Classification and clustering English writing errors based on native language

Flanagan, Brendan
Graduate School of Information Science and Electrical Engineering, Kyushu University

Yin, Chengjiu
Innovation Center for Educational Resource (ICER), Kyushu University Library, Kyushu University

Suzuki, Takahiko
Research Institute for Information Technology, Kyushu University

Hirokawa, Sachio
Research Institute for Information Technology, Kyushu University

出版情報：Proceedings - 2014 IIAI 3rd International Conference on Advanced Applied Informatics, IIAI-AAI 2014, pp.318-323, 2014-01-01. Institute of Electrical and Electronics Engineers Inc.
バージョン：
権利関係：

Classification and Clustering English Writing Errors Based on Native language

Brendan Flanagan
Graduate School of Information Science and Electrical Engineering, Kyushu University
Fukuoka, Japan
b.flanagan.885@s.kyushu-u.ac.jp

Chengjiu Yin
Innovation Center for Educational Resource (ICER), Kyushu University Library Kyushu University
Fukuoka, Japan
yin.academic@gmail.com

Takahiko Suzuki, Sachio Hirokawa
Research Institute for Information Technology, Kyushu University
Fukuoka, Japan
{suzuki, hirokawa}@cc.kyushu-u.ac.jp

***Abstract*— It is important for language learners to determine and reflect on their writing errors in order to overcome weaknesses. Each language learner has their own unique writing error characteristics and therefore has different learning needs. In this paper, we analyze the writing errors of foreign language learners on the language learning SNS website Lang-8 to investigate the characteristics of errors by native language. 142,465 sentences were collected from Lang-8 for analysis. For each native language, the predicted scores of 15 error categories from SVM machine learning models are used as a vector representation of each sentence. These score vectors are then clustered to determine error co-occurrence within the same sentence. The results were then analyzed to determine the error characteristics of different native languages.**

Keywords— Language learning; native language characteristics; writing error categories; machine learning.

I. INTRODUCTION

In order to overcome mistakes, learners need feedback to prompt reflection on their errors [1]. This is a particularly important issue in education systems, as the system effectiveness in finding errors or mistakes could have an impact on learning. Finding errors is essential to providing appropriate guidance in order for learners to overcome their flaws. Traditional classroom-based language study has offered interaction with other learners and feedback from teachers and peers.

In the last decade or so with the global spread of the Internet, the number of people studying languages on the web has increased. Of particular interest are sites that offer a social or collaborative approach to study languages, and are often based on a SNS (Social Networking Service) platform. To some extent these SNS-based websites offer feedback and interaction that might otherwise be absent in autonomous learners studies. Language learning SNS sites work on the language exchange function, where native speakers of the target language offer corrections and feedback to the language learners. In principal, these learners would then correct the writings of a learner studying their native language. For example, person A is a native Japanese speaker who is learning English as a foreign language and posts an English sentence on the website. Person B who is a native English speaker corrects the sentence. Person B is also learning Japanese as a foreign language and posts a sentence on the website in Japanese which is then corrected by person A. This mutually beneficial environment helps learners to achieve their respective goals of learning a foreign language, which in turn is another foreign language learner's mother tongue.

These websites contain numerous foreign language writings that have been created by learners and corrected by speakers of the target language. It could be thought of as a crude crowd-sourced foreign language writing parallel corpus. Taking advantage of this data can help to further enhance the effectiveness of language learning through providing automated feedback and guidance. We have investigated an automated online quiz generation system in previous research [2] for learners to reflect and practice on their errors. Initially error categories for the quiz system had to be manually classified. To overcome this problem the use of a SVM classifier to predict error categories [3,4] and the clustering of predicted error category score vectors for overall co-occurrence analysis [5] was investigated.

In this paper, the writings, in particular the diaries, of language learners on the website Lang-8 are analyzed to investigate the error characteristics of English foreign language writing by native language. Previously we have predicted the scores of 142,465 sentences collected from Lang-8 by using 15 SVM error category models [5]. These scores will be used as a vector representation of the sentences and divided into data subsets by native language of the learner as reported on Lang-8. The subsets will then be clustered to analyze the co-occurrence and independence of foreign writing errors based on the native language of the learner.

