Using Multivariate Statistical Techniques to Analyze the Writing of East Asian Learners of English

Mariko ABE
Chuo University

Yuichiro KOBAYASHI
Ritsumeikan University

Masumi NARITA
Tokyo International University

Abstract

The primary purpose of Contrastive Interlanguage Analysis (CIA) (Granger, 1996) is to reveal first language factors that influence the development of a learner language by comparing the performance of learners who speak different native languages. In the present study, we use CIA to investigate differences in the linguistic features of writing by different L2 East Asian learners of English. Multivariate statistical techniques, namely correspondence analysis and hierarchical cluster analysis, were used to examine the use of a wide range of linguistic features, such as (a) vocabulary, (b) parts-of-speech, (c) semantic categories for the major word classes, and (d) grammatical characteristics, in the writing of East Asian learners of English from different countries, including Hong Kong, Taiwan, Korea, and Japan. Data on the frequency of each linguistic feature are crucial for analyzing differences between learner language groups. Samples were sourced from ICNALE (Ishikawa, 2011), a corpus of a million words writing samples from EFL learners, which is considered to be the largest East Asian composition database. The purpose of this study was to identify the linguistic features that can be used to discriminate between different East Asian EFL learners and to distinguish native and non-native speakers of English to build up a general picture of interlanguage use. We expect our findings to contribute to our understanding of learner language development from multiple linguistic perspectives.

Keywords

Multivariate statistical techniques, Learner language variation, Written language, East Asian learners of English

I Introduction

In recent years, the application of large computational databases of written and spoken samples produced by language learners has developed as a way to reveal the influence of first language (L1) on interlanguage development. This approach was enabled by comparing the digitized performance of language learners who speak different native languages, as well as comparing digitized samples produced by native and non-native speakers, allowing researchers to reveal the distinguishing characteristics of language use in native and non-native speakers. The International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2011) is based on the idea of contrasting digitized linguistic data for interlanguage analysis. It has been released as a data-source for analyzing the writing of East Asian learners of English, enabling researchers to investigate interlanguage studies in Asia from a much wider range of perspectives.

II Literature Review

2.1 Learner Corpus Studies: Differences between First Languages

As the construction of learner corpora flourishes all over the world, researchers have begun to investigate varieties of learner languages in more detail (Ellis, 2008); this research paradigm shift, advocated by Granger (1996), is called Contrastive Interlanguage Analysis (CIA). By using digitized learner performance data, researchers can gain vast amounts of frequency-based information on vocabulary or sentence structures, and this information reveals underuse and overuse patterns across the interlanguages of different L1 groups. It can thereby be used to suggest whether or not learner language is influenced by the L1 (Ortega, 2009). Accordingly, examining the influence of the L1 allows researchers to determine whether certain characteristics of learner language are universal phenomena or unique developmental characteristics indigenous to a specific L1.

Biber and Reppen (1998), for example, investigated the use of complement clauses and reported that four learner groups (French, Spanish, Chinese, and Japanese) shared similar patterns in how to use them: (a) *that*- and *to*-clauses were significantly more common, (b) -*ing* and WH-clause were uncommon, and (c) the use of the *that*-clause in learners' essays was similar to its use in the conversations of English speakers. In the same way, Ringbom (1998) investigated whether there were differences in vocabulary use frequencies among English learners with different L1s, finding that different L1 learners overuse the top 30 to 100 most frequently used words. Aijmer (2002) also found that the overuse of modal auxiliaries, modal adverbials, and lexical verbs with modal meaning was a widespread phenomenon seen in French, German, and Swedish L2 writers, which partly reflects the developmental and interlingual characteristics of

learner language use. What is more, Altenberg (2002) argued that Swedish English learners' overuse of the causative *make* was caused by L1 transfer, which is due to cross-linguistic similarity, comparing it with French learners' underuse of the form.

