THE JOURNAL OF ASIA TEFL
Vol. 21, No. 1, Spring, 2024, 119-138
http://dx.doi.org/10.18823/asiatefl.2024.21.1.7.119

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511

Lexicogrammar of the L2 English Essays Written by Asian College Students: A Corpus-Based Study

Shin'ichiro Ishikawa
Kobe University

By combining a contrastive interlanguage analysis (Granger, 1996, 2015) and a multidimensional analysis (Biber, 1988), this study compared the essays written by Asian college students (in China, Indonesia, Japan, Korea, Taiwan, and Thailand) and L1 English native speakers (ENS). These essays were taken from the International Corpus Network of Asian Learners of English (ICNALE; Ishikawa, 2023a). Through four analyses based on dimension scores, key lexicogrammatical features, clustering, and classification, this study revealed that Asian learners tended to write less informational, less narrative, less overtly expressive, less abstract, and less time-constraint types of essays than ENS. It was also suggested that the lexicogrammatical features of essays were influenced more strongly by learners' L2 proficiency levels than their countries/regions and that learner essays were classified into four archetypes: colloquial and personal, interactive and persuasive, static and descriptive, and dynamic and reflective. These findings will shed new light on our understanding of Asian learners' L2 English use.

本研究においては、中間言語対照分析(Granger, 1996, 2015)および多次元分析法(Biber, 1988)という2つの手法を組み合わせ、International Corpus Network of Asian Learners of English (ICNALE; Ishikawa, 2023)から採取されたアジア圏大学生（中国・インドネシア・日本・大韓民国・台湾・タイ）および英語母語話者の作文を比較した。次元スコア・主要語彙文法特性・クラスタリング・統計的分類に基づく分析により、学習者の作文は母語話者のものより情報性・語り性・明示的表出性・抽象性・時間制約性が低く、学習者の作文の語彙文法特性は習熟度よりも国・地域に影響されていること、学習者作文は口語的・個人的、対人的・説得的、静的・描写的動的・反芻的の4 タイプに分類されることなどが示された。これらの結果は、アジア圏学習者のL2英語使用の理解に新しい光を投げかけるものである。

Keywords: learner corpus, ICNALE, writing, lexicogrammar, multi-dimensional analysis

Introduction

CIA-based Analysis of Learner Essays

When teaching L2 academic writing, many teachers realize that student essays differ in many ways from standard English essays. Past studies have shown that deviance from the native-speaker norm is observed not only in the writings of novice learners but also in those of upper-intermediate and even advanced learners (Lorenz, 1998; Altenberg & Tapper, 1998).

In learner corpus research (LCR), the analytical procedure called contrastive interlanguage analysis (CIA) has been widely adopted (Granger, 1996, 2015). It usually consists of two modules: a comparison between

the outputs of L2 English learners and those of referential writers, who are often chosen from L1 English native speakers (ENS), and a comparison between the outputs of L2 learners with various L1 backgrounds. CIA, which is often combined with statistical keyword analysis, enables researchers to objectively identify the lexical items significantly overused or underused by a particular learner group when compared to the referential writers or other learner groups.

Many of the CIA-based studies in the 1990s and 2000s were based on the International Corpus of Learner English (Granger et al., 2002), an extensive collection of essays produced by advanced college students, mainly in Europe. Altenberg and Tapper (1998), for instance, compared the essays of ENS and students from seven European countries and revealed that (i) the ratio of the top 100 words was 51.7% for ENS, but it reached 52.9–57.3% for students, meaning that learners tend to use a smaller number of high-frequency words repeatedly, (ii) the total adjusted frequency (per 10,000 words) of the top four verbs ("be," "have," "do," and "can") was 682 for ENS, but it was 737–861 for students, (iii) students use "think" 3–5 times as often as ENS because they have a clear tendency to begin the sentence with "I think," and (iv) students tend to overuse "and" and underuse "the" and "that."

