Evidence-Based Statistical Evaluation of Japanese L2-Learners' Proficiency using Principal Component Analysis

Masafumi Arai[1,*], *Hajime* Tsubaki[2], and *Yoshinori* Sagisaka[1]

[1]Department of Pure and Applied Mathematics, Waseda University, Japan
[2]Global Information and Telecommunication Institute, Waseda University, Japan

Abstract. This paper aims at an automatic evaluation of second language (L2) learners' proficiencies and tries to analyze English conversation data having 94 statistics and Global Scale scores of the Common European Framework of Reference (CEFR) given to each participant. The CEFR defines Range, Accuracy, Fluency, Interaction and Coherence as 5 subcategories, which constitute the CEFR Global Scale score. The statistics were classified into the CEFR's 5 subcategories. We used the Principal Component Analysis (PCA), an unsupervised machine learning method, on each subcategory and obtained the participants' principal component scores (PC scores) of the 5 subcategories for estimation parameters. We predicted the participants' CEFR Global scores using the Multiple Regression Analysis (MRA). The proposed prediction method using the PC scores was compared with conventional methods with the 94 statistics. Based on the coefficients of determination (R^2), the value of the proposed method (0.82) was nearly equivalent to one of values obtained by the conventional methods. Meanwhile, as for standard deviation, the proposed method showed the smallest value in the comparison. The results indicated usability of the PCA and PC scores calculated from the CEFR subcategory data for objective evaluation of L2 learners' English proficiencies.

Keywords

Principal Component Analysis; Multiple Regression Analysis; CEFR; L2; Evaluation

1 Introduction

A lot of methods have been proposed to evaluate L2 learners' proficiencies objectively. In educational fields, data mining is usually applied to many research cases. The data mining in these areas plays major roles to find and analyze tendencies and patterns on the L2 learners' learning from speech and text data [1]-[4]. In addition, these learners' proficiencies can be estimated by a combination of data mining methods and L2 learners' data [5]-[8]. Conventionally, a lot of research have tried to predict L2 learners' proficiencies using multivariate analysis or other statistical techniques such as the multiple regression analysis (MRA) and the correlation analysis.

In the previous study, the correlation analysis was used to select the important variables for evaluation [9]. The research used only the variables which have stronger correlations to the Common European Framework of Reference (CEFR) scores for estimation [10]. However, in this way, all the information contained in the statistics cannot be utilized. There is a possibility that other excluded statistics have useful information for objective evaluation of L2 learners' proficiencies. In our study, the Principal Component Analysis (PCA), one of the unsupervised learning methods, was used to analyze L2 learners' conversation data of the previous study and to realize the evaluation based on evidence-based statistics.

PCA is a dimension reduction technique forming new variables which are linear combinations of the original variables. This makes it possible to utilize all the information that the original statistics have with reducing the number of variables used for the MRA. We obtained the participants' principal component scores (PC scores) of 5 subcategories defined in the CEFR, such as Range, Accuracy, Fluency, Interaction and Coherence. Then, those PC scores were used as explanatory variables for estimation of the learners' CEFR Global scores using the MRA.

The study proceeds as follows. In section 2, data used in the research are described. In section 3, the analysis methods are illustrated. The results are presented in section 4. Finally, in section 5, the conclusion and future works are discussed.

2 Research data

The data set for this study is English conversation data of Japanese English learners' groups in educational institutions. The data set were collected and constructed

* Masafumi Arai: araimasa23@gmail.com

Table 1. Educational background of the participants

Institution	Participant	Group
Junior high school	45	(15)
Senior high school	45	(15)
University	45	(15)
Total	**135**	**(45)**

