L2 Learners' Proficiency Evaluation on CEFR Criteria using Recognition-Taguchi Method

Hajime Tsubaki

GITI / Language and Speech Science Research Laboratories, Waseda University, Japan

*Corresponding Author: hjm.tsubaki@gmail.com

Abstract

The purpose of this paper is to evaluate L2 (second language) learners' proficiency objectively. It was examined to estimate language proficiency using 94 statistics extracted from English conversation data on Japanese English learners' groups in educational institutions. To estimate the Japanese learner's English proficiency expressed as Global Rating scores of the Common European Framework of Reference for Languages (CEFR), the statistics were extracted automatically and/or manually, and were classified into 5 subcategories such as range, accuracy, fluency, interaction and coherence.

Considering that each the category included at least 8 items, Recognition-Taguchi method was used for compressing these items and analyzing correlations between them and the Global Rating scores simply. Outputs were calculated in the 5 respective subcategories by the method, and correlations to the scores were analyzed. As the result, one of the outputs, sensitivity β s showed correlations to the scores, and especially the 4 sensitivity β s except for the subcategory accuracy indicated correlations of being more than 0.650 to the scores.

The estimation experiment was carried out using a multiple-regression model trained by data set of 135 learners and the 4 sensitivity β s with higher correlations to the CEFR Global Rating scores in cross-validation. The correlation coefficients of 0.901 was shown between predicted proficiency scores and the L2 learners' actual CEFR Global Rating scores. These results confirmed the usability of the 4 sensitivity β s extracted from the total 94 statistics for the objective evaluation of L2 learner's language proficiency.

Keywords: L2 learners' proficiency evaluation, CEFR, Recognition-Taguchi method, multiple-regression model.

1. Introduction

A number of approaches have been proposed to evaluate second language (L2) learners' proficiency. In the field of education, statistical methods including data mining or machine learning are often adopted. These methods plays major roles in finding, extracting and analyzing both L2 learners' learning information such as patterns or processes from speech data⁽¹⁾⁻⁽³⁾. Furthermore, the L2 learners' proficiencies can be predicted by combining the data mining methods and L2 learner' data⁽⁴⁾⁻⁽⁸⁾. These results are expected to feedback to many educational scenes in order to improve in studying and teaching approaches possible for both learners and educators more effectively.

From the standpoint of quality engineering, ranking candidates and selecting criteria that influence in evaluating best students are conducted based on their suitability for MBA program by using Mahalanobis-Taguchi System (MTS) approach of Mahalanobis-Taguchi (MT) methods⁽⁹⁾ adding Mahalanobis Distance (MD), orthogonal array (OA) and the signal/noise (SN) ratio as its outputs. Correcting evaluation for learning efforts and results of every subjects in medical course students is tried by using Recognition-Taguchi method and the Mahalanobis Distance (MD) as its output⁽¹⁰⁾.

A lot of conventional researched analyze the correlations between evaluation scores rated by human raters and statistical measures or values extracted from evaluating items. These evaluation scores are estimated by consisting of statistical models such as multivariate regression models, based on the correlations between the evaluation scores and data of the evaluating items. As generating these models depends on quality and quantity of learning data highly, shortcomings such as multicollinearity

or overfitting are caused in some cases.

In this study, in order to the shortcomings suppress, an estimation experiment of the L2 learners' proficiency is conducted after compressing learning data and increasing information quantity as much as possible. The learning data of 5 subcategories such as range, accuracy, fluency, interaction and coherence are analyzed using Recognition-Taguchi method, and sensitivity β s are extracted as output. Furthermore correlation between the sensitivity β s and L2 learners' CEFR Global Rating scores are analyzed. The sensitivity β s are used for parameters for the estimation experiment with multiple-regression model and their usability is confirmed for objective evaluation of L2 learners' proficiency.

This study proceeds as follows. In Section 2, data analysis, statistical extraction methods for the estimation parameters of the L2 learners' proficiencies, and experimental setup for predicting their proficiency scores are described. In Section 3, the analytical and experimental results are presented. Finally, in Section 4, the findings and further works for discussion are summarized.

