Journal of Pan-Pacific Association of Applied Linguistics, 20(2), 1-17.

A Corpus-based Approach to the Register Awareness of Asian Learners of English

Yuichiro Kobayashi [*]
Toyo University

Mariko Abe
Chuo University

Kobayashi, Y. & Abe, M. (2016). A corpus-based approach to the register awareness of Asian learners of English. *Journal of Pan-Pacific Association of Applied Linguistics*, *20*(2), 1-17.

The purpose of the present study is to investigate the impact of learners' L1s and proficiency levels on their written production. This study also examined the influence of speech upon their writing. The following research questions were explored: (a) How do L1 and proficiency levels of learners affect their degrees of register awareness? (b) Which linguistic features distinguishing writing and speech registers are characteristic of each Asian learner group? This study draws on four sub-corpora of the International Corpus Network of Asian Learners of English (ICNALE), which is considered to be the largest East Asian composition database. Using the methodology originally developed by Biber (1988) to analyze the differences in the spoken and written registers of English, this study investigated the differences in a wide range of linguistic features among Asian learners of English. The results suggest that the L1s of learners affect the degree of their register awareness. Hong Kong learners display a set of stylistically appropriate features, such as nominalizations, predictive modals, and conjuncts, in their academic prose whereas Japanese learners exhibit many of informal features, such as first person pronouns, private verbs, and independent clause coordination, in their written production. Besides, Korean and Taiwanese learners show several features typical of speech, including second person pronouns, in their writing. In addition, this study demonstrates the effectiveness of Biber's list of linguistic features in the study of spoken nature in L2 writing.

Keywords: register awareness, corpus-based approach, Asian learners of English

---

[*] First author: Yuichiro Kobayashi; second author: Mariko Abe.

[*] An earlier version of the present study was reported at the 11th Teaching and Language Corpora (TaLC) Conference held at Lancaster University on July 22nd, 2014. This work was supported by Grants-in-Aid for Scientific Research Grant Numbers 26770205 and 16H03455.

1 Introduction

The availability of computer learner corpora enables researchers to investigate a vast amount of descriptive data of interlanguage performance. It has led to contrastive interlanguage analysis, which intends to unveil the nonnative characteristics of learner language. Contrastive interlanguage analysis involves two major types of studies: (a) comparison of native language and interlanguage and (b) comparison of different interlanguages (Granger, 1996). The first type of comparison has uncovered *overused* and *underused* linguistic features that distinguish learners from native speakers using statistical tests. The second type of comparison have hatched a new research field called second language (L2) profiling, which aims to describe the developmental patterns of learner language (Meunier, 2015) as well as to identify a set of linguistic features that can be applied to the development of language assessment (Hawkins & Filipović, 2012).

A number of learner corpus studies have underlined the lack of register awareness among L2 learners. Confusing written registers with spoken registers, learners face difficulty in using suitable styles for different production modes. For example, Granger and Rayson (1998) investigated the difference in the use of nine word categories, namely (a) nouns, (b) adjectives, (c) prepositions, (d) articles, (e) determiners, (f) conjunctions, (g) pronouns, (h) adverbs, and (i) verbs, between native speakers and French learners of English, and concluded that French learners exhibited few of the features typical of academic writing and most of those typical of speech. Other studies also detected the influence of speech upon learners' writing with regards to more specific items, such as conjunctions (e.g., Lorenz, 1999), adverbial phrases (e.g., Altenberg & Tapper, 1998; Granger & Petch-Tyson, 1996), and the combination of first person pronouns and private verbs (e.g., Aijmer, 2002; Petch-Tyson, 1998). However, only few attempts have so far been made at investigating the linguistic features characteristic of speech other than lexical items. As Biber (1986) pointed out, limiting the number of linguistic features examined in a corpus-based study can lead to limited results. Thus, it is necessary to address the problem by focusing on other aspects of writing, such as syntax and discourse, pertaining to the lack of register awareness.