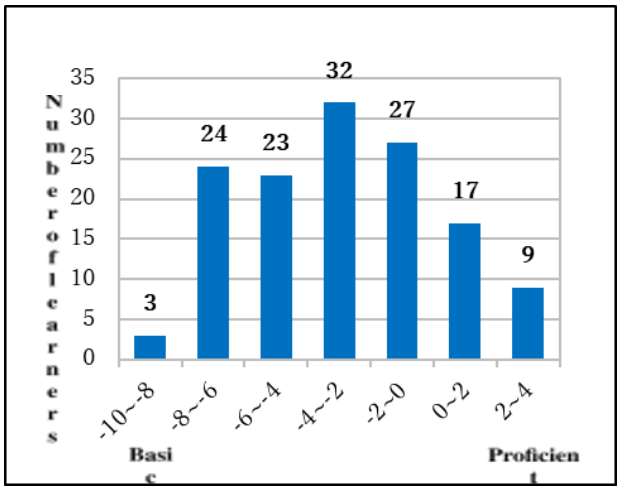

Fig. 1. Normalized distribution of CEFR Global scores

as follows; the participants are 135 students from 7 schools among 3 kinds of educational institutions, namely, 2 junior high schools, 2 senior high schools, and 3 universities in Japan as shown in Table 1. They are divided into a total of 45 groups having 3 students. Each group consists of 3 students who are randomly selected from junior high, senior high and university students. 3 students are interacted orally as a group for 5 minutes on given topics which seemed to be daily and usual for most of the people such as Family, Friends, Hobbies, English, and Culture.

Their English conversations were recorded in video format and transcribed. 94 statistics were extracted from those records. These statistics were determined in the previous study [9]. The research considered the factors to rate learners' levels of proficiency based on CEFR criteria, for example, Number of Tokens, Speaking duration time, Number of errors and so on. These 94 statistics were categorized into 5 subcategories such as Range, Accuracy, Fluency, Interaction and Coherence in the CEFR Rating Scales.

Proficiency rating for the students is conducted by 10 raters who are all Japanese teachers of English holding a minimum of a master's degree in the fields of English education or applied linguistics, and teaching English at either high schools or universities. The rating is conducted for 6 categories which are Global and the 5 subcategories. The rating criteria are 7 levels of the CEFR - Below A1, A1, A2, B1, B2, C1, C2 - on both the Global Oral Assessment Scale and Oral Assessment Criteria Grid. Below A1 corresponds to lower ability or "Not good", and sequentially, C2 applies to higher ability or "Excellent" in general proficiency rating. These categorical scores of Below A1 to C2 obtained from the raters' ratings are changed into numerical values by Rasch model, one of the normalizing methods which is generally used in educational fields [9]. The normalized distribution of CEFR Global scores is shown in Figure 1.

3 Evaluation method

Estimation accuracies of three different evaluation methods were compared in this research (Method 1, 2 and 3). The participants' CEFR global scores are the objective variables in all the cases. Each method uses different explanatory variables for its MRA.

Brief explanations of the three methods are described below. They are explained more precisely in the subsections following (see the subsection 3.1, 3.2 and 3.3).

Method 1; MRA uses as many original statistics as possible.

Method 2; MRA uses 8 variables selected in the previous study [9]. These statistics have stronger correlations to CEFR scores.

Method 3; MRA uses the PC scores. PC1, 2 and 3 scores of the CEFR 5 subcategories are used as explanatory variables.

We applied the principal component analysis (PCA) on each subcategory. PCA is one of the unsupervised machine learning techniques used for dimension reduction. PC scores of the 5 subcategories are used as explanatory variables in the Method 3. The 1st, 2nd and 3rd PC scores (PCi scores, i = 1, 2, 3) are used. The estimation processes are performed by a 9-fold cross-validation. Values of the coefficients of determination (R^2) are used to compare the accuracies of the three estimation methods.

3.1 Method 1 (MRA using Original Variables)

The estimation Method 1 is the MRA using as many original statistics as possible. The statistics are used as explanatory variables. 85 statistics of the 94 original ones are used in Method 1. Due to the multicollinearity, 8 variables are removed. One variable is not used because only 1 participant had a value which is more than 0 meanwhile the others had the value of 0 in this item. The participants' CEFR global scores are the objective variables. In this method, overfittings to training data occur due to the larger number of the explanatory variables.

3.2 Method 2 (MRA using Selected Variables)

The estimation Method 2 is the MRA using only 8 statistics of the 94 original ones. These variables were selected in the previous study [9]. Table 2 shows the items used in this estimation. These 8 statistics are used as explanatory variables in the estimation. The 8 items

have stronger correlations to the CEFR scores of the subcategories into which each item is classified. The variables with the correlation coefficients which are more than 0.6 were extracted.

Table 2. The 8 items selected in the previous study [9]

Subcategory	Item	Correlation coefficients
Range	- Number of types (ID: **RA_2**)	0.62
	- Total number of formulaic sequences (**RA_13**)	0.65
Accuracy	- *Ratio of self-corrections per error* (**AC_12**)	*0.21*
Fluency	- Total speaking time including pause time (**FL_1**)	0.63
	- Number of syllables including dysfluency (**FL_2**)	0.77
	- Number of words (**FL_4**)	0.75
Interaction	- Global Interactional Patterns (group + individual) (**IN_7**)	0.61
Coherence	- Number of words used as a group (**CO_24**)	0.72

It should be noted that one item with the weaker correlation (Ratio of self-corrections per error, 0.21) is included from the subcategory Accuracy. According to the previous study, this is because at least one item from each subcategory should be included when estimating the Global assessment. There are no variables in the subcategory Accuracy having correlation coefficients whose values are more than 0.6. The participants' CEFR global scores are the objective variables, as well as the other methods.