II. Related Work

A. English Writing Error Categories and Corpora

Previous empirical studies on the writings of foreign language students have been undertaken in academic settings to enabled the control of influencing factors, such as: subject and environment. Kroll [6] compared the difference of writings that were conducted in a classroom where learners had a fixed amount of time, and the home environment, where they would have more time and less pressure. English teachers categorized errors manually and the frequency of occurrence was used to compare the writings in the two different environments. Weltig [7] looked at the effect of different categories of errors on the scoring given by English teachers for the writings of foreign language learners. Using similar error categories as Kroll [6], it was found that the frequency of certain error categories had more of an influence on the overall score than others. The sample data in this paper was prepared for machine learning by using similar categories to Kroll and Weltig for manually identifying errors in sample pair sentences from Lang-8. Language learners in academic settings have access to language specialists such as teachers that can provide analysis and corrective feedback, however this is not as readily available to autonomous learners. To fill this gap, we have a goal of creating tools for these learners to enable them to a certain extent to be given some feedback and analysis similar to that provided by language specialists.

Sugiura et al. [8], discuss corpus design and reviewed the International Corpus of Learner English (ICLE). Based on the corpus weaknesses identified, they set about compiling a new English learner corpus and a parallel corpus of native English speakers, called NICE (Nagoya Interlanguage Corpus of English). Using this corpus they performed analysis using mechanical text features, such as: type, token, number of sentences, and average word length to compare the language learners performance with native speakers.

Miki [9] looks at the use of a parallel corpus that is constructed using the essay writings of foreign language learners and exact forms of the sentences that are provided by native English language speakers. NICE was used as a dataset to examine how Japanese English language learner's use "I think" in comparison with native speakers. Unlike other studies on the over usage of expressions which focus on quantifying the errors, by using a parallel corpus they were able to determine how the expression was being inappropriately used to augment the language learners writing. Miyake et al. [10] also used the same method and NICE parallel corpus to examined the use of "there" with the long-term intention of identifying the "Japaneseness" and "nativeness" relating to the use of constructions.

B. SVM Error Categorization

Previous studies have estimated errors in English text by using SVM and other types of machine learning algorithms. Hirano et al. [11] investigated the use of search engine results to detect article errors in English technical papers. The sentences were syntactically parsed to produce a parts of speech tagged sentence, and then a search query was created based on the structure of the sentence. The number of hits from the resulting search query was then counted and used to determine if the input sentence contained an error. Tanimoto et al [12] examined using the number of search results as a indicator in an attempt to identify erroneous words in English sentences. NICE (Nagoya Interlanguage Corpus of English) was used in tri-grams and 4-grams as training data for SVM machine learning to create a model that can determine if an English sentence contains an error.

III. Clustering based on Error Category Prediction

A. Data Collection

Two sets of data were collected for analysis: a set of 142,465 sentences, posted on Lang-8 from October 9, 2011 to January 6, 2012, which are written in English and are corrected in some way. Each sentence is tagged to identify the native language of the author. Figure 1 shows the major groups within the collected data set, and it can be seen that Japanese users wrote roughly 100,000 sentences, which is the largest subset in the collected data.

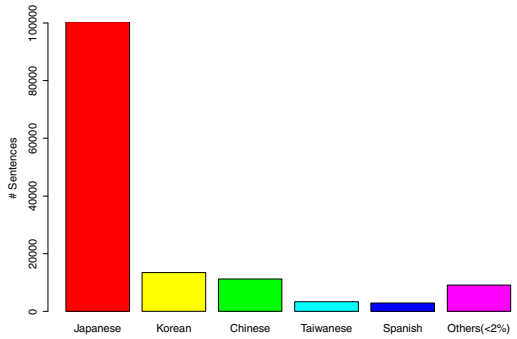

Fig. 1. Number of sentences grouped by native language.

