Classification of English sentences by the degree of difficulty using machine learning

Toshihiko Shimauchi, Ryo Oguri, Hiromi Ban, Hidetaka Nambo, Haruhiko Kimura

***Abstract*— This study aims to develop a system to classify English sentences by the degree of difficulty by using English textbooks of Finland, Japan, and South Korea. First, the data sets are built by extracting features from English sentences included in 20 paragraphs from English textbooks used in Finland. The Random Forests algorithm is applied to the data set to build a classifier. This method leads to a classifier which is able to classify sentences with higher accuracy. Second, a two-tier classifier method is applied to wider datasets from textbooks of Finland. The experiment shows the effectiveness of implementing multi-tier classifiers. These two new methods are applied to textbooks used in Japan and South Korea. The results of the experiments show that a model which classifies English sentences with higher accuracy can be developed by following the proposed methods.**

***Index Terms*— Random Forest, Machine Learning, Text Mining, Readability Score, TESOL**

I. INTRODUCTION

In recent years, English has increasingly gained the importance. Out of world population of approximately 73 billion, 21 billion people live in countries where English is an official or semi-official language [1], making English the most widely spoken language in the world.

In Japan, where English is not used as a semi-official language, there are several developments to promote learning or utilizing the language. In educational realm, since 2014, top-tier universities and high schools are designated respectively as super global universities and super high schools to educate students who will play leading roles in global society [2], [3]. In business realm, since 2010, many listed companies such as Rakuten, Fast Retailing, and Honda Motor have started to introduce English as an official in-house language [4].

Additionally, many people in Japan take various types of certificate exams for various purposes, one of which is for self-cultivation. Table 1 shows the three top ranking exams arranged according to the number of examinees. EIKEN and TOEIC, both certificate exams for English, are most popular. This suggests there are huge demand for English learning.

All these developments suggest that, in Japan, English has been given significant priority over other languages. However, English is not a semi-official language. Those who want to be good at English have to study harder compared to people living in countries where English is widely used. In order to study English better, it is important to look at not only amounts of study hours but also methods of learning and teaching. There already exist wide range of studies on English learning which suggest the importance of using study materials appropriate to the proficiency of each learner [5]-[7]. However, it is not easy to know beforehand the exact difficulty level of a given material, making it difficult for each learner to select the material appropriate for his or her proficiency.

Table 1: Number of certificate examinees in Japan

	Certificate Exams	Examinees in 2015
1st	EIKEN	3,225,358
2nd	TOEIC	2,779,300
3rd	KANKEN	2,103,271

English textbooks used in school take into account the proficiency of English learners. These school textbooks "are written and edited with proper consideration for the order of learning by meticulously controlling vocabularies and sentence structures for the sake of learners' aptitude" [5]. Chujo et al. calculate the difficulty of English texts by using readability scores and percentages of words not covered in school textbooks in Japan and UK [6]. Chujo et al. classify difficulty of English textbooks by using corpus data [5]. Ban and Oyabu analyze English textbooks by applying quantitative linguistics method and find features which changes according to grade [8].

In this study, by using features extracted from text data of school textbooks as learning data, we propose to develop a system which can classify difficulty levels of English textbooks.

II. PROPOSED SYSTEM

A. Outline

In this study, classifiers are built by using features extracted from English text and then develop a system to classify difficulty level of given English textbooks. Fig. 1 shows the process of building classifiers. First, features of English textbooks are extracted to develop training datasets. After building classifiers, the training datasets are used to validate the accuracy of the classifiers. Leave-one-out cross validation method is applied.

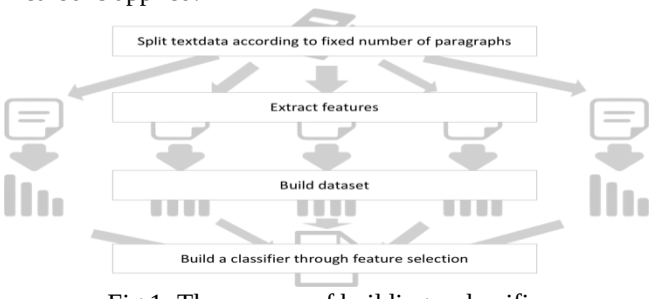

Fig.1: The process of building a classifier

Toshihiko Shimauchi, Department of Regional Design and Development, Komatsu College, Komatsu, Ishikawa, Japan.
Ryo Oguri, Rinnai Corporation, Japan.
Hiromi Ban, Graduate School of Engineering, Nagaoka University of Technology, Nagaoka, Niigata, Japan.
Hidetaka Nambo, Graduate School of Natural Science and Technology, Kanazawa University, Kanazawa, Ishikawa, Japan.
Haruhiko Kimura, Faculty of Production Systems Engineering and Sciences, Komatsu University, Komatsu, Ishikawa, Japan