Another problem is that, while many previous learner corpus-based studies have focused on European learners of English, fewer studies have targeted East Asian learners of English. Oi (2016) made remarks on the current state of L2 writing research, pointing out the differences between English as Second Language (ESL) and English as Foreign Language (EFL) environment and those between European and Asian countries. In EFL context, learners have less opportunities to use English outside the classroom than ESL context. Besides, in Asian educational settings, *writing to learn language* is required rather than *learning to write about the content*. Given

the possible impact of learners' first language (L1) on L2 performance, it is important to shed light upon the nature and characteristics of Asian learners' English, which might differ from those of European learners. Moreover, since the lack of register awareness is attributable to L1 transfer as well as developmental factors (Gilquin & Paquot, 2008), it is essential to compare multiple learner groups from different L1 backgrounds and developmental stages.

2 Research Questions

The present study investigated the impact of learners' L1s and proficiency levels on their written production. In particular, this study examined the influence of speech upon their writing. The following research questions were explored:

(1) How do L1 and proficiency levels of learners affect their degrees of register awareness?
(2) Which linguistic features distinguishing writing and speech registers are characteristic of each Asian learner group?

By pursuing these research questions, this study aims to address the problems discussed in the previous section. The investigation employed multiple statistical techniques to identify the spoken features of learner writing.

3 Data and Methodology

3.1 Corpus data

The present study draws on four sub-corpora of the International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2011), which is considered to be the largest East Asian composition database. The corpus contains argumentative essays written in response to two different prompts, namely, (a) *It is important for college students to have a part time job* and (b) *Smoking should be completely banned at all the restaurants in the country*. Writers are required to compose their essays using word processing software without the use of dictionaries and other references. They are asked to use an electronic spell-checker before submitting the essay. The essays are required as extra homework for their English classes. The data analyzed in the present study is a subset from this corpus, including the written compositions of 2,000 English as foreign language learners in Hong Kong (HKG), Korea (KOR), Taiwan (TWN), and Japan (JPN). Their Common European Framework of Reference for Languages (CEFR) levels were assessed, varyingly, as A2, B1_1, B1_2, B2, and C1. Table 1 shows the size of each sub-corpus compared in this study.

Table 1. The Numbers of Learners and Words

CEFR level	numbers	HKG	KOR	TWN	JPN	Total
A2	learners	2	150	58	308	518
	(words)	(519)	(33,095)	(12,776)	(67,902)	(114,292)
B1_1	learners	60	122	174	358	714
	(words)	(14,403)	(26,651)	(39,520)	(78,775)	(159,349)
B1_2	learners	104	176	122	98	500
	(words)	(24,530)	(39,823)	(28,286)	(22,135)	(114,774)
B2	learners	30	116	44	34	224
	(words)	(6,969)	(26,964)	(10,397)	(7,895)	(52,225)
C1	learners	4	36	2	2	44
	(words)	(944)	(8,914)	(518)	(529)	(10,905)
Total		200	600	400	800	2,000
		(47,365)	(135,447)	(91,497)	(177,236)	(451,545)

3.2 Linguistic features

Corpus-based analysis can complement the flaw of traditional language studies, which focus on a small number of linguistic features, by conducting more comprehensive descriptions of multiple linguistic features (Biber, Conrad, & Reppen, 1998). For instance, the set of linguistic features selected by Biber (1988), are widely used in corpus-based studies to explore various types of linguistic variation (e.g., Conrad & Biber, 2001; Sardinha & Pinto, 2014). This trend can be applied to learner corpus studies which compare English essays written by learners from different L1 backgrounds (Abe, Kobayashi, & Narita, 2013), describe the developmental patterns of interlanguage (Abe, 2014), and automatically assess L2 spoken performances (Kobayashi & Abe, 2016). Further, since Biber's framework was originally developed for investigating variation across speech and writing, it can be utilized to study learners' register awareness (Aguado-Jiménez, Pérez-Paredes, & Sánchez, 2012). In the present study, 58 linguistic features were selected from the original list of 67 linguistic features in Biber (1988), and they were used to analyze differences between learner groups in the ICNALE. The features can be classified into fifteen major grammatical categories: (a) tense and aspect markers, (b) place and time adverbials, (c) pronouns and pro-verbs, (d) questions, (e) nominal forms, (f) passives, (g) stative forms, (h) subordination, (i) prepositional phrases, adjectives, and adverbs, (j) lexical classes, (k) modals, (l) specialized verb classes, (m) reduced forms and dispreferred structures, (n) coordination, and (o) negation. Nine features: (a) demonstratives, (b) gerunds, (c) present participal clauses, (d) past participal clauses, (e) present participal WHIZ deletion relatives, (f) sentence relatives, (g) type/token ratio, (h) word length, and (i) subordinator-that deletion, were