Other main subsets include: Korean, Chinese, Taiwanese and Spanish. The analysis in this paper will focus on these five major native language subsets within the collected data.

TABLE I. Error Category Numbers and Descriptions

Category	Description
2	Subject formation
3	Verb missing
6	Dangling/misplaced modifier
11	Word order
13	Extraneous words
17	Tense
19	Verb formation
25	Ambiguous/unlocatable referent
28	Lexical/phrase choice
30	Word form
33	Singular for plural
36	Preposition
37	Genitive
38	Article
42	Spelling

The second set consists of 399 corrected sentence pairs that have been manually classified into error categories. We analyzed the data to train and test SVM classifiers for 15 error categories in previous research [3-5]. The error category numbers and descriptions that were used are shown in Table 1.

10 models constructed for each error category. Each sentence from the first set of data was scored with respect an error category is calculated as the average of the score of the result obtained by applying the 10 models. 15 scores corresponding to 15 error categories form a vector representation of a sentence.

B. *Error Co-occurrence Analysis by Clustering*

We have previously investigated the co-occurrence of errors from an overall perspective [5], and did not take into account other factors, such as the learners' native language, etc. The score vector representations of the sentences were analyzed by clustering into 20 clusters using the high-dimension clustering tool CLUTO [13]. This research identified co-occurring and non-co-occurring errors, an overview of which can be seen in the dendrogram (clustering tree) in Figure 2. The darker colored squares represent clusters of sentences with high averages in parts of the score vector. For example, cluster 0 has a high average score for error category 38 (Article errors), and cluster 3 has a high average score for error category 36 (preposition errors). On the vertical axis is a cluster hierarchy tree of the 20 resulting clusters. The clusters are leaf nodes of the tree where the number of sentences in the cluster is represented in the brackets next to the cluster ID. The error categories are represented on the horizontal axes at a cluster hierarchy tree. This visualization is very helpful to understand the huge amount of target data. We can see that the lower branch of the tree contains 1/3 of all the data and corresponds mainly to error category 38 (article errors). Cluster 0 is a core part of this branch whose sentences contain mainly article errors. Cluster 17 contains lexical or phrase choice errors (category 28). The cluster 16 contains preposition errors (category 36). Thus, the tree represents not only the clustering of sentences but also the clustering of error categories. Indeed, we can interpret that article errors (category 38) are the largest errors and occur with preposition error (category 36) and lexical/phrase choice error (category 28). For further details please refer to [5].

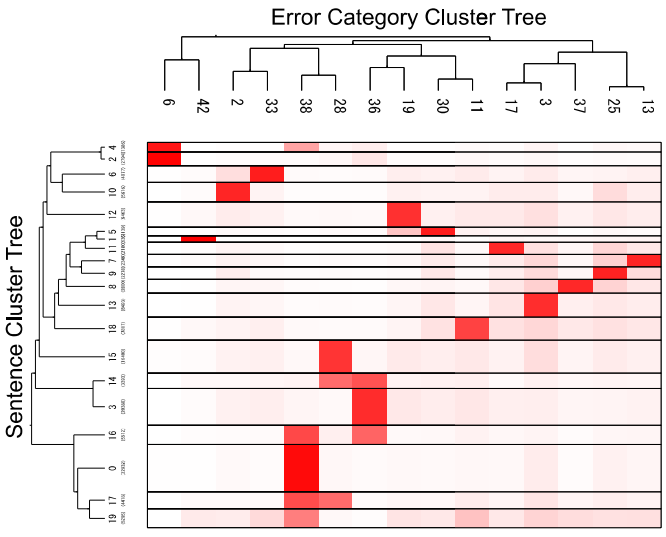

Fig. 2. Clustering of all writing error data

While presenting this previous research, it was recommended by an attendee that it is important to perform the analysis of error characteristics with regard to the native language of the learner. Therefore this paper investigates co-occurring error categories by native language.