In this method, overfittings to training data can be alleviated due to the smaller number of the explanatory variables. However, information that the remaining 84 variables have, are not used for the analysis.

3.3 Method 3 (MRA using PC scores)

The estimation Method 3 is the MRA using the PC scores. We applied the PCA on the 5 subcategories to obtain the participants' PC scores of the subcategories. It is possible to reduce the number of values used for the MRA without disposing of information that the original statistics have, because PC scores are linear combinations of the original variables. The PC 1, 2 and 3 scores of the 5 subcategories are used as the explanatory variables. The number of the explanatory variables in the Method 3 (15), is fewer than that of the Method 1 (85), which can alleviate overfittings or selection bias. The participants' CEFR global scores are the objective variables, as well as the Method 1 and 2.

3.4 Principal Component Analysis

PCA is a dimension reduction technique forming new variables which are linear combinations of the original variables [11].

Let $X = (x_1, x_2, \ldots, x_p)$ the original random vector. And let $y_1, y_2, \ldots, y_p$ represent the linear combinations of the

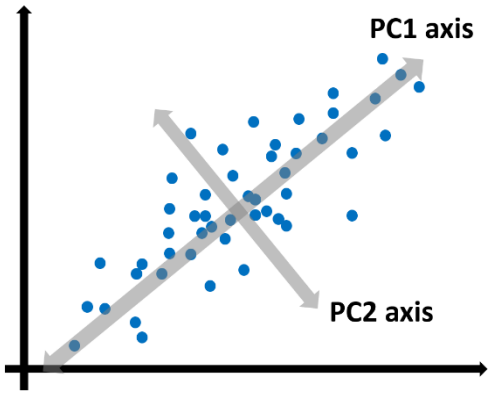

Fig. 2. PCA gives a new set of orthogonal axes. (e.g. $p = 2$)

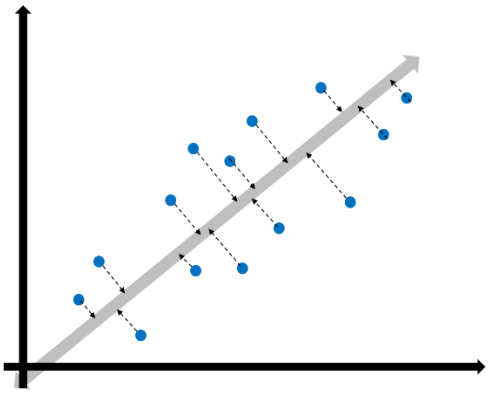

Fig. 3. PC scores are coordinates on new axes.

original variables so that they are new variables. Then,

$$y_1 = w_{11}x_1 + w_{12}x_2 + \ldots + w_{1p}x_p$$

$$y_2 = w_{21}x_1 + w_{22}x_2 + \ldots + w_{2p}x_p$$

$$\ldots$$

$$y_p = w_{p1}x_1 + w_{p2}x_2 + \ldots + w_{pp}x_p$$

where the coefficients w_{ij} are the weights of the j^{th} variables ($j = 1, 2, \ldots, p$) for the i^{th} new variable. The values of w_{ij} ($j = 1, 2, \ldots, p$) are called factor loadings of i^{th} principal component. They are determined to maximize the variance of the new variables, meeting equations below.

$$\sum_{j=1}^{p} w_{ij}^2 = 1 \sum_{j=1}^{p} w_{ij}^2 = 1$$

Equation 1

$$\sum_i \sum_j w_i w_j = 0 \sum_i \sum_j w_i w_j = 0$$

Equation 2

Equation 1 is used to fix the scale of the new variables and Equation 2 means that the new axes are orthogonal to each other.
Geometrically, PCA gives a new set of orthogonal axes as shown in Figure 2. The coordinates of the