not included in the present analysis due to differences in the software used to annotate part-of-speech tags. The frequencies of 58 linguistic features were counted using the TreeTagger (Schmid, 1994). The Perl program, which was originally developed for the multi-dimensional analysis of English textbooks by Murakami (2009), were modified by the authors for more accurate processing of L2 performance data.

3.3 Statistical method

This study applied the linguistic feature list used by Biber (1988), but instead of employing factor analysis, we used two multivariate methods, correspondence analysis, and hierarchical cluster analysis, since, as McEnery and Hardie (2012) pointed out, Biber's multi-dimensional analysis, which is based on factor analysis, has been criticized for the difficulty it poses while replicating the findings. In contrast, correspondence analysis can show higher reproducibility than factor analysis because it requires simpler calculation processes and has fewer options to reduce the dimensionality of data. This statistical technique reveals frequency-based associations between corpora and those between variables, and graphically represents them on a two- or three-dimensional scatter plot (Glynn, 2014). The scatter plot is helpful for investigating similarities among corpora and/or variables included in a frequency table. However, it is sometimes arbitrarily interpreted by "an informal way, grouping 'by eye' the points lying one near the other on the plot," and thus, "a more formal method" may be required for the better understanding of the plot (Alberti, 2013, p. 40). In this study, hierarchical cluster analysis was used for interpreting the relationships among sub-corpora in the ICNALE. This method organizes information about how similar items are, so that clusters can be formed (Divjak & Fieller, 2014). Results of the method show tree-like categorizations where small groups of highly similar items are included within much larger groups of less similar items (Oakes, 1998). Furthermore, Cramér's V (Gries, 2014) and z score (Jarvis, Grant, Bikowski, & Ferris, 2003) were checked to supplement the findings of these two multivariate methods. Cramér's V can be used as a keyness, which identifies linguistic features that can distinguish different learner groups, and z score can be used as a measure to investigate frequent or infrequent features in each learner group. All statistical analyses in this study were conducted using R (Ihaka & Gentleman, 1996), a programming language for data analysis and graphics.

4 Results and Discussion

This study began by examining whether language use, as reflected by the 58 linguistic features discriminating written and spoken registers, differed between essays written by learners from different L1 backgrounds and

different proficiency levels. Correspondence analysis allows a visual representation of the similarity between learner groups in a scatter plot. Figure 1 shows the result of correspondence analysis in the two most powerful dimensions, which account for 71.36% of total variation in the frequency table. The coordinates in the diagram reflect the interrelationship between learner groups, and the relative distance between groups indicates the similarity of co-occurrence patterns of 58 linguistic features used for the analysis. Hong Kong learners (HKG) are clustered on the right-hand side of the diagram and Japanese learners (JPN) on the left. Korean learners (KOR) and Taiwanese learners (TWN) were positioned between Hong Kong and Japanese learner groups. These results suggest that L1 of learner groups differ in their use of the linguistic features.