IV. Co-occurrence Analysis By Native Language

A. *Principal Component Analysis (PCA)*

To investigate if there are any underlying correlations between native languages and the predicted error category scores, we analyzed the score vector and the native language of the learners using Principal Component Analysis (PCA).

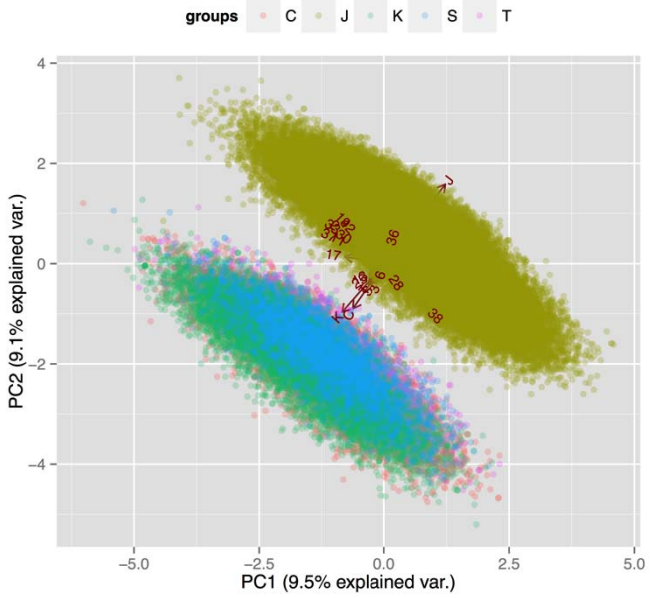

Fig. 3. Principal Component Analysis of all writing error data

The results of the PCA are shown in Figure 3, where there is a slight association between error category 36 (preposition) and Japanese when compared to other languages: Korean, Chinese, Taiwanese and Spanish. However other than this observation there are no significant correlations between error categories and native languages.

B. *Error Co-occurrence Analysis of Native Languages by Clustering*

The dataset of predicted error category score vectors was divided into subsets based on the native language of the learner. The top five native languages by number of sentences were then clustered into 20 clusters to analyze possible differences in error co-occurrence. Figure 4 shows the clustering results for sentences written by Japanese natives.

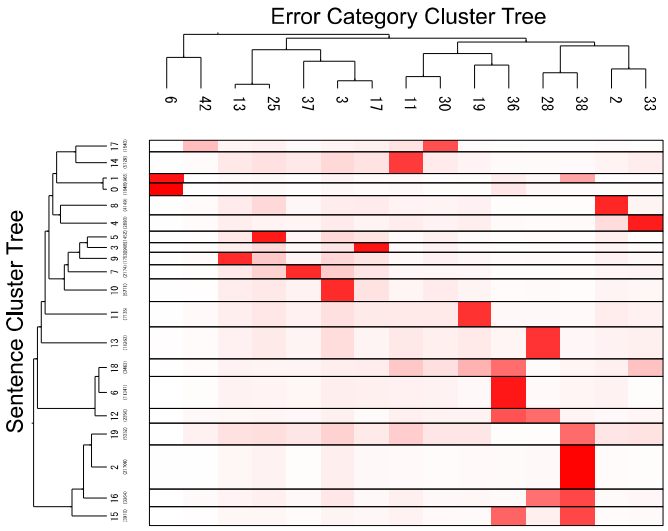

Fig. 4. Clustering of writing errors by Japanese natives

Clusters 15 and 16 are both made up of sentences that contain error categories 36 (preposition) and 28 (lexical/phrase choice) respectively that co-occur with error category 38 (article). Clusters 12 and 18 are also both made up of sentences that contain error category 36 (preposition) that co-occur with 19 (Verb formation) and 28 (Lexical/phrase choice). Other notable co-occurrences are seen in cluster 1 which contains error category 6 (Dangling/misplaced modifier) and 38 (article), and cluster 17 which contains error category 42 (spelling) and 30 (word form).