2. Research Data, Analysis Method and Experimental Design

2.1 Research Data

The data collection for this study is composed of English conversation data recorded from groups of Japanese English learners at educational institutions. The data set are collected and constructed as follows. The participants are 135 students from seven schools among three types of educational institutions: two junior high schools, two senior high schools, and three universities in Japan. They are divided into a total of 45 groups of three students each. Each group comprises three students who are each randomly chosen from one of the participating junior high schools, senior high schools, and universities, respectively. The three students interact orally as a group for five minutes on a given topic, such as family, friends, hobbies, English, and culture. Their English conversations are recorded in video format, transcribed manually, and extracted into 94 statistical measures, including the number of used tokens, types and parts of speech, utterance duration time, and number of syllables for each the student. And then, the extracted 94 statistical values are respectively classified into the following 5 subcategories in the CEFR rating scales: range, accuracy, fluency, interaction and

Table 1. Research data structure.

Learners' ID	CEFR Global Rating score	Subcategory				
		Range (Rating item: 18)	Accuracy (Rating item: 12)	Fluency (Rating item: 29)	Interaction (Rating item: 8)	Coherence (Rating item: 27)
1						
}	}	}	}	}	}	}
135

Fig. 1. CEFR Global Rating score distribution.

coherence, as shown in Table 1 ⁽¹¹⁾.

Proficiency ratings are given to the students by ten raters who are all Japanese teachers or lecturers of English having at least a master's degree in the field of English education or applied linguistics and teaching English at either high schools or universities. The ratings are assigned for a global category and 5 subcategories in the CEFR rating scales: range, accuracy, fluency, interaction, and coherence. The rating criteria corresponding to the seven CEFR levels: Below A1, A1, A2, B1, B2, C1 and C2 on both the Global Oral Assessment Scale and Oral Assessment Criteria Grid. Below A1 shows the lowest ability and C2 indicates the highest ability in the general proficiency rating. The categorical scores of Below A1 up to C2 given by the raters are converted into numerical values using the Rasch model, which is a normalization method generally used in educational fields ⁽¹²⁾. Global of these values are used as the evaluation values for analysis and estimation in this research. The normalized CEFR Global Rating scores' distribution is shown in Figure 1.

2.2 Analysis Method

Recognition-Taguchi (RT) method is used for an analysis method in this study. The method is one of Mahalanobis-Taguchi (MT) methods used in quality engineering field as a pattern recognition method⁽¹³⁾. When using the method, data in a certain condition is measured in many times and a unit space for the condition is generated. Information composed of the unit space is called a member. As for members comprising the unit space, quantities known as standard S/N ratio η and sensitivity β are defined based on a model of quality engineering. Using these quantities, distances related to the unit space are defined in accordance with the RT method. By being calculated and compared with distances in relation to unit spaces on unknown data, it is determined which unit space the unknown data is classified to. Standard S/N ratio η is a concept which evaluates stability of action or reaction under a non-linear system. Its quantity is also an index showing a degree of adaptation to an origin regression in the action or reaction. Sensitivity β is a value indicating gradient of the origin regression formula.

In this study, the total 94 statistics of L2 learner's language proficiency are narrowed to the 2 quantities of the standard S/N ratio η and the sensitivity β . Each the learner's relationship between average values of the total 94 statistics and those values is transformed to an origin regression formula. It is expected that proficient-graded learner's sensitivity β value is more than that of poor-graded learner as the proficient-graded learner's statistics may exceed the average values of the total 94 statistics in many cases. When correlation between the sensitivity β and L2 learners' actual CEFR Global Rating scores is found out, L2 Learners' proficiency evaluation can be conducted using the RT method.

In the correlation analysis between the learners' CEFR Global Rating scores and the L2 learners' sensitivity β s, the statistical characteristics of the sensitivity β s are researched. The applicability of the sensitivity β s for parameters estimating the L2 learners' proficiencies is investigated in the predicting experiment.

2.3 Predicting Experiment

An experiment for estimating the learners' Global scores is conducted based on the L2 learners' sensitivity β s. A multiple-regression model is used as linear-regression method. The estimated scores are compared with the learner's actual Global scores to verify the precision of the

Table 2. Sensitivity β correlation.

Subcategory	Correlation with Global scores
Range	0.728
Accuracy	-0.010
Fluency	0.675
Interaction	0.735
Coherence	0.852

Vertical axis: L2 learners' CEFR Global Rating scores.
Horizontal axis: Learners' sensitivity β values.

Fig. 2. Correlations between L2 learners' CEFR Global Rating scores and learners' sensitivity β .

multiple-regression model. In addition, 2 comparative experiments with a multiple-regression model and neural network are executed for verifying prediction accuracy between the estimating method of this study and conventional methods.