B. Data Used

Text data from English school textbooks of Finland, Japan and South Korea are used in this study. Table 2 lists all the textbooks analyzed in this study. For the textbook used in 3rd grade of South Korea high schools which include both reading and writing sections, only the reading section is used. Hereafter, data will be described by using country and grade, such as "from E3 to E6 in Finland".

Table 2: Textbooks used

Title	Grade*	Country**	Year	Publisher
Wow! 3	E3	FIN	2002	WSOY
Wow! 4	E4	FIN	2003	WSOY
Wow! 5	E5	FIN	2005	WSOY
Wow! 6	E6	FIN	2006	WSOY
KEY 7	J1	FIN	2002	WSOY
KEY 8	J2	FIN	2003	WSOY
KEY 9	J3	FIN	2004	WSOY
NEW HORIZON English Course 1	J1	JPN	2010	Tokyo Shuppan
NEW HORIZON English Course 2	J2	JPN	2010	Tokyo Shuppan
NEW HORIZON English Course 3	J3	JPN	2010	Tokyo Shuppan
UNICORN ENGLISH COURSE I	H1	JPN	2010	Bun-eido
UNICORN ENGLISH COURSE II	H2	JPN	2010	Bun-eido
UNICORN ENGLISH COURSE READING	H3	JPN	2010	Bun-eido
MIDDLE SCHOOL ENGLISH 1	J1	KOR	2008	Genius Education
MIDDLE SCHOOL ENGLISH 2	J2	KOR	2009	Genius Education
MIDDLE SCHOOL ENGLISH 3	J3	KOR	2010	Genius Education
HIGH SCHOOL ENGLISH I	H1	KOR	2009	Genius Education
HIGH SCHOOL ENGLISH II	H2	KOR	2009	Genius Education
HIGH SCHOOL ENGLISH READING AND WRITING	H3	KOR	2009	Genius Education

* E: elementary school, J: Junior High School, H: High school

** FIN: Finland, JPN: Japan, KOR: Korea

C. Features

Table 3 shows features used to generate a dataset. Of 13 features in the table, 11 are used in the study by Ban et al. 2012, and other two are average syllables used to calculate readability score and "average syllables x 84.6" used in Flesch Reading Ease Score, one of the most widely used readability scores.

Table 3: Features used in the experiments

Total letters	Average word length
Total letter types	Words / sentence
Total words	Sentences / paragraph
Total word types	Words / word types
Total sentences	Comma / sentence
Total paragraphs	Average syllables
Average syllables x 84.6	

D. Process of Generating Datasets

Fig. 2 shows the process of generating a dataset by using 25 paragraphs for one instance. First, text data are preprocessed to fit one paragraph data into one line. 25 lines are used as a unit to extract features. Extracted features are aggregated to make dataset. Fig. 3 shows a sample of text data and extracted features. The dataset generated from the process are partially listed in Fig. 4. Labels of dataset are manually adjusted for each grade.

E. Proposed Method 1: Refining Process of Dataset Building

In an existing study, datasets used in the experiments are generated by extracting features from sentences contained in one page [9]. However, using a page as a unit hinders accurate classification due to the difference in the number of sentences between textbooks for training and those for test. This leads to a less versatile model. To solve this issue, we propose a new method: using a paragraph as a unit and extract data from sentences included in the appropriate number of paragraphs. By using paragraphs as a unit, this method can be applied to wide range of documents beyond textbooks used in the study, possibly leading to a development of system which can classify difficulty of wide range of books.