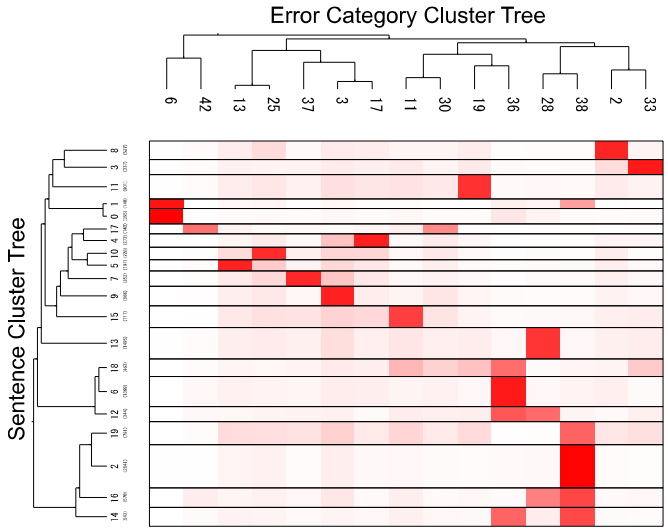

Fig. 5. Clustering of writing errors by Korean natives

The results of the clustering analysis of Korean natives, as shown in Figure 5, share some similarities with Japanese natives. It contains the same co-occurring errors as Japanese, except instead of error categories 19 (Verb formation) and 36 (preposition) co-occurring, Korean has a more prominent co-occurrence between error category 11 (word order) and 36 (preposition).

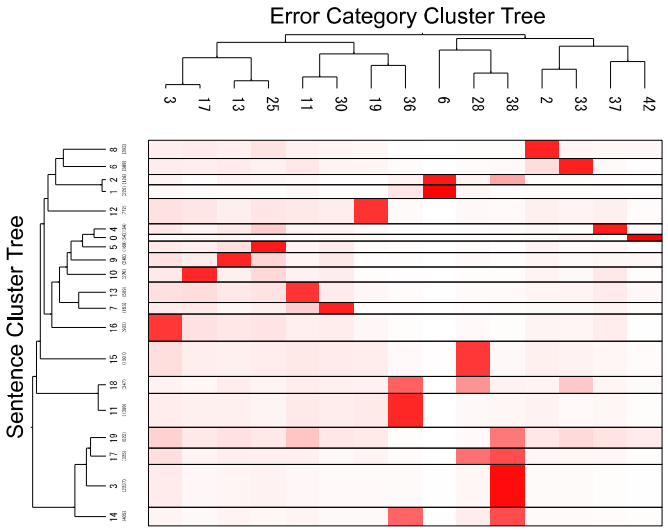

Fig. 6. Clustering of writing errors by Chinese natives

For Chinese natives the results displayed in Figure 6 have some similarities to the results for Korean and Japanese as error category 38 (article) co-occurs with 6 (Dangling/misplaced modifier), 28 (Lexical/phrase choice), and 36 (preposition), and also error category 28 (Lexical/phrase choice) with 36 (preposition).

Fig. 7. Clustering of writing errors by Taiwanese natives

As with all the preceding results, sentences written by Taiwanese natives, as seen in Figure 7, have error category 38 (article) co-occurring with 6 (Dangling/misplaced modifier), 28 (Lexical/phrase choice), and 36 (preposition), and also error category 28 (Lexical/phrase choice) with 36 (preposition). In addition it can be seen that error categories 3 (Verb missing), 13 (extraneous words), and 25 (ambiguous/unlocatable referent) co-occur in the same sentences.