Recent CIA studies have come to expand the coverage of target learners. Many studies focus on the usage of a particular word in the essays of Asian learners. Lin and Chung (2022) focused on the usage of "just" as a polysemous adverb and revealed that the gap between Taiwanese learners and ENS was seen not in its total frequency but in the frequency of each of its different usage types. Taiwanese learners, both young and adult, tended to overuse the depreciatory "just" (meaning "merely" and "simply") and underuse the specificatory "just" (meaning "very recently/soon," "by a small amount/time/distance," etc.). Zhi (2022) compared the usage of various types of "make" in essays of Chinese learners at intermediate and upper intermediate levels and revealed that the former use "make + pron. + v." (transformation) and the latter use "make + n._abstract" (supportive verb) more often than the other. Whitty, Parkinson, and Pham (2022) focused on the usage of "can/could" in essays of L1 Vietnamese undergraduate students (UGS), MA-level EFL students (MAS), and L1 English students (L1S). The analysis showed that the UGS overused "can" and underused "could," though they made rather few usage mistakes, while the MAS used "could" as much as L1S, but with many "awkward" deviances.

In addition, more and more studies pay attention to phrases or a broader set of words. Lu (2017) compared the occurrence patterns of grammatical and lexical collocations in essays of non-English major Chinese students and ENS and found that Chinese students adopted an approximately equivalent number of grammatical collocations but a greater number of lexical collocations, and they used both types of collocations with much less formal variety. Lu suggests the existence of an L1 influence behind Chinese learners' deviant L2 collocation use. Gao (2023) examined the use of phrasal verbs in the academic articles written by Chinese and ENS scholars and revealed that phrasal verbs tended to appear more in ENS writings than in Chinese writings and in soft sciences (linguistics and management) than in hard sciences (physics and computer sciences), while the analysis showed no significant differences between the two writer groups in terms of the high-frequent types and construction patterns of phrasal verbs. Yoo and Shin (2022) examined the lexical bundles used in the academic essays of Korean learners at three different proficiency levels and revealed that novice learners often used VP-based clausal bundles (e.g., "I agree with the," "that I want to") seen typically in casual speeches, while advanced learners came to use more complex NP/PP-based phrasal bundles often with post-modifiers (e.g., "the development of the," "that fact that there," "in the center of," "for a long time"). Leedham (2015) analyzed the high-rated academic essays produced by Chinese and British students studying at UK colleges and revealed that Chinese students tended to overuse a variety of connectors—informal connectors, in particular— ("in other words," "meanwhile," / "besides,"" what's more," "lots"), first-person plural pronouns ("we"), and the words with references to data and visuals ("formula," "figure," "appendix"/ "according to," "as below," "refer to the"). The author interprets that Chinese learners' dependence on tables, figures, formulae, lists, and list-likes in the articles may come from their writing strategy to reduce the quantity of connected prose writing (Leedham, 2015, p. 61).

The lexical and linguistic features that learner corpus research has focused on are not limited to the abovementioned topics. Ishikawa (2023a), for instance, scrutinized the ICNALE corpus to discuss the vocabulary, grammar, and pragmatics (pragmatic markers, politeness control, and gesture use) seen in Asian learners' L2 English use from an integrative analytical viewpoint. Kwon (2022) surveyed major learner corpus studies published in Korea and reported that the recent studies had covered a more comprehensive range of topics related to (a) nouns and pronouns (e.g., nominal modifiers, demonstratives, first-person pronouns), (b) verbs and tenses/aspects (e.g., basic verbs, prepositional verbs, V-ing forms, overpassivization, verb-noun collocations), (c) adjectives and adverbs (e.g., conjunctive/stance adverbials), (d) modality, (e) function words (e.g., "of," "because," logical connectors, conjunctive adjuncts, contrastive conjunctions, cohesive devices), (f) collocations/lexical bundles, and (g) constructions (e.g., dative, main/adverbial clause orderings).