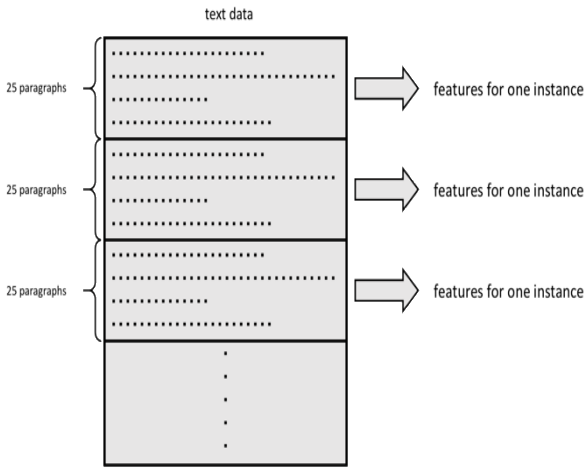

Fig.2: Process of build dataset

Fig.3: Feature extraction

Total letters	Total letter types	Total words	Total word types	Total sentences	' ' '	average syllables	average syllables * 84.6	label
884	46	154	83	48	' ' '	1297	109.76	a
576	52	105	56	37	' ' '	1106	93.543	a
667	46	111	55	40	' ' '	1395	117.983	a
765	47	131	70	44	' ' '	1298	109.851	a
1041	44	201	89	56	' ' '	1196	101.141	a
747	47	132	87	40	' ' '	1132	95.752	a
673	49	119	78	45	' ' '	1187	100.454	a
' ' '	¥	' ' '	' ' '	' ' '	' ' '	' ' '	' ' '	' ' '

Fig.4: Dataset (partial)

F. *Experiments and Validation*

Training datasets are loaded to Weka to find feature subsets with the highest feature importance by using a feature selection method. As a feature selection method, brute force search is applied. Random Forest is used to build a model [10], [11]. The feature subsets located by the search are used as training data. To validate the accuracy of classifiers, leave-one-out cross validation is used, since the datasets do not contain sufficient number of instances. Accuracy and F-measure are used as validation indices.

III. EXPERIMENT 1

A. *Outline*

This experiment aims to find optimum amount of text data required to extract features for one instance used in training dataset. Text data used are from four English textbooks from E3 to E6 grade in Finland. Five datasets are generated according to the number of paragraphs: from 5, 10, 15, 20 and 25. Table 4 shows datasets generated in this experiment.

Table 4: Datasets outline

Number of paragraphs in one instance	Number of instances				total
	E3	E4	E5	E6	
5	73	98	114	110	395
10	36	49	57	55	197
15	24	32	38	36	130
20	18	24	28	27	97
25	14	19	22	22	77

B. *Result*

Table 5 shows the results of the experiment. Both accuracy and F-measure are improved with the increase of the number of paragraphs for one instance. However, there is only a slight improvement between 20 paragraphs and 25 paragraphs.

Table 6 details the classification result of the experiment. There are few instances to misclassify lower grades (E3 and E4) as higher grades (E5 and E6) and vice versa. However, there are more misclassifications between E3 and E4, and between E5 and E6. Table 7 shows the selected features in the model which is built by training dataset based on 20 paragraphs as a unit.

Table 5: Result of experiment 1

The number of paragraphs in one instance	accuracy (%)	F -measure
5	52.658	0.525
10	56.853	0.568
15	57.692	0.578
20	64.949	0.650
25	64.935	0.645

Table 6: Result of classification

		Actual grade			
		E3	E4	E5	E6
Predicted grade	E3	14	5	1	0
	E4	3	14	2	0
	E5	1	4	17	9
	E6	0	1	8	18

Table 7: Selected features

Total letter types
Total words
Total sentences
Sentences / paragraph
Words / word types

C. *Discussion*

A certain amount of training data is required to build a classifier. Based on the number of available data, accuracy and F-measure, 20-paragraph is adopted as a proper unit for one instance to run further experiments. Also, compared to the existing study using page as a unit of analysis, this study shows higher F-measure. This result shows that, by using paragraph as a unit of analysis, the system can be developed which not only has wider applicability to many study materials but also can classify with higher accuracy.

Average syllables per word, a feature used in existing readability scores, is not selected by the feature selection. Instead, the number of words per sentence is selected. This result indicates that in primary school, there are little changes in the number of syllables per word and that after junior high school, the number would increase which leads to the rise in difficulty levels. Also, the misclassifications between E3 and E4 and between E5 and E6 indicate that, although the selected features are appropriate to classify data with higher accuracy in general, they are different from features that can classify E3 and E4 and E5 and E6 more accurately.

IV. PROPOSED METHOD 2: TWO-TIER CLASSIFICATION

The result of the experiments 1 shows several pairs of grades which are mutually misclassified and those which are not misclassified at all. The result indicates that multi-tier classifications by using several models can perform better classification than one-tier, single model classification into 4

grades. Based on this finding, we propose second method: two-tier classification with first stage classifier to perform general classification and second stage classifier to perform finer classification.