2.2 Problems in Previous Literature

Many previous CIA studies have used the International Corpus of Learner English (ICLE), which is one of the most well-known written learner corpora, with samples from learners from more than 20 native language backgrounds. The ICLE has a reference corpus consisting of production data from tasks also completed by native speakers of English, which enables researchers to compare the writing of native and non-native speakers. However, fewer learner corpus-based studies have targeted East Asian learners of English, because a sufficiently large-scale learner corpus with proficiency level information based on objective rubric was not available to the public.

Additionally, despite early work on Second Language Acquisition (SLA), relatively few researchers have been concerned with describing interlanguage development using multiple linguistic features. More specifically, the number of targeted linguistic features in previous SLA studies was limited (Biber, Conrad, & Reppen, 1998), which has resulted in an insufficient understanding of the general picture of interlanguage. This problem is possibly caused by methodological limitations, as language processing technologies, such as the automatic detection of relevant linguistic features and multivariate statistical analyses for assessing interlanguage variation, have not been used to their full potential.

III Research Design

3.1 The Purpose of this Study and Research Questions

In order to address the problems discussed in the previous section, this study aims to profile the language use characteristics of East Asian learners of English and to identify linguistic features that can be used to discriminate between different L1 groups and native speakers of English. The present study is based on the methodology originally developed by Biber (1988) to analyze the differences in the spoken and written language of native speakers of English. As in Biber's (1988) study, we used a large amount of linguistic data showing multiple linguistic features, identifying the linguistic features noted by this earlier study to investigate learner language in this new study group.

3.2 Corpus Data

Since it was crucial to have sufficient frequency information of targeted linguistic features, data for the present study were sourced from ICNALE, a corpus of a million words written samples from EFL learners and native speakers. The present study used part of this corpus, including written compositions from 2,000 EFL learners and 400

native speakers of English. Figure 1 provides an indication of the general research design and Table 1 shows the size of each sub-corpus compared in this study.

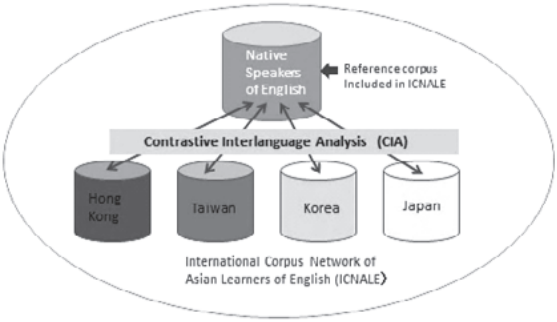

Fig. 1. Diagrammatic representation of the research design of the present study.

Table 1. Corpus Size of Native Speakers of English and Four Groups of East Asian EFL Learners.

	Native Speakers (EN)	Hong Kong (HK)	Korea (KOR)	Taiwan (TWN)	Japan (JPN)	Total
Participants	400	200	600	400	800	2,400
Word frequency	89,950	47,365	135,447	91,497	177,236	541,495
Mean of word frequency	225	237	226	229	222	226

3.3 Linguistic Features

In the present study, 58 linguistic features were selected from the original list of 67 linguistic features in Biber (1988) to analyze differences between written samples. Seven features: (a) demonstratives, (b) gerunds, (c) present participial clauses, (d) past participial clauses, (e) present participial WHIZ deletion relatives, (f) sentence relatives, and (g) subordinator-that deletion, could not be included in the present analysis because of differences in the software used to annotate part-of-speech tags. In addition, type/token ratio (TTR) and word length were excluded because they cannot be categorized as linguistic features but as word variation. A full list of the linguistic features analyzed in the present study can be found in the Appendix.

3.4 Statistical methods

The present study applies the linguistic feature list used by Biber (1988) to learner language, but instead of employing factor analysis, we used multivariate methods, such

as correspondence analysis and cluster analysis, to investigate differences between different L1 groups and native speakers, since these methods are more suitable for investigating similarities among variables (Oakes, 1998; McEnery & Hardie, 2012). Nakamura (1995) also recommended other multivariate statistical methods, including Hayashi's Quantification Method Type III, which is similar to correspondence analysis, and he pointed out that factor analysis produces different results because of its numerous steps and alternative choices for statistical processing.