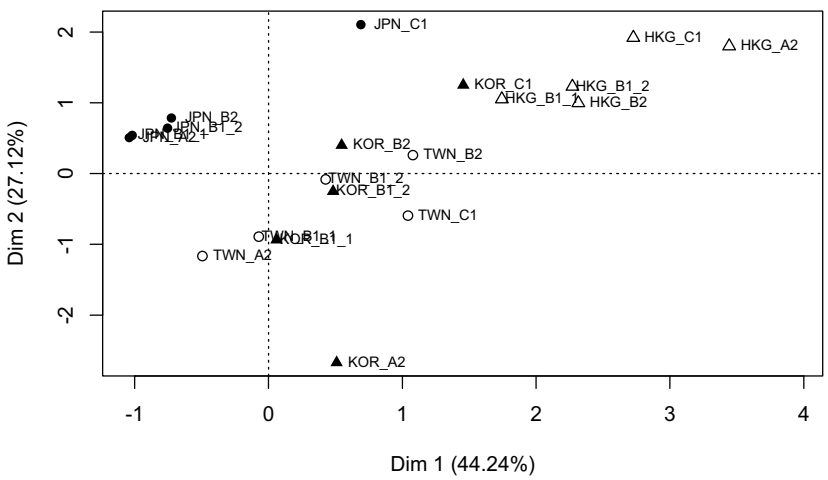

Figure 1. Scatter plot showing the results of correspondence analysis

The grouping of East Asian learners of English can be more clearly seen in the dendrogram representing the results of hierarchical cluster analysis. The results of clustering are displayed in Figure 2, which was obtained from the resulting coordinates of the two strongest dimensions of correspondence analysis. Gower's distance was used for measuring the dissimilarities between learner groups, and Ward's method was selected for forming clusters.

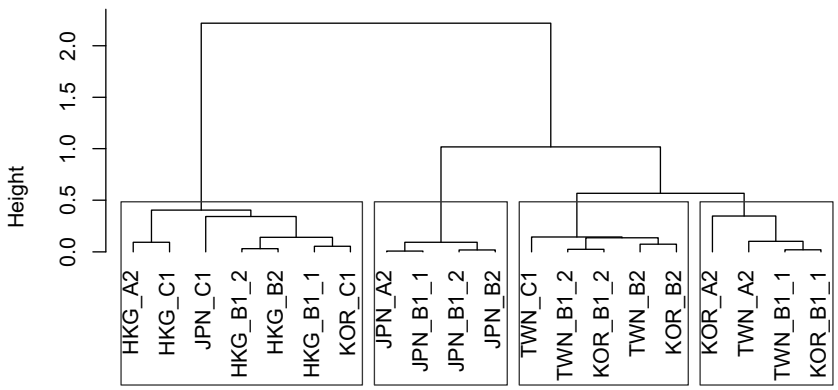

Figure 2. Dendrogram representing the results of hierarchical cluster analysis

In this diagram, the vertical axis represents the distance between each learner group, and the positions of horizontal lines on the scale indicate the distance at which clusters are joined. The number of clusters can be determined by specifying the cutting point on the vertical axis. In the present study, the dendrogram was terminated at the height of 0.5. As a result, East Asian learners were classified into four clusters: (a) all Hong Kong learners, Japanese C1-level learners and Korean C1-levels, (b) Japanese A2- to B2-level learners, (c) Taiwanese B1_2- to C1-level learners and Korean B1_2- to B2-level learners, and (d) Taiwanese and Korean A2- to B1_1-level learners. These results suggest that there is a major influence of L1 on the output of learners from Hong Kong and Japan. While the distinction between Taiwanese and Korean learners is not clear, proficiency levels appear to affect upon their language use.

The next step was to identify linguistic features that can distinguish four different clusters in Figure 2. Chi-square test and log-likelihood ratio test are usually used to compare the frequency patterns in two or more corpora (Baker, Hardie, & McEnery, 2006). However, the results of these methods are strongly affected by the sample size (Gries, 2014), and it is problematic for corpus-based studies which compare very high-frequency words. Therefore, Cramér's *V*, which is independent of the sample size, was used as keyness for comparing the frequency patterns in four clusters shown in Figure 2. Furthermore, *z* scores were checked to investigate frequent or infrequent linguistic features in each learner group. Table 2 summarizes the results of Cramér's *V* and *z* score for the top 20 linguistic features, whose frequencies are considerably different among clusters.

