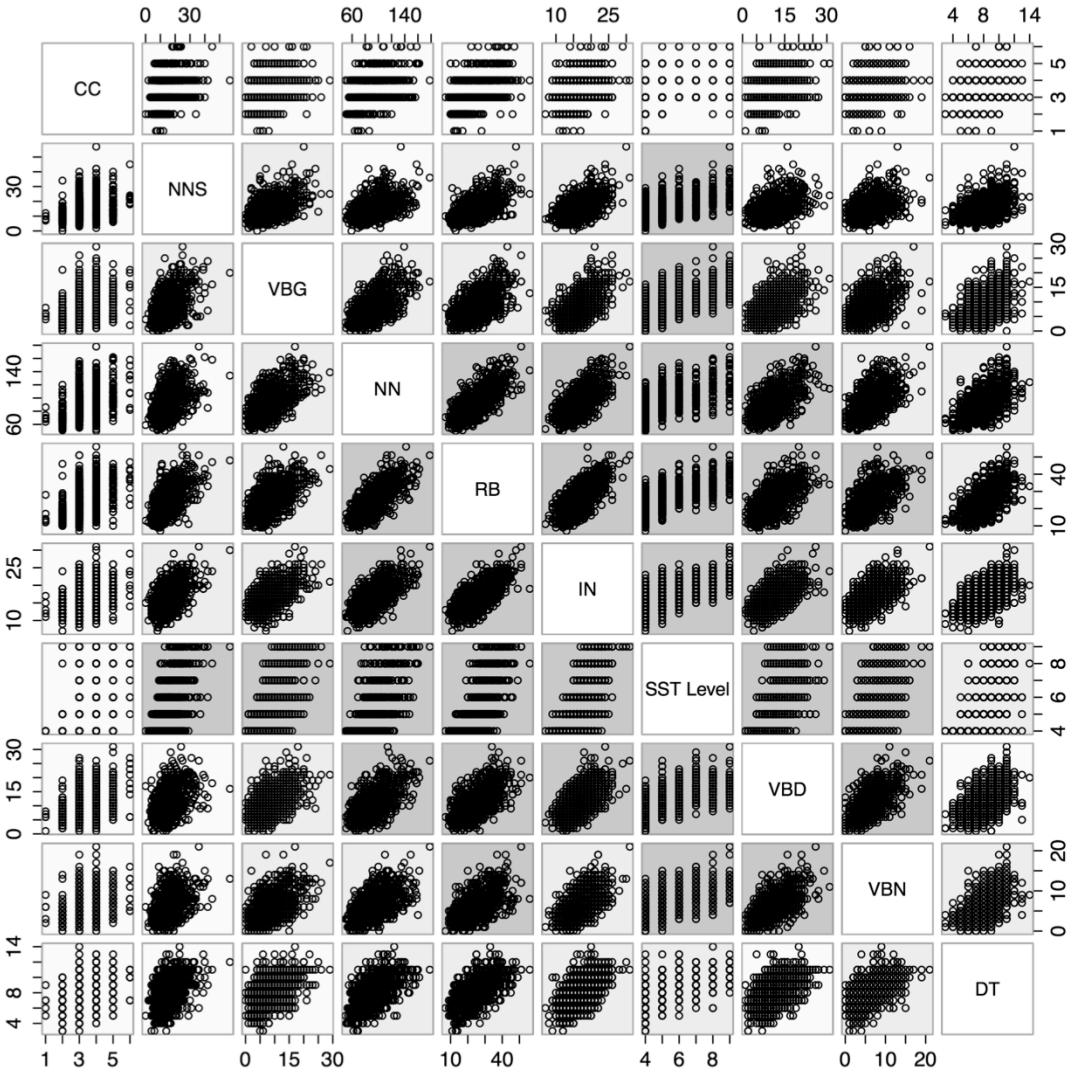

Figure 2: Correlation scatter plot matrix of Learner proficiency versus vocabulary size for top 9 POS subsets.

In Figure 3, a correlation scatterplot matrix of our proposed measure versus SST level shows a finer level of granularity when compared to the vocabulary size plots.