Combination of CIA and MDA: Beyond Individual Words

The findings in the previous studies are undoubtedly of much pedagogical importance, but several methodological concerns seem to exist in some of them. It is because they tended to focus on individual words rather than larger constructions beyond words, specific local structures in the text rather than the text as a whole, and lexis alone rather than lexis and grammar in a combination.

When considering these limitations in the traditional CIA approaches, integration of the CIA into the multidimensional analysis (MDA; Biber, 1988) seems to be a promising direction (Ishikawa, 2023a, pp. 108–109). MDA was originally proposed by Douglas Biber as a method to investigate register variation in English texts. With the notion that "linguistic variation must be analyzed in terms of sets of co-occurring features" (Biber, 1988, p. 21), Biber collected 481 kinds of spoken and written English texts taken from 23 registers and counted the frequency of 67 types of lexicogrammatical features chosen from the literature in each of them. These "potentially important linguistic features" (Biber, 1988, p.64) are classified into 16 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 features, I: Prepositional patterns, adjectives and adverbs, J: Lexical specificity, K: Lexical classes, L: Modals, M: Specialized verb classes, N: Reduced forms and dispreferred structures, O: Coordination, and P: Negation) and 67 sub-categories (Biber, 1988, pp. 73–75). Then, he conducted a factor analysis and identified the six key dimensions determining the text variation, which include Dim 1: Informational versus Involved Production, Dim 2: Narrative versus Non-Narrative Concerns, Dim 3: Explicit versus Situation-Dependent Reference, Dim 4: Overt Expression of Persuasion, Dim 5: Abstract (versus) Non-Abstract Information, and Dim 6: On-line Informational Elaboration (Biber, 1988, p. 115). Although the total MDA consists of all the four steps shown above—identification of relevant variables, extraction of factors from the variables, functional interpretation of factors as dimension, and placement of registers on the dimensions—researchers often skip the first three processes and just compare the result reported in Biber (1988) with that obtained from a new corpus dataset (Brezina, 2018, p.161).

When integrated into CIA, MDA enables researchers to focus on larger constructions beyond individual words, analyze learner texts as a whole, and discuss lexis and grammar as a unit. Bearing this in mind, this paper aims to apply a new CIA/MDA integrative approach to L2 English essays produced by Asian college students from six English as a Foreign Language (EFL) countries/regions and at four different L2 proficiency levels. An analysis based on a broader range of lexicogrammatical features rather than a limited set of individual words is expected to deepen our understanding of the linguistic characteristics of Asian learners' L2 English use.

Method

Aim and RQs

By combining CIA and MDA, this study aims to reconsider the possible gaps between Asian EFL learners from different countries/regions and at different proficiency levels and ENS in writing from the viewpoint of lexicogrammar (Halliday, 1991). Thus, the following four research questions (RQs) were formulated for the current study:

RQ1. What differences are seen between writer groups in terms of lexicogrammatical dimension scores? (Differences in dimension scores)
RQ2. What lexicogrammatical features do learners commonly overuse/underuse? (Key lexicogrammatical features)
RQ3. Which of a country/region and an L2 proficiency level influences the clustering of writer groups? (Primary factors for clustering)
RQ4. How are the writer groups and lexicogrammatical features classified? (Writer/Feature classification)

Data

This study analyzed the topic-controlled essays written by Asian college students (including some graduate students) from six EFL countries/regions (China: CHN, Indonesia: IDN, Japan: JPN, Korea: KOR, Thailand: THA, Taiwan: TWN) as well as ENS, which include college students (ENS1), English teachers (ENS2), and others (ENS3). The topic was "It is important for college students to have a part-time job." Although the ICNALE includes essays written about two topics (a part-time job and non-smoking), this study focused only on the former, which is due to the need to control the possible influence of the topic on the essay content. The essay lengths were all between 200 words and 300 words. These were taken from the ICNALE Written Essays (Ishikawa, 2023a).