Table 2. Cramér's *V* and *z* Scores of the Top 20 Linguistic Features whose Frequencies are Considerably Different among Clusters

Linguistic feature	*z* score				Cramér's *V*
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	
first person pronouns	-1.203	1.236	0.091	-0.124	0.062
second person pronouns	-1.016	-0.659	0.581	1.094	0.046
nominalizations	1.458	-0.754	-0.179	-0.525	0.039
attributive adjectives	0.685	-1.001	1.018	-0.702	0.028
third person pronouns	0.713	0.036	0.677	-1.425	0.023
contractions	0.713	0.036	0.677	-1.425	0.022
past tense	-1.384	-0.072	0.828	0.628	0.022
private verbs	-1.361	0.820	0.675	-0.134	0.021
predictive modals	1.236	-1.127	0.259	-0.369	0.021
total prepositional phrases	0.712	0.245	0.515	-1.472	0.021
other adverbial subordinators	1.358	-0.916	0.109	-0.551	0.019
present tense	-1.004	0.782	0.934	-0.712	0.019
indefinite pronouns	-1.203	1.236	0.091	-0.124	0.018
split auxiliaries	1.362	-0.979	0.045	-0.428	0.018
emphatics	-0.141	-1.062	1.353	-0.149	0.017
other total nouns	-0.074	-0.247	1.361	-1.040	0.017
independent clause coordination conjuncts	-1.138	1.270	0.140	-0.272	0.017
WH relatives in object position	-0.773	1.468	-0.397	-0.298	0.015
that verb complements	-0.254	1.142	0.343	-1.231	0.015

The *z* score is a measure of how far a given frequency value is from the mean, expressed as a number of standard deviations (Oakes, 1998). In Table 2, positive *z* scores represent that the relative frequency of given linguistic features is higher than the average frequency of all four clusters, and negative scores represent that the frequency is lower than the average. According to Jarvis, Grant, Bikowski, and Ferris (2003), *z* scores above 0.5 and below 0.5 indicate noteworthy deviation from the central tendency, and they can characterize the cluster membership (Friginal, Li, & Weigle, 2014). Table 3 lists the frequent and infrequent linguistic features in each cluster identified with *z* scores shown in Table 2.

Table 3. Frequent and Infrequent Linguistic Features in Four Clusters

	Frequent	Infrequent
Cluster 1	nominalizations	first person pronouns
	attributive adjectives	second person
	third person pronouns	pronouns
	contractions	past tense
	predictive modals	private verbs
	total prepositional	present tense

	phrases	indefinite pronouns
	other adverbial subordinators	independent clause coordination
	split auxiliaries	WH relatives in object position
	conjuncts	second person pronouns
Cluster 2	first person pronouns	nominalizations
	private verbs	attributive adjectives
	present tense	predictive modals
	indefinite pronouns	other adverbial subordinators
	independent clause coordination	split auxiliaries
	WH relatives in object position	emphatics
	that verb complements	
Cluster 3	second person pronouns	
	attributive adjectives	
	third person pronouns	
	contractions	
	past tense	
	private verbs	
	total prepositional phrases	
	present tense	
	emphatics	
	other total nouns	
Cluster 4	second person pronouns	nominalizations
	past tense	attributive adjectives
		third person pronouns
		contractions
		total prepositional phrases
		other adverbial subordinators
		present tense
		other total nouns
		conjuncts
		that verb complements

As shown in Table 3, the written compositions of Cluster 1, which contains all Hong Kong learners, as well as Japanese and Korean C1-level learners, exhibit the tendency to use linguistic features prominent in academic writings, such as nominalizations, predictive modals, and conjuncts more frequently than other learner groups. This is seen in the following essay sample.

> (1) Someone believed that having a part-time job for a university student is important, but I cannot totally agree. Certainly, life and study in university is quite different from one's previous

> **education** experience as more individual **motivation** is required and less **rigidity** to decide one's own learning **activities**. Having a part-time job **will** be a good choice as students can earn their tuition fee and gain some experiences which can make their resumes more impressive and attractive. **However**, for the students who want to be involved in academic research, having a part time job may not be that essential and even at certain circumstances, can be counterproductive. I **shall** concede that they can gain some experience in team work, time **management** and **communication** skills. But these can also be acquired by doing research projects in the laboratory. **Furthermore**, working part-time can have negative influence on their academic performance as people's energy is not inexhaustible and there **will** be inevitable time conflicts. After all, their results in academic are far more important than working experience. In addition, at the university level, most students are not expected as experienced or equipped with apt skills. Most of the time, they are not likely to be on the appropriate positions where they **will** be after **graduation**. So a part-time job **won't** be efficient to tell them what their future careers look like. (HKG_C1)

It is noteworthy that the register awareness of Hong Kong learners who belong to the Outer Circle is higher than that of other learners in East Asia who belong to the Expanding Circle in terms of three concentric circles of English language (Kachru, 1992).