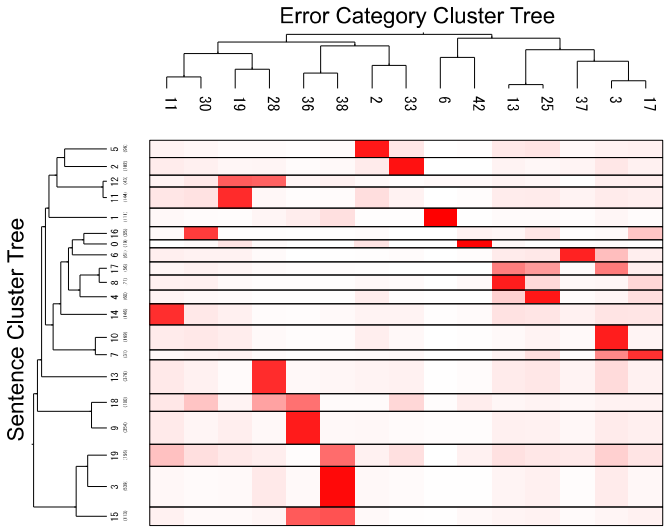

Fig. 8. Clustering of writing errors by Spanish natives

In contrast with the previous results, sentences by Spanish natives that contain error category 38 (article) co-occur only with 36 (preposition) as displayed in Figure 8. Another similarity is the co-occurrence of error category 28 (Lexical/phrase choice) and 36 (preposition). There are some similarities with Taiwanese in that they share the same error co-occurrence between error categories 3 (Verb missing), 13 (extraneous words), and 25 (ambiguous/unlocatable referent). A co-occurrence that is unique to Spanish is that between 19 (Verb formation) and 28 (Lexical/phrase choice), and 3 (Verb missing) and 17 (Tense).

C. *Analysis by Tree Distance*

This analysis aims to identify differences in error categories between native languages. Initial analysis was conducted for similar and different category distances, however it was found that a large number of error categories shared similar tree distances across most of the native languages. Therefore we will focus on the difference of error category tree distances in this analysis.

$$\max\left\{\left|d(C_i, C_j, N_p) - d(C_i, C_j, N_q)\right|\right\} \quad (1)$$

The tree distance between two leaves of the error category cluster tree was calculated to investigate the difference in error category tree distances. Formula 1 was used to search for the errors with the greatest difference between languages, where $N_p \neq N_q$ and $C_i \neq C_j$, and $d(C_i, C_j, N_p)$ is the distance between the nodes of error category C_i and C_j for the native language N_p.

The results of this analysis are shown in Table 2 as a distance matrix. It can be seen that there are no discernable differences in the distances of error category nodes between Japanese and Korean. This would suggest that the errors of Japanese and Korean learners have similar characteristics. The difference between Taiwanese and Spanish is also low, only differing in the distance between errors 17 (tense) and 19 (verb formation). These two similar groups seem to be at extremes as they have the greatest number of differences. Chinese has two distances that are different when compared to all the other languages. The differences between Chinese and the Taiwanese/Spanish group are error categories 19 (verb formation) and 36 (preposition), along with 36 (preposition) and 38 (article). The Japanese/Korean group has different distances to Chinese in error categories 3 (verb missing) and 37 (genitive), and 17 (tense) and 37 (genitive).

TABLE II. ERROR CATEGORY DIFFERENCE OF NATIVE LANGUAGES BASED ON TREE DISTANCE.

	J	K	C	T	S
J		NA	3,37; 17,37;	3,37; 17,19; 19,36;	19,28; 19,36; 28,38; 36,38;
K	NA		3,37; 17,37;	3,37; 17,19; 19,36;	19,28; 19,36; 28,38; 36,38;
C	3,37; 17,37;	3,37; 17,37;		19,36; 36,38;	19,28; 36,38;
T	3,37; 17,19; 19,36;	3,37; 17,19; 19,36;	19,36; 36,38;		17,19;
S	19,28; 19,36; 28,38; 36,38;	19,28; 19,36; 28,38; 36,38;	19,28; 36,38;	17,19;	