The ICNALE's learner-participants are classified into four CEFR-linked proficiency bands—A2, B1_1 (B1 lower), B1_2 (B1 upper), and B2+—according to their scores in the standard English proficiency test such as TOEIC and TOEFL or in the common vocabulary size test. The score/band conversion is based on the conversation table by each of the test agencies. Although test scores and actual performances in L2 may not always be in accordance, CEFR-band estimation from test scores has been widely conducted, and it holds validity to some extent. The number of writers in each proficiency group is shown in Table 1.

TABLE 1
The Number of Essays Analyzed in This Study

Countries/regions	A2	B11	B12	B2+	Sum
CHN	50	232	105	13	400
IDN	32	82	83	3	200
JPN	154	179	49	18	400
KOR	75	61	88	76	300
THA	119	179	100	2	400
TWN	29	87	61	23	200
ENS	ENS1 (college students): 100 ENS 2 (English teachers): 44 ENS 3 (others): 56	200			
Ttl					2,100

The total number of words in the 2,100 essays reached approximately 500,000 words.

Analytical Procedure

Firstly, learner essays were merged into 24 files based on the countries/regions and proficiency levels, and ENS essays were merged into three files (ENS1, ENS2, ENS3). These 27 files were processed with the Multidimensional Analysis Tagger (MAT; Nini, 2015a).

MAT, which is a duplication of the original Biber tagger, automatically assigns the Stanford parts-of-speech (POS) tags and Biber's VASW tags—a set of the function-based lexicogrammatical tags proposed in Biber's *Variation across Speech and Writing* (1988)— to texts and computes the dimension scores based on the frequencies of those tags. The sample below, which is from a part-time job essay written by a Chinese student (No. 001), shows how plain text is tagged.

(1) … I think a part-time job is one of the most important things in college life.

(2) … I_PRP think_VBP a_DT part-time_JJ job_NN is_VBZ one_CD of_IN the_DT most_RBS important_JJ things_NNS in_IN college_NN life_NN ._.

(3) … I_FPP1 think_VPRT [PRIV] [THATD] a_DT part-time_JJ job_NN is_VPRT [BEMA] one_CD of_PIN the_DT most_EMPH important_JJ things_NN in_PIN college_NN life_NN._.

(1) is a plain text before tagging. (2) is a text with the Stanford POS tags, where PRP, VBP, and DT, for instance, represent a personal pronoun, a verb (non-3rd person singular present), and a determiner, respectively. Then, (3) is a text with Biber's VASW tags, where VPRT, PRIV, THATD, and BEMA, for instance, represent a present-time verb, a private verb (i.e., a verb referring to a person's internal intellectual behavior), complementizer *that* deletion, and a main verb *be*, respectively. As seen in (3), a word is sometimes assigned plural tags, when the second and the subsequent tags appear in the square brackets. See the appendix for major tag codes used in this analysis.

In Biber's MDA, the type/token ratio (TTR), which usually represents a text's lexical diversity, is also used to calculate the dimension scores. This study computed TTR from the first 100 words of each file. Z score correction, which normalizes raw frequency, was not applied (Figure 1).

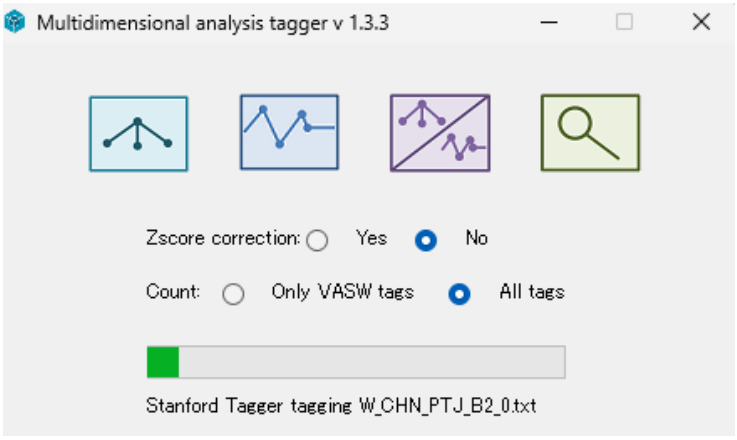

Figure 1. MAT setting.