In contrast, Cluster 2, which contains A2- to B2-level Japanese learners, shows a very frequent use of linguistic features typical of spoken language, such as first person pronouns, private verbs, present tense, and independent clause coordination. In particular, some influence from spoken language can be found in *I think* in the sentence-final position and *and* in the sentence-initial position in the following essay.

> (2) **I** agree with this topic that smoking should be banned at all the restaurants in the country. There are two reasons why **I think** so. First, smoking is bad for our health, especially lung. Smoking is known as the major cause of lung cancer. Smoking also causes many heart or bronchus diseases. If all the restaurants in the country would prohibit smoking, the number of smokers, or the number of smoking times might be decreasing, **I think**. Second, some of the nonsmoker spaces are insignificant. Many restaurants in Japan have nonsmoking space. But actually, sometimes smoke from tobacco comes to nonsmoking space from smoking space. It is very bad for nonsmoker because most of they are not like smoking smell. **I**'m nonsmoker, **and I** hate

> smoking smoke too. **And** generally speaking, it is more danger around smokers who have a smoke than smoker themselves. So they hate tobaccos smoke, all the more they are eating. However, for smokers, it is stressful that they cannot smoke, **and** maybe it is bad for their health. But smokers should stop smoking even if they are eating. So **I**'m in favor of this topic. **I** wish it would come true some day, **and** other public institution should be banned too, **I think**. (JPN_A2)

Moreover, Cluster 3 and 4, which contain all Taiwanese learners and Korean A2- to B2-level learners, show the usage of second person pronouns more frequently than other clusters. Second person pronouns are also salient in conversation (Biber, Johansson, Leech, Conrad, & Finegan, 1999), and they reflect a more informal style of writing than first person pronouns (Smith, 1986).

> (3) I agree that it is important for college students to have a part-time job. There are several reasons. First, **you** can earn money in **your** own. When we grow up, the entire bill about studying was paid by our parents. Usually one enters the college at the age of eighteen. Eighteen years old means **you** are an adult. **You** should share the economical load of **your** family. **You** can use the money to pay for **your** registration and learn to be independent. Second, **you** can learn how to deal with problems encountered in **your** work. When **you** solve these problems, **you** can get experiences. Those experiences will one day help **you** when **you** graduate from school. Last, **you** can make friends from **your** work. A good friend will be help of **you** future work, especially when **you** are in trouble. He can save **you** when **you** need help. But as a student, **your** primary goal is get good grade to pass the exam. So don't spend too much time at part-time job. **You** need time to study and go to class. Try to distribute **your** time among class, part-time job, playing and rest! Those things must be balanced. Or **you** will be regret in the future. (TWN_B2)

The results of present study indicate linguistic features that learners from particular L1 groups have used frequently and infrequently, which can contribute to our understanding of the variation across learner language. The results also include a global description of interlanguage variation across proficiency levels, which have the potential to be used to develop syllabi, teaching materials, and language tests targeted for learners at a particular proficiency level of a particular L1 background.

5 Conclusion and Future Work

The purpose of this study was to investigate the influence of speech on Asian learners' written production. The results suggest that the L1s of learners primarily affect the degree of their register awareness. Hong Kong learners display a set of stylistically appropriate features, such as nominalizations, predictive modals, and conjuncts, in their academic prose whereas Japanese learners exhibit many of informal features, such as first person pronouns, private verbs, and independent clause coordination, in their written production. Besides, Korean and Taiwanese learners show several features typical of speech, including second person pronouns, in their writing. In addition, this study demonstrates the effectiveness of Biber's list of linguistic features in the study of spoken nature in L2 writing. A further direction of this study is to identify the register-related problems specific to each learner group by meticulously considering the effect of writing tasks as well as individual differences in L2 writing.

References

Abe, M. (2014). Frequency change patterns across proficiency levels in Japanese EFL learner speech. *Journal of Applied Language Studies, 8*(3), 85-96.