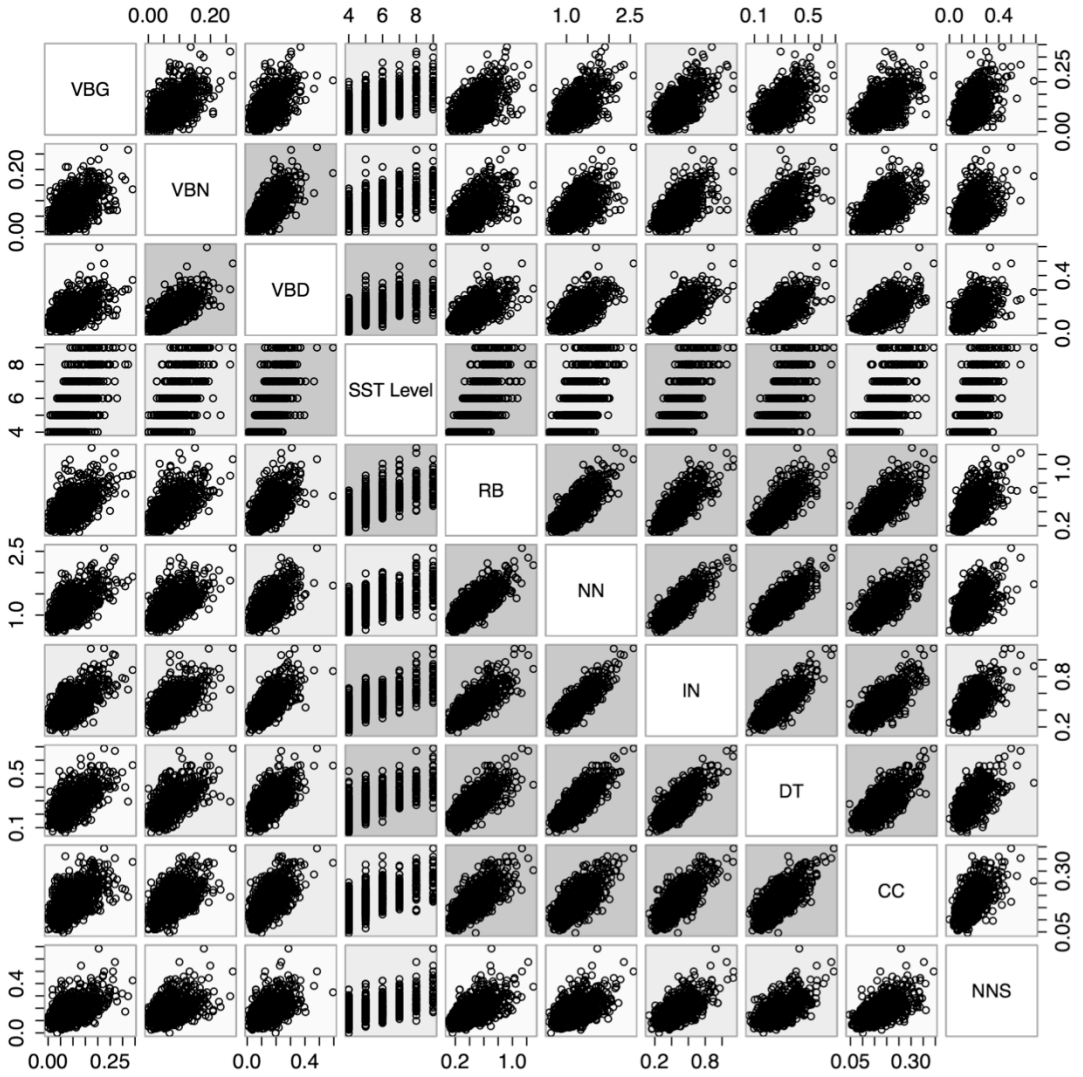

Figure 3: Correlation scatter plot matrix of Learner proficiency versus the proposed measure for top 9 POS subsets.

4.3 Correlation Analysis

The Pearson product-moment correlation coefficient can be used to measure the linear correlation between two variables. In this section, the correlation between learner proficiency and vocabulary size, entropy, and our proposed measure are compared.

Table 1: Pearson correlation coefficient *r* for Entropy, Proposed measure, and Vocabulary size of learner speaking exams.

SST Level	Entropy	Proposed Measure	Vocabulary Size	N
1 – 9	0.8021619	0.843445	0.840757	1114

4 – 9	0.7464927	0.7816775	0.7810395	890

In Table 1, the correlation coefficient *r* of all SST proficiency levels is higher than that of the intermediate and advanced levels. This is most likely due to greater variation in word usage rather than vocabulary at higher proficiency levels.

Table 2: Pearson correlation coefficient *r* for each SST Level vs Vocab Size/Entropy of the top 20 parts of speech.

Parts Of Speech	**Proposed Measure**	**Vocabulary Size**	**Non-zero Samples**
IN (preposition/subord. conj.)	0.7448775	0.6893652	890
DT (determiner)	0.7150026	0.5642325	890
RB (adverb)	0.7092389	0.7085002	890
VBD (verb be, past)	0.7017322	0.6591255	890
VBN (verb be, past participle)	0.6837663	0.6272495	877
CC (coordinating conjunction)	0.6794182	0.2868362	890
VBG (verb be, gerund/participle)	0.6569454	0.6510437	884
NN (noun, singular or mass)	0.6472380	0.6233473	890
NNS (noun plural)	0.6437256	0.6038592	889
VB (verb be, base form)	0.6252794	0.5643526	890
PP (personal pronoun)	0.6248875	0.4182833	890
VBP (verb be, pres non-3rd p.)	0.6139417	0.5422852	890
JJ (adjective)	0.5415856	0.5284353	890
TO (to)	0.5338780	*0.0193595*	890
MD (modal)	0.5242069	0.4547834	871
PP$ (possessive pronoun)	0.4955024	0.4486139	890
WP (wh-pronoun)	0.3967646	0.3573016	727
VBZ (verb be, pres, 3rd p. sing)	0.2494197	0.3847656	890
RBR (adverb, comparative)	0.3284726	0.2460167	502
FW (foreign word)	0.3265978	0.1129750	798