Correspondence analysis was employed in this study as it has key advantages over other multivariate statistical methods. The first advantage lies in its simplicity of calculation, since the mathematical solution can be gained by one eigenvalue calculation and singular value decomposition. Moreover, there are no options in the calculation process, so its reproducibility is high compared with factor analysis and principal component analysis (Kobayashi, 2010). The second advantage is that it displays similarities and dissimilarities among variables in a scatter plot, which is beneficial because the characteristics of variation are summarized in two to three large groups. It reduces the complexity of the data, and thus it is can be used as a first step to identify analysis points for a detailed investigation (Baayen, 2008).

In order to supplement the findings of correspondence analysis, another multivariate statistical procedure, hierarchical cluster analysis, was then conducted. This statistical method classifies the groups in the scatterplot of correspondence analysis into larger meaningful groupings; it also specifies the linguistic features that discriminate different L1 groups.

IV Results and Discussions

The primary results of correspondence analysis are shown in Figure 2. In this scatter plot, Dim 1 and Dim 2, represent the two strongest factors of linguistic variation and linguistic items that are similar to each other cluster into groups. The cumulative contribution rate of Dim 1 and Dim 2 was 78.41%, which indicates the extent that each variable contributes to explaining the variance among factors (Baayen, 2008). The present study follows the statistical methodology of Biber (1988), which focuses on the prominent axis of the results of factor analysis. In this study, Dim 1 explained 51.65% of the linguistic variation in the texts analyzed, so this prominent dimension will be the main focus of further discussion. As indicated in the scatter plot in Figure 2, speakers of different L1s are distributed along the horizontal axis, Dim 1, so this dimension can be interpreted as representing the different L1 groups, with JPN on the right side, TWN and KOR in the middle, and HK and EN on the left side of the dimension. Consequently, we can conclude that linguistic features in Biber (1988) can be used to distinguish different L1 groups and the difference between native and non-native speakers of English.

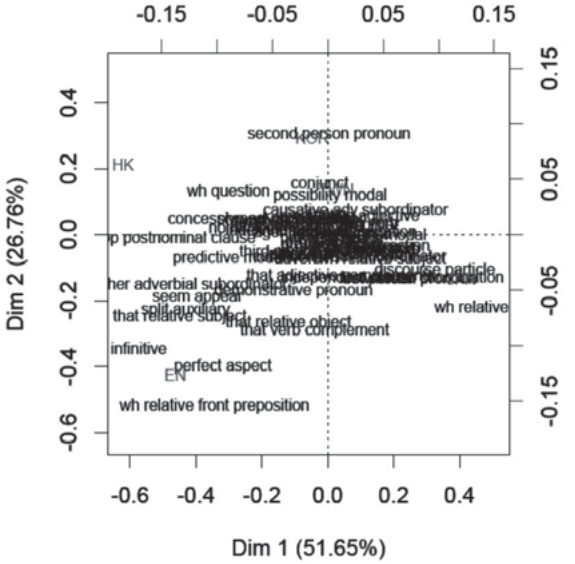

Fig. 2. Scatterplot showing the results of correspondence analysis.

The results of hierarchical cluster analysis are also presented in a dendrogram (see Figure 3). This dendrogram was obtained from the resulting coordination scale of the most powerful dimension of correspondence analysis. The Ward method, which maximizes the variance within and between groups to minimize the sum of squares of clusters, tends to create smaller clusters, and was used by applying Euclidean distance as the distance or similarity measure. As shown in the dendrogram, "native speakers of English and HK" and "JPN, KOR, and TWN" were clearly divided into different groups.