One-tier classification used in the existing study and two-tier classification we propose are applied to the following experiments to compare the accuracy of each method.

V. EXPERIMENT 2

A. Outline

The result of experiment1 leads to the hypothesis that feature subsets which can classify accurately between E3 and E4 and between E5 and E6 are different from those that can better classify entire grades are different. In order to verify the hypothesis, second experiment is conducted. At the first stage, classifier 1 is placed to make binary classification between lower-grade group (E3 or E4) and higher-grade group (E5 or E6). At the second stage, two classifiers, classifier 2 and 3, respectively classify lower-grade group into E3 and E4 and higher-grade group into E5 and E6 to obtain 4 classes. Data used are same as experiment 1. Fig. 5 shows the process of the two-tier classification.

B. Result

The results of the first stage and second stage classifications are respectively shown in Tables 8 and 9. Table 10 shows the comparison of the accuracy and the F-measures of one-tier and two-tier classifications. Two-tier classification shows better result: 19.46 points higher accuracy and 0.057 higher F-measure. This result proves the effectiveness of two-tier classification. Feature subsets selected by the algorithm are listed in bold in Fig. 6.

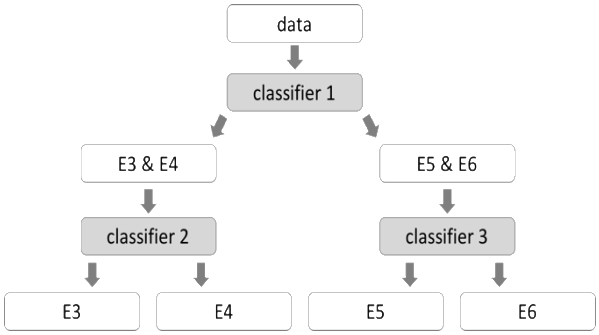

Fig.5: Process of two-tier classification for Finland

Table 8: Result of first stage classification

		Actual grades	
		E3&E4	E5&E6
Predicted grades	E3&E4	38	3
	E5&E6	4	52

Table 9: Result of second stage classification

		Actual grade			
		E3	E4	E5	E6
Predicted grade	E3	14	3	1	0
	E4	3	18	2	0
	E5	0	1	20	9
	E6	1	2	5	18

Table 10: Result comparison: one-tier vs two-tier

classifier	accuracy (%)	F-measure
One-tier	64.949	0.650
1st / Two-tier	92.784	0.928
2nd / Two-tier	72.165	0.722

One-tier	Two-Tier		
	classifier 1	classifier 2	classifier 3
Total letters	Total letters	Total letters	Total letters
Total letter types	Total letter types	Total letter types	Total letter types
Total words	Total words	Total words	Total words
Total word types	**Total word types**	Total word types	Total word types
Total sentences	Total sentences	Total sentences	Total sentences
Average word length	Average word length	Average word length	Average word length
words / sentence	words / sentence	words / sentence	words / sentence
sentences / paragraph	sentences / paragraph	sentences / paragraph	sentences / paragraph
words / word types	words / word types	words / word types	words / word types
comma / sentence	comma / sentence	comma / sentence	comma / sentence
average syllables	average syllables	average syllables	**average syllables**
average syllables * 84.6	average syllables * 84.6	average syllables * 84.6	**average syllables * 84.6**

Fig. 6: Comparison of feature subsets

C. Discussion

Using three classifiers to run two-tire classification to obtain 4 classes results in higher accuracy compared to using one-classifier-single-stage classification, suggesting the effectiveness of the two-tier classification method.

Fig. 6 shows two types of features: those which function at each classification, such as the total number of sentences in 20 paragraphs, and those which function at specific classification, such as average syllables per one word.

Classifiers 2 and 3 use distinct feature subset, while the subsets of the classifier of one-tier classification and the classifier 1 of two-tier classification are composed of similar features. This result supports the finding in the experiment 1 which suggests that feature subsets to classify lower-grade group into E3 and E4 and higher-grade group into E5 and E6 are different from the subsets for classifying data into lower and higher grades.

VI. EXPERIMENTS TO EXPAND THE RANGE OF GRADES

A. Outline

This experiment uses dataset for 7 years: from elementary E3 to Junior-High J3 in Finland to run one-tier and two-tier classifications. Dataset are generated by using text based on 20 paragraphs. Table 11 shows the number of the instances for each grade. Fig. 7 shows the process of two-tier classification. At the first stage, the data are classified into three classes: E3 and E4, E5 and E6 and J1 to J3.