The Pearson correlation coefficient *r* for each of the relations is shown in Table 2. Both the proposed measure and vocabulary size data contained 890 sample pairs, except transcripts that did not contain a particular POS tag. All correlations are significant at $p < 0.01$, except for the vocabulary size of the part of speech "TO" which is written in *italic* text. The table is sorted by strongest correlation to weakest, with the strongest correlation for each part of speech bolded. The correlation between the proposed measure and the learner's proficiency is stronger than vocabulary size for the majority of parts of speech. This confirms that the proposed measure of transcripts is a stronger indicator of learner proficiency than vocabulary size for SST proficiency equal to or higher than level 4.

5 Conclusion and Future Work

In this paper, we proposed a measure based on the entropy of the sentence occurrence frequency of words in transcripts of English speaking proficiency exams. The proposed measure was compared with the vocabulary size and entropy of the same transcripts. It was found that the proposed measure has a stronger correlation with SST learner proficiency than both vocabulary size and entropy. The correlations were then compared on parts of speech subsets. It was found that the proposed measure has a stronger correlation with proficiency in a majority of subsets. In future work we will undertake a comparison of prediction with other speaking and writing learner corpora, and assess it usefulness in the enhancement of learner error detection.

Acknowledgement

This work was partially supported by JSPS KAKENHI Grant Number 15J04830.

References

[1] B. Flanagan, C. Yin, T. Suzuki, and S. Hirokawa, Classification and Clustering English Writing Errors Based on Native Language, Proc. 2014 IIAI 3rd International Conference on Advanced Applied Informatics (IIAIAAI), 2014, pp. 318-323.

[2] E. B. Page, The use of the computer in analyzing student essays, International Review of Education, vol. 14, no. 2, 1968, pp. 210-225.

[3] T. Supnithi, K. Uchimoto, T. Saiga, E. Izumi, S. Virach, and H. Isahara, Automatic proficiency level checking based on SST corpus, Proc. RANLP, 2003, pp. 29-33.

[4] L. Chen, J. Tetreault, and X. Xi, Towards using structural events to assess non-native speech, Proc. NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications, 2010, pp. 74-79.

[5] M. Chen, and K. Zechner, Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech, Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, 2011, pp. 722-731.

[6] D. Higgins, X. Xi, K. Zechner, and D. Williamson, A three-stage approach to the automated scoring of spontaneous spoken responses, Computer Speech & Language, vol. 25, no. 2, 2011, pp. 282-306.

[7] K. Zechner, K. Evanini, S. Y. Yoon, L. Davis, X. Wang, L. Chen, and C. W. Leong, Automated Scoring of Speaking Items in an Assessment for Teachers of English as a Foreign Language, ACL 2014, 2014, pp. 134-143.

[8] S. A. Crossley, T. Salsbury, D. S. McNamara, and S. Jarvis, Predicting lexical proficiency in language learner texts using computational indices, Language Testing, vol. 28, no. 4, 2011, pp. 561–580.

[9] S. A. Crossley, and D. S. McNamara, Predicting second language writing proficiency: the roles of cohesion and linguistic sophistication, Journal of Research in Reading, vol. 35, no. 2, 2012, pp. 115-135.

[10] S. Y. Yoon, and S. Bhat, Assessment of ESL learners' syntactic competence based on similarity measures, Proc. 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012, pp. 600-608.

[11] M. Abe, Frequency Change Patterns across Proficiency Levels in Japanese EFL Learner Speech, Apples: Journal of Applied Language Studies, vol. 8, no. 3, 2014, pp. 85-96.

[12] E. Izumi, K. Uchimoto, and H. Isahara, SST speech corpus of Japanese learners' English and automatic detection of learners' errors, ICAME Journal, vol. 28, 2004, pp. 31-48.

[13] Y. Tono, T. Kaneko, H. Isahara, T. Saiga, E. Izumi, and M. Narita, The Standard Speaking Test (SST) Corpus: A 1 million-word spoken corpus of Japanese learners of English and its implications for L2 lexicography, Proc. Second Asialex International Congress, 2001, pp. 257-262.

[14] C.E. Shannon, A Mathematical Theory of Communication, Bell system technical journal, vol. 27, no. 3, 1948, pp. 379-423.

[15] H. Schmid, Probabilistic part-of-speech tagging using decision trees, Proc. international conference on new methods in language processing 12, 1994, pp. 44-49.