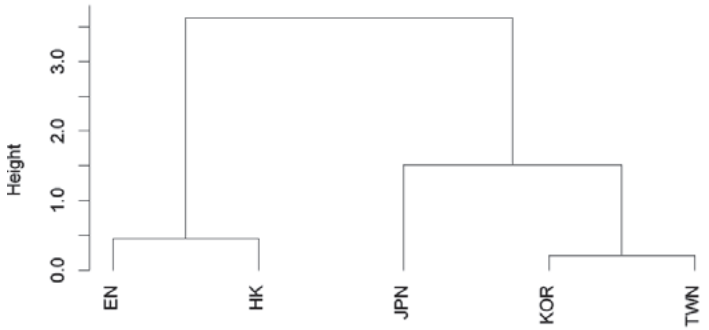

Fig. 3. Dendrogram representing the results of hierarchical cluster analysis.

The clusters in the dendrogram are likely to be divided by the frequency of certain linguistic features. The following box plots show differences in the frequency of nouns (see Figure 4) and personal pronouns (see Figure 5) among the different L1 groups. As indicated in the figures, these two features represent large-scale group differences in frequency. It is interesting to note that learners tend to use nouns more frequently than native speakers of English (EN). Also, Japanese learners of English (JPN) seem to have a tendency to use first person pronouns more than other L1 groups. These findings need to be interpreted through qualitative analysis in future work.

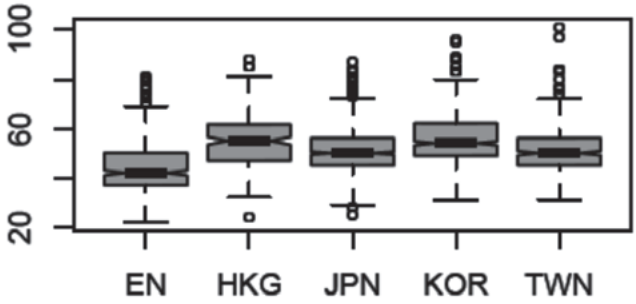

Fig. 4 Boxplots of noun for different L1 groups

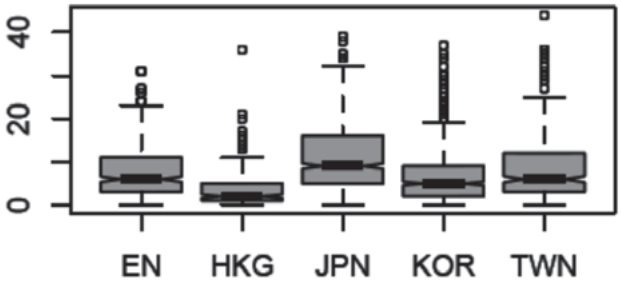

Fig. 5 Boxplots of first person pronoun for different L1 groups

Finally, it was possible to identify linguistic features that showed characteristic tendencies in different learners. Japanese learners of English infrequently used (a) attributive adjectives, (b) emphatics, (c) other adverbial subordinators, and (d) predictive modals; however, they frequently used (a) independent clause coordination, (b) the present tense, and (c) first person pronouns. In contrast, native speakers of English frequently used the following linguistic features: (a) the perfect aspect, (b) split auxiliaries, (c) adverbs, and (d) the relative subject.

V Conclusion

The purpose of this analysis was to explore the linguistic features that can discriminate different L1 East Asian learners, and native and non-native speakers of English. Frequency patterns of key linguistic features were identified and used to distinguish the variation of learner languages among different L1 groups. More detailed qualitative analysis of these linguistic features in future research may suggest whether learner language is influenced by universal phenomena or developmental characteristics of specific L1s. The present study contributes to our understanding of the nature and characteristics of learner language variation and shows that a methodological approach combining a large learner corpus, language processing techniques, and multivariate statistical analysis is useful in exploring this variation.

Acknowledgments

This work was supported by Grants-in-Aid for Scientific Research Grant Numbers 24320101 and 24520631.

References

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

Altenberg, B. (2002). Using bilingual corpus evidence in learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 37-53). Amsterdam: Benjamins.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

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., & Reppen, R. (1998). Comparing native and learner perspectives on English grammar: A study of complement clauses. In S. Granger (Ed.), *Learner English on computer* (pp. 145-158). London: Addison Wesley Longman.