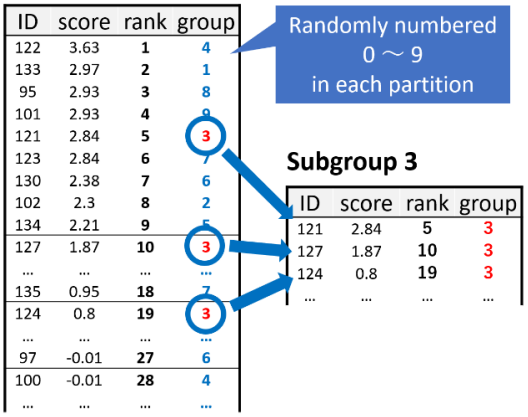

Fig. 4. Procedure for making 9 subgroups

observations with respect to each of the axes equal the values for the new variables. The new axes and the new variables are called principal components and the values of the new variables are called PC scores.
The y_i (i = 1, 2, …, p) in the formulas are defined as PC_i scores, and these are the coordinates on the new i^{th} axis as shown in Figure 3. The set of PC1 scores accounts for the maximum variance in the data, according to the above-described definition.

3.5 Cross-Validation

The prediction experiments by Method 1, 2 and 3 are conducted in a 9-fold cross-validation. The cross-validation is a model assessment technique used to measure the accuracy of a prediction method for analyzing unknown data which is not included in training data. In a k-fold cross-validation, the original sample dataset is divided into k subgroups. One of them is retained as a testing set and the others (k-1 subgroups) are used for training data. After building a model by using the training set, the model is tested on the testing data. This process is repeated k times, in which each of the k subgroups is used as testing data once. The number of the subgroups 9 was determined on the basis that the number of the participants (135) is a multiple of 9. In an alternative option, the dataset can be divided into 3 or 5 subgroups. However, we put our priority on keeping the number of samples for training, so we adopt 9-fold cross-validation. The number of training data is 90 in case of 3-fold cross-validation, and 120 in the event of 9-fold, respectively.
A procedure for making 9 groups is as follows;

(1) The samples are sorted in a descending order according to the CEFR Global scores.
(2) Then, the dataset was divided into every 9 sample from the beginning to the end.
(3) Next, in each partition, samples are numbered from 1 to 9 randomly.
(4) Finally, samples which have the same numbers are collected into a partition and 9 subgroups are formed as shown in Figure 4.

Even though the sort process (1) is not included in general k-fold cross-validations, this procedure makes sure that every subgroup has participants with higher and lower levels of proficiency well in balance.

3.6 Coefficients of Determination

Values of the coefficients of determination (R^2) are used to compare the accuracies of the three estimation methods.
Let y_i (i = 1, 2, …, 135) observed values of the Global CEFR scores. And let fi (i = 1, 2, …, 135) predicted values calculated by a model. Then, the definition of R^2 is as follows:

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

The numerator of the second term is the summation of squares of residuals, differences between observed values and predicted ones. And the denominator is the summation of squares of deviations. The range of the value R^2 is $0 \leqq R^2 \leqq 1$ by the definition. The R^2 values provide assessments of a prediction model's accuracy, based on the proportion of residuals to total variations. The R^2 values approximate to 0 correspond to lower accuracy, and values close to 1 apply to higher precision sequentially.

4 Results

Table 3 shows the results of the learners' proficiency estimation methods. The Method 2 and 3 showed higher values of the R^2 in their cross-validations, which means that the model made by the Method 2 and 3 estimated the CEFR Global scores more accurately than the Method 1. The mean value of the R^2 on Method 1 was 0.61, and those of Method 2 and 3 were 0.82. As for Method 2 and 3, these R^2 mean values were not exactly the same but had no critical differences considering the significant figures and randomness of grouping processes in the validation. In addition, the Method 3 showed the smallest standard deviation. The results showed usability of the PCA and PC scores calculated from the CEFR subcategory data for objective evaluation of L2 learners' English proficiencies.

Table 4 is a summary of the estimation model by Method 2. The model shown in this table is made by the MRA on all the dataset (All the 135 samples are used to analyze). The abbreviations B, SE and β in the table stand for the partial regression coefficient, the standard error and the standard partial regression coefficient, respectively. The original names of the variables with their IDs are presented in Table 2 in section 3.

Table 5 shows a summary of the estimation model by Method 3. The model presented in this table is made by the MRA using all the 135 samples. The variable names xxPC*i* in the table represent the PCi score of a subcategory whose name begins with the letters xx. For instance, raPC1 means the PC1 scores of the subcategory Range.