Table 11: Instances for each grade (Finland)

Grade	Instances
E3	18
E4	24
E5	28
E6	27
J1	21
J2	27
J3	32

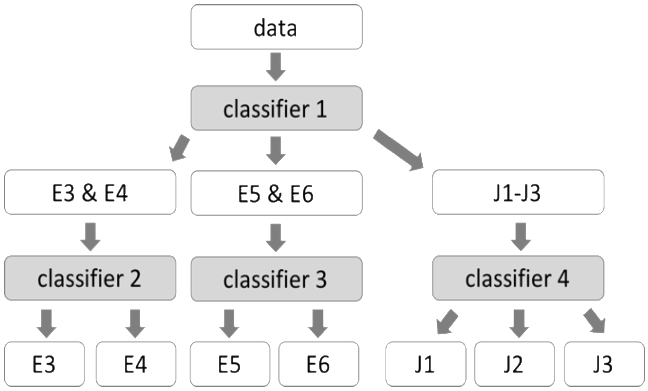

Figure 7: Two-tier classification for Finland (expanded)

B. *Result*

The results of one-tier and two-tier classifications are respectably shown in Tables 12 and 13. Table 14 shows the final result of the two-tier classification. Table 15 shows the comparison of the accuracy and the F-measures of one-tier and two-tier classifications. Feature subsets selected by the algorithm are listed in bold in Fig. 8.

Table: 12 Result of the one-tier classification

		Actual grade						
		E3	E4	E5	E6	J1	J2	J3
Predicted Grade	E3	11	9	2	0	0	0	0
	E4	4	12	3	0	0	0	0
	E5	2	2	16	8	2	0	2
	E6	1	1	5	12	3	4	1
	J1	0	0	1	2	9	3	2
	J2	0	0	0	4	4	14	7
	J3	0	0	1	1	3	6	20

Table13: Result of the 1st stage classification

		Actual grades		
		E3&E4	E5&E6	J1-J3
Predicted grades	E3&E4	38	5	0
	E5&E6	4	44	6
	J1 -J3	0	6	74

Table14: Result of the 2nd stage classification

		Actual grade						
		E3	E4	E5	E6	J1	J2	J3
Predicted grade	E3	14	3	2	0	0	0	0
	E4	3	18	3	0	0	0	0
	E5	0	1	18	9	3	1	0
	E6	1	2	3	14	1	0	1
	J1	0	0	1	0	8	7	5
	J2	0	0	1	3	3	15	5
	J3	0	0	0	1	6	4	21

Table15: Comparison of the result of experiment2

classifier	accuracy (%)	F-measure
One-tier	53.107	0.531
1st / Two-tier	88.136	0.881
2nd / Two-tier	61.017	0.609

One-tier	Two-Tier			
	classifier 1	classifier 2	classifier 3	classifier 4
Total letters	Total letters	Total letters	Total letters	Total letters
Total letter types	Total letter types	Total letter types	Total letter types	Total letter types
Total words	Total words	Total words	Total words	Total words
Total word types	Total word types	Total word types	Total word types	Total word types
Total sentences	Total sentences	Total sentences	Total sentences	Total sentences
Average word length	Average word length	Average word length	Average word length	Average word length
words / sentence	words / sentence	words / sentence	words / sentence	words / sentence
sentences / paragraph	sentences / paragraph	sentences / paragraph	sentences / paragraph	sentences / paragraph
words / word types	words / word types	words / word types	words / word types	words / word types
comma / sentence	comma / sentence	comma / sentence	comma / sentence	comma / sentence
average syllables	average syllables	average syllables	average syllables	average syllables
average syllables * 84.6	average syllables * 84.6	average syllables * 84.6	average syllables * 84.6	average syllables * 84.6

Fig. 8: Comparison of feature subsets

C. *Discussion*

Compared to one-tier classification, two-tier classification yields better result with accuracy higher than 60% and F-measure larger than 0.6. This result shows the effectiveness of two-tier classification. Also, the first stage of two-tier classification results in F-measure of 0.881. This high accuracy suggests that this classifier can accurately classify the difficulty of English sentences.