Ellis, R. (2008). *The study of second language acquisition*. Oxford: Oxford University Press.

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.

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.

Kobayashi, Y. (2010). The correspondence analysis: Summarize the structure among the data. In S. Ishikawa, T. Maeda, & M. Yamazaki (Eds.), *An introduction to statistics for linguistic research* (pp. 245-264). Tokyo: Kuroshio Shuppan.

Nakamura, J. (1995). Text typology and corpus: A critical review of Biber's methodology. *English Corpus Studies*, *2*, 75-90.

Ortega, L. (2009). *Understanding second language acquisition*. London: Hodder Education.

Ringbom, H. (1998). Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In S. Granger (Ed.), *Learner English on computer* (pp. 41-52). London: Addison Wesley Longman.

Appendix. *Linguistic Features Analyzed in the Present Study*

Linguistic category and examples
A. Tense and aspect markers
1. past tense
2. perfect aspect
3. present tense
B. Place and time adverbials
4. place adverbials (e.g., across, behind, inside)
5. time adverbials (e.g., early, recently, soon)
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 (that, this, these, those)
11. indefinite pronouns (e.g., anybody, nothing, someone)
12. pro-verb *do* (e.g., the cat did it)
D. Questions
13. direct WH-questions
E. Nominal forms

14. nominalizations (ending in -tion, -ment, -ness, -ity)
15. total other nouns (except for nominalizations)

F. Passives

16. agentless passives
17. by-passives

G. Stative forms

18. be as main verb
19. existential there (e.g., there are several explanations . . .)

H. Subordination

H1. Complementation

20. that verb complements (e.g., I said that he went)
21. that adjective complements (e.g., I'm glad that you like it)
22. WH-clauses (e.g., I believed what he told me)
23. infinitives (*to*-clause)

H2. Participial forms

24. past participial postnominal (reduced relative) clauses
(e.g., the solution produced by this process)

H3. Relatives

25. that relatives in subject position (e.g., the dog that bit me)
26. that relatives in object position (e.g., the dog that I saw)
27. WH relatives in subject position (e.g., the man who likes popcorn)
28. WH relatives in object position (e.g., the man who Sally likes)
29. WH relatives with fronted preposition (e.g., the manner in which he was told)

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)
(e.g., since, while, whereas)

I. Prepositional phrases, adjectives, and adverbs

34. total prepositional phrases
35. attributive adjectives (e.g., the big horse)
36. predicative adjectives (e.g., the horse is big)
37. total adverbs (except conjuncts, hedges, emphatics, discourse particles, downtoners, amplifiers)

J. Lexical classes

38. conjuncts (e.g., consequently, furthermore, however)
39. downtoners (e.g., barely, nearly, slightly)
40. hedges (e.g., at about, something like, almost)

41. amplifiers (e.g., absolutely, extremely, perfectly)
42. emphatics (e.g., a lot, for sure, really)
43. discourse particles (e.g., sentence initial well, now, anyway)

K. Modals

44. possibility modals (can, may, might, could)
45. necessity modals (ought, should, must)
46. predictive modals (will, would, shall)

L. Specialized verb classes

47. public verbs (e.g., acknowledge, admit, agree)
48. private verbs (e.g., anticipate, assume, believe)
49. suasive verbs (e.g., agree, arrange, ask)
50. *seem* and *appear*

M. Reduced forms and dispreferred structures

51. contractions
52. stranded prepositions (e.g., the candidate that I was thinking of)
53. split infinitives (e.g., he wants to convincingly prove that …)
54. split auxiliaries (e.g., they are objectively shown to …)

N. Coordination

55. phrasal coordination (e.g., NOUN and NOUN, ADJ and ADJ)
56. independent clause coordination (clause initial and)
(e.g., It was my birthday and I was excited.)

O. Negation

57. synthetic negation (e.g., no answer is good enough for Jones)
58. analytic negation: not (e.g., that's not likely)