3. Analytical and Experimental Results

3.1 Correlation Analysis Result

The L2 learners' sensitivity β s were calculated in each the 5 subcategories. 4 of them showed correlation coefficients of being higher than 0.650 to the learners' CEFR Global Rating scores, as illustrated in Table 2. As for the accuracy, its sensitivity β showed few correlation coefficients to the learners' CEFR Global Rating scores. The reason is that accuracy rating items in the 5 subcategories have a lot of data with low differentials, and that each the accuracy rating items' correlation coefficients to the scores are ranged from 0.000 to 0.230. For the

example, Figure 2 scatter plot shows correlation between Global scores and the sensitivity β s extracted from the coherence.

3.2 Predicting Experiment Result

In the predicting experiment, in accordance with the above-mentioned 4 of 5 sensitivity β s on the range, fluency, interaction and coherence were chosen as estimation parameters for the learners' proficiencies. The research data were divided in to each 3 groups of 45 learners. Based on a linear-regression method, a multiple-regression model was built and a predicting experiment was conducted in cross-validation. The estimated scores obtained from the experiment were compared with the L2 learner's actual Global rating scores to verify the precision of the model. The overall correlation with the actual values of the estimated scores was 0.901, as shown in both Table 3 and Figure 3.

On the first comparative experiment using a multiple-regression model, 13 statistics with higher correlations to the CEFR Global Rating scores were selected from the 4-subcategory-item data except the accuracy. 7 of the statistics had positive correlations of being more than 0.700 to the scores. The 3 others contained negative correlations of being less than 0.400 to the scores as there was no statistics indicating negative ones of being less than 0.500. The experiment was performed according to the same procedure as above, and the overall correlation with the actual values of the estimated scores was 0.892, as shown in both Table 4 and Figure 4.

As for the second comparative experiment using a neural network, 82 statistics, namely all of the 4-subcategory-item data except the accuracy were used as input information. A three-layer-feedforward neural network was applied, and the number of units in the middle layer was 82. The experiment was carried out by following the same procedure as above, and the overall correlation with the actual values of the estimated scores was 0.918, as shown in Figure 5.

As the experimental results, the accuracy of the estimating method used in this study is lower than that of non-linear prediction method. However, it is considered that calculation amount can be reduced and sufficient precision is assured from the viewpoint of process efficiency.

Table 3. Estimation summary based on RT method.

Variable	B	SD	β	p-value	
Range	1.7823	0.6203	0.3062	0.0047	**
Fluency	-0.7230	1.0015	-0.0731	0.4716	
Interaction	2.6716	0.5299	0.2490	p < 0.001	**
Coherence	4.4061	0.3930	0.5561	p < 0.001	**
Constant	-11.0098	0.6240		p < 0.001	**
R^2	0.8306				

*, P < 0.05, **, P < 0.01

Vertical axis: CEFR Global scores
Horizontal axis: Estimated scores

Fig. 3. Correlations based on RT method.

Table 4. Estimation summary using multiple regression.

Variable	B	SD	β	p-value	
Tokens	0.0059	0.0213	0.0782	0.7822	
Types	0.0093	0.0697	0.0541	0.8940	
Guiraud index	0.4076	0.6402	0.1299	0.5256	
Tokens / clause	0.2774	0.0580	0.1964	p < 0.001	**
syllables incl. dysfluency	-0.0210	0.0066	-0.4945	0.0019	**
syllables excl. dysfluency	0.0297	0.0105	0.5262	0.0055	**
Mean pauses	-0.2545	0.1510	-0.0582	0.0945	
Japanese insert	-0.0431	0.0368	-0.0435	0.2439	
Group	0.1352	0.2882	0.0406	0.6399	
Total	0.3286	0.2269	0.1176	0.1501	
Abrupt change	-0.0910	0.0641	-0.0500	0.1585	
Words	0.0112	0.0016	0.3839	p < 0.001	**
Constant	-11.6313	1.6223		p < 0.001	**
R^2	0.8930				

*, P < 0.05, **, P < 0.01

Vertical axis: CEFR Global scores
Horizontal axis: Estimated scores

Fig. 4. Correlations using multiple regression.

Vertical axis: CEFR Global scores
Horizontal axis: Estimated scores

Fig. 5. Correlations using neural network.