Fig. 8 lists feature subsets for classifiers used in one-tier and two-tier classifications. Every classifier uses total number of sentences. This result suggests that the number of sentences used in 20 paragraphs changes according to the difficulty level of the textbooks. Also, the feature subsets used in each classifier are different. This result shows the necessity to use multiple classifiers.

VII. EXPERIMENTS BY USING TEXTBOOKS OF OTHER COUNTRIES

A. *Outline*

In this section, the methods used in the previous experiments are applied to the datasets from textbooks of Japan (subsections B to D) and South Korea (E to G).

B. *Experiment using textbooks of Japan*

One-tier and two-tier classifications are conducted by using Japan textbook data. Dataset are generated by using text based on 20 paragraphs. Table 16 shows the number of instances for each grade. Fig. 9 shows the process of two-tier classification.

Table 16: Instances for each grade (Japan)

Grade	instances
J1	11
J2	11
J3	8
H1	8
H2	13
H3	12

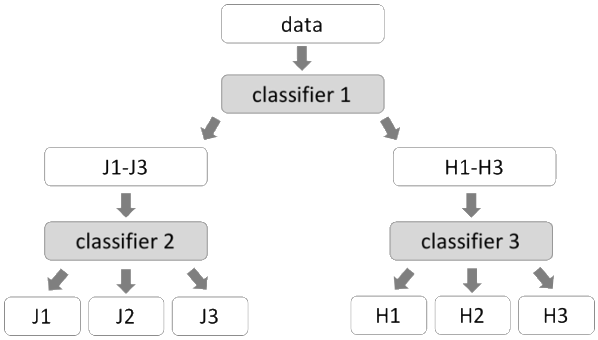

Fig. 9: Two-tier classification for Japan and Korea

C. Result

Table 17 shows the result of one-tier classification. The first stage of two-tier classification classifies all the data accurately as shown in table 18. Table 19 shows the final result of the two-tier classification. Table 20 shows the comparison of the accuracy and the F-measures of one-tier and two-tier classifications. Two-tier classification shows higher result: 17.905 points higher accuracy and 0.028 higher F-measure. Feature subsets used by each classifier are listed in bold in Fig. 10.

D. Discussion

Out of 4 classifiers, only the classifier 1 uses "comma per sentence" feature. This result suggests during the compulsory educational years of junior high school, the number of commas per one sentence remains relatively stable. When the high school year starts, this feature jumps to a higher level. This non-linear change in the number of commas in one sentence means the average length of one sentence in high school textbook is significantly longer than that in junior high textbook. The longer sentence is difficult to read and understand for a non-native learner.

Table 17: Result of one-tier stage classification

		Actual grade					
		J1	J2	J3	H1	H2	H3
Predicted grade	J1	11	1	0	0	0	0
	J2	0	10	2	0	0	0
	J3	0	0	6	0	1	0
	H1	0	0	0	6	3	1
	H2	0	0	0	1	5	3
	H3	0	0	0	1	4	8

Table 18: Result of 1st stage of two-tier classification

		Actual grades	
		J1 -J3	H1-H3
Predicted grades	J1-J3	30	0
	H1-H3	0	33

Table 19: Result of 2nd stage of two-tier classification

		Actual grade					
		J1	J2	J3	H1	H2	H3
Predicted grade	J1	10	1	0	0	0	0
	J2	0	10	2	0	0	0
	J3	0	0	6	0	0	0
	H1	0	0	0	6	2	1
	H2	0	0	0	2	9	5
	H3	0	0	0	0	2	6

Table 20: Results comparison: one-tier vs two-tier

Classifier	accuracy (%)	F-measure
One-tier	73.016	0.721
1st / Two-tier	100.000	1.000
2nd / Two-tier	75.806	0.749

One-tier	Two-Tier		
	classifier 1	classifier 2	classifier 3
Total letters	Total letters	**Total letters**	Total letters
Total letter types	Total letter types	Total letter types	Total letter types
Total words	Total words	Total words	Total words
Total word types	Total word types	Total word types	Total word types
Total sentences	**Total sentences**	Total sentences	Total sentences
Average word length	**Average word length**	Average word length	**Average word length**
words / sentence	**words / sentence**	words / sentence	**words / sentence**
sentences / paragraph	sentences / paragraph	sentences / paragraph	**sentences / paragraph**
words / word types	**words / word types**	words / word types	words / word types
comma / sentence	**comma / sentence**	comma / sentence	comma / sentence
average syllables	average syllables	**average syllables**	**average syllables**
average syllables * 84.6	average syllables * 84.6	average syllables * 84.6	average syllables * 84.6

Fig. 10: Comparison of feature subsets

E. Experiment using textbooks of South Korea

One-tier and two-tier classifications are conducted by using South Korea textbook data. Datasets are generated by using text based on 20 paragraphs. Table 21 shows the number of instances for each grade. Two-tier classification is run in the process illustrated in Fig. 9, same as in the previous experiment.