V. CONCLUSION AND FUTURE WORK

The present paper clustered the predicted scores of the writing error categories of 142,465 sentences by English learners' writing from the language learning SNS web site Lang-8. To investigate the differences in error characteristics of native languages, the data was divided into subsets based on the native language of the learner. These subsets were then each clustered based on the predicted error category score vectors. The clustering results of five major subsets: Japanese, Korean, Chinese, Taiwanese and Spanish were then analyzed and compared to determine the error characteristics of the native languages. All of the native languages have two co-occurring errors in common (28 and 36, 36 and 38). Asian languages have the co-occurring errors 6 and 38, and 28 and 38 in common. Japanese and Korean also have co-occurring errors 30 and 42 in common. Taiwanese and Spanish have three error categories 3, 13, and 25 that co-occur. The error categories of each of the results was clustered and analyzed by the tree distance between nodes. There were no observed differences between Japanese and Korean. For Taiwanese and Spanish, only the distance between errors 17 and 19 is different, suggesting a degree of similarity. Chinese has two different error pairs difference when compared with all other languages. These results suggest that the error characteristics of Japanese and Korean learners are quite similar, as are those of Taiwanese and Spanish learners to a lesser degree. These differences in co-occurring errors are characteristic of the learner's native language. This could be used in teaching and learning to focus on co-occurring errors that are characteristic

of their native language. In future work we will undertake detailed analysis and evaluate the validity of the results and investigate the different error characteristics of native languages for non-co-occurring errors categories and the words that are features of these errors.

ACKNOWLEDGMENT

This work was partially supported by JSPS KAKENHI Grant Number 24500176.

REFERENCES

[1] Lo, J. J., Wang, Y. C., Yeh, S. W., "WRITE: Writing Revision Instrument for Teaching English." Technologies for E-Learning and Digital Entertainment (2008), pp.1-8.

[2] Flanagan, B., Yin, C., Hirokawa, S., Hashimoto, K., Tabata, Y., An Automated Method to Generate e-Learning Quizzes from Online Language Learner Writing, International Journal of Distance Education Technologies (to be published)

[3] Flanagan, B., Yin, C., Suzuki, T., Hirokawa, S., Intellegent Computer Classification of English Writing Errors, Proc. KES-IIMS2013, (2013), pp.174-183.

[4] Flanagan, B., Yin, C., Suzuki, T., Hirokawa, S., Classification of English Language Learner Writing Errors Using a Parallel corpus with SVM, International Journal of Knowledge and Web Intelligence (to be published)

[5] Flanagan, B., Yin, C., Hashimoto, K., Hirokawa, S., Clustering English Writing Errors based on Error Category Prediction, ISEEE2013, (2013), pp.733-738.

[6] Kroll, B., What does time buy? ESL student performance on home versus class compositions, In B. Kroll (Ed.), Second language writing: Research insights for the classroom, Cambridge: Cambridge University Press (1990), pp.140-154.

[7] Weltig, M. S., Effects of language errors and importance attributed to language on language and rhetorical-level essay scoring, Spaan Fellow Working Papers in Second or Foreign Language Assessment Volume 2 2004, 1001 (2004), pp.53-81.

[8] M. Sugiura, M. Narita, T. Ishida, T. Sakaue, R. Murao, K. Muraki, A Discriminant Analysis of Non-native Speakers and Native Speakers of English. Proceedings of the Corpus Linguistics Conference CL2007, (2008) 27.

[9] [9] N. Miki, A new parallel corpus approach to Japanese learners' English, using their corrected essays. Themes in Science and Technology Education, 3(1-2), (2011) 159.

[10] [10] H. Miyake, T. Tsushima, On the features of there constructions used by Japanese speakers of English, The Journal of Humanities & Natural Sciences, 132, (2012) 55.

[11] [11] T. Hirano, Y. Hirate, H. Yamana, Detecting Article Errors in English using Search Engines, DBSJ Letters 6.3, (2007) 13. (in Japanese)

[12] [12] T. Tanimoto, M. Ohta, Examination of English Error Detection Using the Number of Search Results, DEIM Forum 2012, 9.1 (2012). (in Japanese)

[13] Karypis, G. *CLUTO - a clustering toolkit*. No. TR-02-017. Minnesota University, Minneapolis, Dept of Computer Science, 2002.