Abe, M., Kobayashi, Y., & Narita, M. (2013). Using multivariate statistical techniques to analyze the writing of East Asian learners of English. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world – Vol.1* (pp. 55-65). Kobe: School of Language and Communication, Kobe University.

Aguado-Jiménez, P., Pérez-Paredes, P., & Sánchez, P. (2012). Exploring the use of multidimensional analysis of learner language to promote register awareness. *System, 40*(1), 90-103.

Aijmer, K. (2002). Modality in advanced Swedish learners' written interlanguage. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 55-76). Amsterdam: John Benjamins.

Alberti, G. (2013). An R script to facilitate correspondence analysis: A guide to the use and the interpretation of results from an archaeological perspective. *Archeologia e Calcolatori, 24,* 25-53.

Altenberg, B., & Tapper, M. (1998). The use of adverbial connectors in advanced Swedish learners' written English. In Granger, S. (Ed.), *Learner English on computer* (pp. 80-93). London: Longman.

Baker, P., Hardie, A., & McEnery, T. (2006). *A glossary of corpus linguistics.* Edinburgh: Edinburgh University Press.

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language, 62,* 384-414.

Biber, D. (1988). *Variation across speech and writing.* Cambridge: Cambridge University Press.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* Harlow: Pearson Education.

Conrad, S., & Biber, D. (Eds.) (2001). *Variation in English: Multi-dimensional studies.* Harlow: Longman.

Divjak, D., & Fieller, N. (2014). Cluster analysis: Finding structure in linguistic data. In D. Glynn & J. A. Robinson (Eds.), *Corpus methods in cognitive semantics: Quantitative studies in polysemy and synonymy* (pp. 405-441). Amsterdam: John Benjamins.

Friginal, F., Li, M., & Weigle, S. C. (2014). Revisiting multiple profiles of learner compositions: A comparison of highly rated NS and NNS essays. *Journal of Second Language Writing, 23,* 1-16.

Gilquin, G., & Paquot, M. (2008). Too chatty: Learner academic writing and register variation. *English Text Construction, 1*(1), 41-61.

Glynn, D. (2014). Correspondence analysis: An exploratory technique for identifying usage patterns. In D. Glynn & J. A. Robinson (Eds.), *Corpus methods in cognitive semantics: Quantitative studies in polysemy and synonymy* (pp. 443-485). Amsterdam: John Benjamins.

Granger, S. (1996). From CA to CIA and back: An integrated contrastive approach to bilingual and learner computerized corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in contrast: Text-based cross-linguistic studies* (pp. 37-51). Lund: Lund University Press.

Granger, S., & Petch-Tyson, S. (1996). Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes, 15,* 19-29.

Granger, S., & Rayson, P. (1998). Automatic profiling of learner texts. In Granger, S. (Ed.), *Learner English on computer* (pp. 119-131). London: Longman.

Gries, S. Th. (2014). Frequency tables: Test, effect sizes, and explorations. In D. Glynn & J. A. Robinson (Eds.), *Corpus methods in cognitive semantics: Quantitative studies in polysemy and synonymy* (pp.365-389). Amsterdam: John Benjamins.

Hawkins, J. A., & Filipović, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the Common European Framework.* Cambridge: Cambridge University Press.

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics, 5*(3), 299-314.

Ishikawa, S. (2011). A new horizon in learner corpus studies: The aim of the ICNALE project. In G. Weir, S. Ishikawa, & K. Poonpon (Eds.), *Corpora and language technologies in teaching, learning and research* (pp. 3-11). Glasgow: University of Strathclyde Press.

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing, 12*(4), 377-403.

Kachru, B. (1992). *The other tongue: English across cultures.* Urbana, IL: University of Illinois Press.

Kobayashi, Y., & Abe, M. (2016). Automated scoring of L2 spoken English with random forests. *Journal of Pan-Pacific Association of Applied Linguistics, 20*(1), 55-73.

Lorenz, G. (1999). Learning to cohere: Causal links in native vs. non-native argumentative writing. In W. Bublitz, U. Lenk, & E. Ventola (Eds.), *Coherence in spoken and written discourse: How to create it and how to describe it* (pp. 53-66). Amsterdam: John Benjamins.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice.* Cambridge: Cambridge University Press.