Table 21: Instances for each grade (South Korea)

Grade	Instances
J1	22
J2	20
J3	21
H1	17
H2	17
H3	13

F. Result

Table 22 shows the result of one-tier classification. Table 23 shows the result of first stage classification of two-tier experiment. Table 24 shows the final result of the two-tier classification. Table 25 shows the comparison of the accuracy and the F-measures of one-tier and two-tier classifications. Two-tier classification shows higher result: 27.926 points higher accuracy and 0.045 higher F-measure. Feature subsets used by each classifier are listed in bold in Fig. 11.

Table 22: Result of one-tier classification

		Actual grade					
		J1	J2	J3	H1	H2	H3
Predicted grade	J1	18	4	3	0	0	0
	J2	2	13	2	0	3	0
	J3	1	2	14	2	2	3
	H1	0	0	1	10	3	6
	H2	1	1	0	4	8	2
	H3	0	0	1	1	1	2

Table 23: Result of 1st stage of two-tier classification

		Actual grades	
		J1-J3	H1-H3
Expected grades	J1-J3	60	4
	H1-H3	3	43

Table 24: Result of 2nd stage of two-tier classification

		Actual grade					
		J1	J2	J3	H1	H2	H3
Predicted grade	J1	19	2	2	0	1	0
	J2	2	14	7	0	2	0
	J3	1	3	10	1	0	0
	H1	0	0	2	8	1	4
	H2	0	1	0	4	12	2
	H3	0	0	0	4	1	7

Table 25: Result comparison: one-tier vs two-tier

classifier	accuracy (%)	F-measure
One-tier	59.091	0.575
1st / Two-tier	93.636	0.936
2nd / Two-tier	63.636	0.631

One-tier	Two-Tier		
	classifier 1	classifier 2	classifier 3
Total letters	Total letters	Total letters	Total letters
Total letter types	Total letter types	Total letter types	Total letter types
Total words	Total words	Total words	Total words
Total word types	Total word types	Total word types	Total word types
Total sentences	Total sentences	Total sentences	Total sentences
Average word length	Average word length	Average word length	Average word length
words / sentence	words / sentence	words / sentence	words / sentence
sentences / paragraph	sentences / paragraph	sentences / paragraph	sentences / paragraph
words / word types	words / word types	words / word types	words / word types
comma / sentence	comma / sentence	comma / sentence	comma / sentence
average syllables	average syllables	average syllables	average syllables
average syllables * 84.6	average syllables * 84.6	average syllables * 84.6	average syllables * 84.6

Figure 11: Comparison of feature subsets

G. *Discussion*

Similar to the previous experiment using Japan textbooks, only the classifier 1 which classifies junior high and high school uses "comma per sentence" feature. This result suggests the number of commas used in one sentence does not increase gradually but leaps from junior high years to high school years.

This non-linear change means the sentences used in the high school textbook are longer and/or more complicated than those used in junior high, making it difficult to comprehend for non-native English learners. This finding shows that until the compulsory education ends, the difficult of the textbook is more strictly controlled compared to the textbooks used for post-compulsory education.

VIII. CONCLUSION

In this study, we develop a system which can classify English sentences according to difficulty level by using features of dataset generated from school textbooks. The purpose of our study is to assist English learners to find appropriate level of reading materials. The contributions of our study are as follows:

- We propose a new method to set paragraph as a unit of analysis for one instance when building dataset. In order to find appropriate number of paragraphs for better classification, an experiment is run by making 5 datasets with a range of paragraphs from 5, 10, 15, 20 and 25. The result shows 20 paragraphs yields the highest accuracy. The proposed method also leads to more accurate classification compared to the existing study which employs page as a unit of analysis, suggesting the effectiveness of using the paragraph as a unit.
- We propose a two-tier classification method which improves classification accuracy compared to the existing one-tier method. The first stage of two-tier classification shows considerably higher accuracy in the experiments.
- We expand the years and countries for the investigation to build a better classifier. Also, this expansion leads to a new finding about the selected feature which could shed light on the educational policy of Japan and South Korea.