Table 3. Method 3 shows higher R^2 and the smallest σ

Testing Set	Method 1	Method 2	Method 3
1	0.60	**0.90**	0.76
2	0.81	**0.84**	**0.84**
3	0.38	**0.78**	0.76
4	0.68	0.73	**0.79**
5	0.78	0.89	**0.91**
6	0.63	**0.83**	0.81
7	0.63	0.68	**0.73**
8	0.56	0.84	**0.85**
9	0.45	**0.91**	0.90
Mean	**0.61**	**0.82**	**0.82**
σ	**0.13**	**0.073**	**0.058**

Table 4. Summary of the estimation model by Method 2

Variable ID	B	SE	β	P-value	
RA_2	0.1007	0.0192	0.5850	P < 0.001	**
RA_13	0.1239	0.0372	0.1881	0.0011	**
AC_12	1.0407	0.5924	0.0633	0.0814	
FL_1	-0.0016	0.0061	-0.0217	0.7887	
FL_2	-0.0036	0.0071	-0.0846	0.6149	
FL_4	-0.0203	0.0131	-0.2681	0.1255	
IN_7	0.4608	0.2623	0.0732	0.0814	
CO_24	0.0170	0.0015	0.5847	P < 0.001	**
Constant	-11.1590	0.5733		P < 0.001	**
R^2	**0.8472**			*: P<0.05 **: P<0.01	

Table 5. Summary of the estimation model by Method 3

Variable ID	B	SE	β	P-value	
raPC1	0.3832	0.1149	0.3450	0.0011	**
raPC2	-0.1354	0.0997	-0.0820	0.1771	
raPC3	-0.4370	0.1352	-0.1906	0.0016	**
acPC1	-0.1445	0.0706	-0.0792	0.0429	*
acPC2	0.1387	0.0934	0.0615	0.1403	
acPC3	0.0638	0.1156	0.0245	0.5817	
flPC1	0.0466	0.0983	0.0464	0.6366	
flPC2	-0.1611	0.0983	-0.0919	0.1040	
flPC3	-0.1058	0.0834	-0.0557	0.2073	
inPC1	-0.5172	0.1184	-0.2611	P < 0.001	**
inPC2	0.1403	0.1150	0.0578	0.2251	
inPC3	0.0044	0.1146	0.0014	0.9696	
coPC1	-0.2038	0.0785	-0.1539	0.0106	*
coPC2	0.3477	0.1050	0.2139	0.0012	**
coPC3	0.3832	0.0974	0.1837	P < 0.001	**
Constant	-2.8728	0.1132		P < 0.001	**
R^2	**0.8402**			*: P<0.05 **: P<0.01	

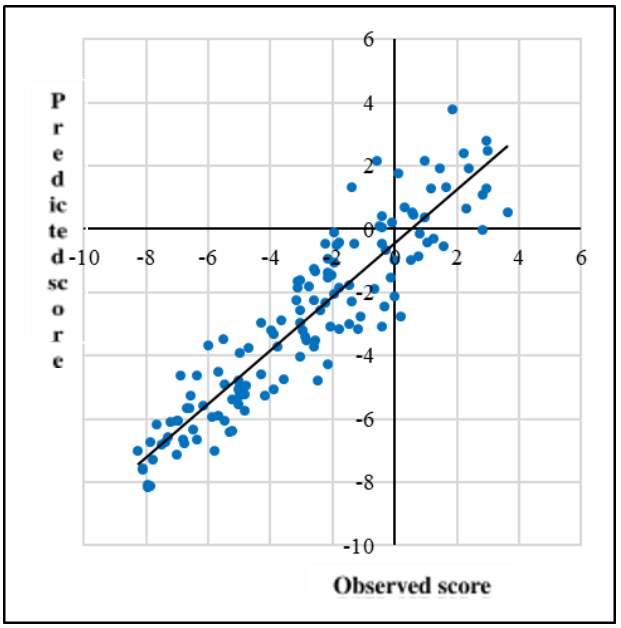

Fig. 5. Correlations between predicted and observed score (Method 2)