Meunier, F. (2015). Developmental patterns in learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 379-400). Cambridge: Cambridge University Press.

Murakami, A. (2009). *A corpus-based study of English textbooks in Japan and Asian countries: Multidimensional approach* (Unpublished master's thesis). Tokyo University of Foreign Studies, Tokyo, Japan.

Oakes, M. P. (1998). *Statistics for corpus linguistics.* Edinburgh: Edinburgh University Press.

Oi, K. (2016). Introduction. In K. Oi (Ed.), *EFL writing in East Asia: Practice, perception and perspectives* (pp. 4-12). Tokyo: Seisen University.

Petch-Tyson, S. (1998). Writer/reader visibility in EFL written discourse. In S. Granger (Ed.), *Learner English on computer* (pp. 107-118). London: Longman.

Sardinha, T. B., & Pinto, M. V. (Eds.) (2014). *Multi-dimensional analysis, 25 years on: A tribute to Douglas Biber.* Amsterdam: John Benjamins.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing 12*(4), 44-49.

Smith, E. L. (1986). Achieving impact through the interpersonal component. In B. Couture (Ed.), *Functional approaches to writing* (pp. 108-119). London: Pinter.

Appendix
Linguistic features analyzed in the present study (Based on Biber, 1988)

A. Tense and aspect markers
    1. past tense
    2. perfect aspect
    3. present tense
B. Place and time adverbials
    4. place adverbials
    5. time adverbials
C. Pronouns and pro-verbs
    6. first person pronouns
    7. second person pronouns
    8. third person pronouns (excluding *it*)
    9. pronoun *it*
    10. demonstrative pronouns
    11. indefinite pronouns
    12. pro-verb *do*
D. Questions
    13. direct WH-questions
E. Nominal forms
    14. nominalizations (ending in *-tion, -ment, -ness, -ity*)
    15. other total nouns (except for nominalizations)
F. Passives
    16. agentless passives
    17. *by*-passives
G. Stative forms
    18. *be* as main verb
    19. existential *there*
H. Subordination
H1. Complementation
    20. *that* verb complements
    21. *that* adjective complements
    22. WH-clauses
    23. infinitives (*to*-clause)
H2. Participial forms
    24. past participial postnominal (reduced relative) clauses
H3. Relatives
    25. *that* relatives in subject position
    26. *that* relatives in object position
    27. WH relatives in subject position
    28. WH relatives in object position
    29. WH relatives with fronted preposition
H4. Adverbial clauses

30. causative adverbial subordinators: *because*
31. concessive adverbial subordinators: *although, though*
32. conditional adverbial subordinators: *if, unless*
33. other adverbial subordinators: (having multiple functions)

I. Prepositional phrases, adjectives, and adverbs

34. total prepositional phrases
35. attributive adjectives
36. predicative adjectives
37. total adverbs (except conjuncts, hedges, emphatics, discourse particles, downtoners, amplifiers)

J. Lexical classes

38. conjuncts
39. downtoners
40. hedges
41. amplifiers
42. emphatics
43. discourse particles

K. Modals

44. possibility modals
45. necessity modals
46. predictive modals

L. Specialized verb classes

47. public verbs
48. private verbs
49. suasive verbs
50. seem and appear

M. Reduced forms and dispreferred structures

51. contractions
52. stranded prepositions
53. split infinitives
54. split auxiliaries

N. Coordination

55. phrasal coordination
56. independent clause coordination (clause initial *and*)

O. Negation

57. synthetic negation
58. analytic negation: *not*

Yuichiro Kobayashi
Faculty of Sociology
Toyo University
5-28-20, Hakusan, Bunkyo-ku, Tokyo, 112-8606, Japan

Phone: 03-3945-8660
E-mail: kobayashi0721@gmail.com

Mariko Abe
Faculty of Science and Engineering
Chuo University
1-13-27, Kasuga, Bunkyo-ku, Tokyo, 112-8551, Japan
Phone: 03-3817-1958
E-mail: abe.127@g.chuo-u.ac.jp

Received: June 29, 2016
Revised: December 4, 2016
Accepted: December 7, 2016