For a future research, following three points are worth exploring:

- Random Forest is used for every experiment. Compared to this algorithm, SVM is better suited to binary classification. Hence, for the future research, more accurate classification can be achieved by employing SVM in the cases for binary classification in two-tier methods.
- Maximum range of grades used to build datasets are 7 years: from E3 to J3 in Finland. Difficulty level of English sentences continues to advance with the advancement of the grade. It is necessary to widen the range in order to develop a better system.
- Several feature subsets are generated which allow more accurate classification. By analyzing these subsets, new findings can be obtained regarding how the sentences or structures would change in the process of the rise of difficulty level.

APPENDIX

Readability is defined as "scores calculated by combining factors which make sentence easier to read, such as words difficulties and length, and sentence length, and by substituting it to the formula. The scores are used to find appropriate school level for reading" [12]. Following are partial list of the readability scores:

- •Flesch Readability Score
- •Flesch-Kincaid Grade Level
- •Gunning's Fog Index
- •SMOG Formula
- •FORECAST Readability Formula
- •Powers-Summer-Kearl Formula
- •Fry Index

Among these scores, Flesch Readability Score (FRS) is the most widely used. The score is calculated as follows:

$$FRS = 206.835 - (1.015 \times \alpha) - (84.6 \times \beta)$$

where α= average number of words per one sentence, and β = average number of syllables per one word.

Flesch-Kincaid Grade Level (FKG) is a score obtained by adopting FRS to the level of the grade in the United States. Various readability scores use the number of syllables to calculate the score. This feature has significance in determining difficulty level of English text.

REFERENCES

[1] Ministry of Education, Culture, Sports, Science and Technology (hereafter MEXT), "Countries using English as official or semi-official language," 2006, http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo3/004/siryo/attach/1379959.htm. (accessed on 2018/5/21)

[2] MEXT, "Support for Super Global Universities," 2014a http://www.mext.go.jp/a menu/koutou/kaikaku/sekaitenkai/1360288.html. (accessed on 2018/5/21)

[3] MEXT, "Super Global High Schools," 2014b, http://www.mext.go.jp/a_menu/kokusai/sgh/(accessed on 2018/5/22)

[4] M. Sasaki, "English as in-house language: five years after introduction: Do staff of Fast Retailing speak English fluently?" *Diamond Online*, http://president.jp/articles/-/21640. (accessed on 2018/5/22)

[5] K. Chujo, C. Nishigaki, M. Yamaho, and K. Amano, "Identifying the suitability of textbook English for beginner-level corpus data," *Journal of the College of Industrial Technology of Nihon University*, Vol.44(B), 2011, 13-23.

[6] K. Chujo, A. Shirai, M. Utiyama, C. Nishigaki, and S. Hasegawa, "A study on classifying texts in English-Japanese parallel corpora according to linguistic difficulty," *Journal of the College of Industrial Technology of Nihon University*, Vol.37(B), 2004, 57-68.

[7] L. Wang, "The prospect and challenge of textbook research in English language education," *Bulletin of the Graduate School of Education, University of Tokyo*, Vol.53, 2013, 247-254

[8] H. Ban and T. Oyabu, "Text mining of English textbooks in Finland," *Proceedings of the Asia Pacific Industrial Engineering & Management Systems Conference*, 2012, 1674-1679.

[9] H. Ban, R. Oguri, and H. Kimura, "Difficulty-level classification for English writings," *Transactions on Machine Learning and Artificial Intelligence*, Vol. 3, No 3, 2015, 24-32

[10] L. Breiman, "Bagging predictors," *Machine Learning*, Vol. 24, No.2, 1996, 123-140.

[11] L. Breiman, "Random Forests," *Machine Learning*, Vol. 45, No. 1, 2001, 5-32.

[12] Y. Takanashi, and Y. Ushiro, *Handbook of English Reading*, Kenkyu-sha, 2000.

Toshihiko Shimauchi, Department of Regional Design and Development, Komatsu College, Komatsu, Ishikawa, Japan.

Ryo Oguri, Rinnai Corporation.

Hiromi Ban, Graduate School of Engineering, Nagaoka University of Technology, Nagaoka, Niigata, Japan.

Hidetaka Nambo, Graduate School of Natural Science and Technology, Kanazawa University, Kanazawa Japan.

Haruhiko Kimura, Faculty of Production Systems Engineering and Sciences, Komatsu University, Komatsu, Ishikawa, Japan