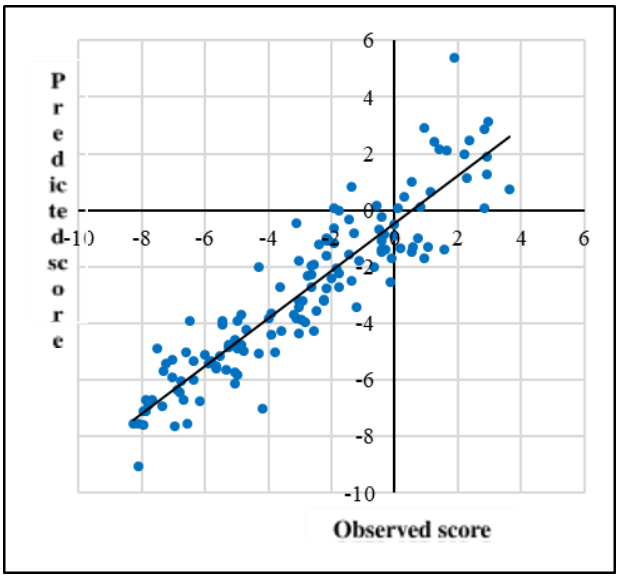

Fig. 6. Correlations between predicted and observed score (Method 3)

Correlations between the observed CEFR Global scores and the predicted values estimated by Method 2 and 3, were illustrated in Figure 5 and 6, respectively.

5 Conclusion and future works

For objective evaluation of L2 learners' English proficiencies, usability of the PCA and PC scores of the CEFR subcategories were observed. The MRA using PC scores of the CEFR 5 subcategories, the prediction method proposed in our research, estimated the CEFR Global scores more accurately than the MRA using the original variables. Dimension reductions by the PCA make it possible to reduce the number of values used for the MRA (or other conventional multivariate analysis techniques) without disposing of information that the original statistics have, because PC scores are linear combinations of the original variables. Overfittings and selection bias can be alleviated by this method with making use of all the information that the original statistics have.

In our future works, we plan to analyze characteristics and tendencies of the participants whose proficiencies are estimated more (or less) accurately by the evaluation method 2 and 3, so that we can propose efficient ways to use these methods together to evaluate L2 learners' proficiencies more precisely.

We would like to express our gratitude to Dr. Junko Negishi at Tsurumi university, who offered the research data used in our study.

References

1. K.Hirabayashi, S.Nakagawa: Automatic evaluation of English pronunciation by Japanese speakers using various acoustic features and pattern recognition techniques. Proc.Interspeech, 598-601 (2010)
2. H.Wang, C.J.Waple, T.Kawahara: Computer assisted language learning system based on dynamic question generation and error prediction for automatic speech recognition. J. Speech Communication, Vol.51, No.10, 995-1005 (2009)
3. S. Nakamura, S. Matsuda, H. Kato, M. Tsuzaki and Y.Sagisaka: Objective evaluation of English learners' timing control based on a measure reflecting perceptual characteristics. Proc. IEEE ICASSP, 4837-4840 (2009)
4. S. Nakamura, H. Kato and Y. Sagisaka: Effects of Mora-timing in English Rhythm Control by Japanese Learners. Proc. INTERSPEECH 2009 1539-1542 (2009)
5. H.Wang, T.Kawahara: Effective prediction of errors by non-native speakers using decision tree for speech recognition-based CALL system. IEICE Trans., Vol.E92-D, No.12, 2462-2468 (2009)
6. Keiji Yasuda, Eiichiro Sumita, Seiichi Yamamoto, Masuzo Yanagida, Kikuo Maekawa, and Fumiaki Sugaya: A Proposal for Automatically Gauging of English Language Proficiency. IPSJ SIG Technical Report, Vol.2003-NL-155:65-70 (2003)
7. Sakata Kosuke, Shimbo Masashi, Matsumoto Yuji: Automatic estimation of English proficiency level using corpora. Information Processing Society of Japan SIG Technical Report 2007-NL-181, 113-119 (2007)
8. Mashael A. Al-Barrak and Muna Al-Razgan: Predicting Students Final GPA Using Decision Trees: A Case Study. International Journal of Information and Education Technology, Vol. 6, No. 7, July 2016, 528-533 (2016)
9. Junko Negishi: Multi-faceted Rasch analysis for the assessment of group oral interaction using CEFR criteria. Annual Review of English Language Education in Japan, 21, 111-120 (2010)
10. Council of Europe: "Common European framework of reference for languages: Learning, teaching, assessment", Cambridge: Cambridge University Press (2001)
11. Aboagye, E.A. and Mensah Cynthia: Principal Component Analysis of Students' Academic Performance in Mathematics and Statistics. American Based Research Journal, Vol-5-Issue-10. (2016)
12. M. Arai, T. Hajime, Y. Sagisaka: Principal Component Analysis on English Conversation Data of Japanese L2-Learners. International Workshop of Intelligent Data Analytics and Applications Joint with JSAI International Symposia on AI. (2018)