4. Conclusions

To estimate the English learner's proficiencies, the adaptability of statistical measures extracted from English conversation data of groups of Japanese English learners in educational institutions is examined. The statistical measures are based on five subcategories in the CEFR Rating Scales: range, accuracy, fluency, interaction and coherence. The statistical measures called sensitivity β s were calculated in each subcategories by Recognition-Taguchi method. A correspondence relation between these sensitivity β s and the CEFR Global Rating scores were analyzed, and the correlation was observed.

By using a multiple-regression model, the learners' English proficiency scores (CEFR Global Rating scores) were predicted. The predicted scores of the learners showed a high overall correlation of 0.901 to their actual scores. In addition, 2 comparative experiments with conventional methods were executed for verifying prediction accuracy. These results indicate the usefulness of the statistical measures that were obtained with the Recognition-Taguchi method for the estimation of L2 learners' proficiencies.

In this study, predictions of L2 learners' proficiency scores were attempted by using both a Recognition-Taguchi method and a linear-regression method. Discrimination of L2 learners' proficiencies will be researched using Mahalanobis-Taguchi (MT) methods in future studies. Regarding discrimination, a decision tree method is conventionally used. This method has attempted to improve discrimination accuracy by manipulating the complexity parameter through the pruning and reduction of rating items. However, reducing rating items limits the quantity of information available for evaluation. As a result, on the borders of neighboring rating categories, the accuracy rate decreases. Outputs of Recognition-Taguchi method are summarized from the L2 learners' data, which means that all information in each rating category is collected without information loss. In the next work, the Recognition-Taguchi method will be applied to this kind of discrimination problem in L2 learners' proficiency evaluation instead of the decision tree method.

References

- (1) K. Hirabayashi and S. Nakagawa : "Automatic evaluation of English pronunciation by Japanese speakers using various acoustic features and pattern recognition techniques", Proceedings of INTERSPEECH 2010, pp. 598–601, 2010
- (2) H. Wang, C. J. Waple, and T. Kawahara : "Computer assisted language learning system based on dynamic question generation and error prediction for automatic speech recognition", Speech Communication, vol. 51, No. 10, pp. 995-1005, 2009
- (3) T. Cincarek et al. : "Automatic Pronunciation Scoring of Words and Sentences Independent From the Nonnative's First Language", Computer Speech and Language, Vol. 23, Issue 1, pp. 65-88, 2009
- (4) H. Wang and T. Kawahara : "Effective prediction of errors by non-native speakers using decision tree for

- speech recognition-based CALL system”, IEICE Transactions, vol. E92-D, No. 12, pp. 2462-2468, 2009
- (5) S. Kosuke, S. Masashi, and M. Yuji : “Automatic estimation of English proficiency level using corpora”, Information Processing Society of Japan SIG Technical Report 2007-NL-181, pp. 113-119, 2007
 - (6) T. Mishra, D. Kumar, and S. Gupta : “Mining students’ data for prediction performance”, Fourth International Conference on Advanced Computing Communication Technologies, pages 255-262, 2014
 - (7) M. S. Amirah., H.Wahidah, and A.R. Nuraini : “A Review on Predicting Student’s Performance using Data Mining Techniques”, Procedia Computer Science, vol. 72, pp. 414-422, 2015
 - (8) M. A. Al-Barrak and M. Al-Razgan : “Predicting students final GPA using decision trees: A case study”, International Journal of Information and Education Technology, vol. 6, No. 7, pp. 528-533, 2016
 - (9) Manisha Ketkar, Dr. Omkarprasad S Vaidya : “Evaluating and Ranking Candidates for MBA program: Mahalanobis Taguchi System Approach”, Procedia Economics and Finance, vol. 11 pp. 654-664, 2014
 - (10) Shinichiro Uetake, Hisato Nakajima, Kouya Yano, Naofumi Kimura, Kaoko Nagasawa, Eiji Kobayashi, Kazuyuki Abe, Ichiro Takagi, Kuninobu Yokota: “Examination of a New Rating System for Medical Student Results using the Mahalanobis · Taguchi System”, Tokyo Jikeikai Medical Journal, vol. 127 pp. 91-103, 2012
 - (11) Council of Europe : “Common European framework of reference for languages: Learning, teaching, assessment”, Cambridge: Cambridge University Press, 2001
 - (12) J. Negishi : “Multi-faceted Rasch analysis for the assessment of group oral interaction using CEFR criteria”, Annual Review of English Language Education in Japan, 21, pp. 111-120, 2010
 - (13) G. Taguchi, S. Chowdhury, and Y. Wu : “The Mahalanobis-Taguchi System”, McGraw-Hill, 2001