Regarding RQ1, in order to scrutinize the possible differences between (i) learners and ENS, (ii) learners' proficiency levels, and (iii) their countries/regions of origin, this study compared the averages of the dimension scores of (i') 24 learner groups and three ENS groups, (ii') four proficiency-based learner groups, and (iii') six region-based learner groups. Regarding (iii'), to control the possible effect of proficiency, this study analyzed the data of learners only at B1 (B1_1 and B1_2) level.

Regarding RQ2, this study focused on the 49 tags used by all 27 writer groups. Analyzing the commonly used tags rather than the tags that only some particular groups frequently used seems to enhance the validity of the discussion.

TABLE 2
All the Tags Used for the Current Analysis

POS/Biber tags
AMP, AWL, [BEMA], CAUS, CC, CD, COMMA, COND, CONJ, [CONT], DEMO, DEMP, DT, EMPH, EX, FPP1, GER, IN, JJ, NEMD, NN, NOMZ, [PEAS], PERIOD*, PHC, PIN, PIT, POMD, PRED, [PRIV], PRMD, [PROD], QUAN, QUPR, RB, [SPAU], SPP2, [THATD], THVC, TIME, TO, TPP3, TTR, VB, VBG, VBN, VPRT, WDT, XX0

Note. In the original Stanford tag set, "." represents sentence final punctuations that include periods and question marks. This study renamed it to PERIOD to avoid confusion. See the Appendix for the meaning of each tag.

Next, the frequency of each of the 49 features for each of the 24 learner groups was examined. It was compared to the average frequency for ENS 1–3, and then a learner/ENS ratio (percentile value) was obtained. For example, per-million-word (PMW) adjusted frequencies of SPP2 (second person pronouns) were 1.09, 1.12, 0.18, and 0.24 for A2 Chinese learners, ENS_1, ENS_2, and ENS_3, respectively. In this case, the average of the three ENS subgroups was 0.51. Thus, the ratio was calculated as 213.73% (1.09/0.51*100). This suggests that A2 Chinese learners use SPP2 more than twice as much as ENS.

All the frequencies were converted to the learner/ENS ratios. Then, the features whose ratios exceeded 150% and those whose ratios were smaller than 50% were regarded as positive and negative key features characterizing learners' L2 use. Compared to the traditional keyword identification method, which often extracts all the items showing statistical significance in the frequency comparison, the current method could be more robust.

Then, for RQ3, based on the frequency table with 27 writer groups as variables and 49 features as cases, a hierarchical agglomerative cluster analysis was conducted. Cluster analysis is a procedure in which "[w]e take the individual data points and in a step-by-step (hierarchical) procedure join (i.e., agglomerate) the closest ones until we create one large cluster containing all the data points" (Brezina, 2018, p. 154). The distance was defined as the square root of (2-2r), and the Ward method was adopted. The result of the analysis was shown in the tree diagram, which visually presents how different writer groups are agglomerated. Analytical attention is paid to whether different writer groups are clustered according to the countries/regions or the proficiency levels.

Finally, regarding RQ4, a correspondence analysis was applied to the same frequency table. Correspondence analysis is a statistical measure to reduce the number of dimensions of variation and output a simple visual depiction of the frequency table in two-dimensional space, where the chi-squared distance is regarded as a measure of closeness/remoteness of the item category data included in the table (Brezina, 2018, p. 200). The result of the analysis was shown in the scatter plot, where the top two dimensions (Z1 and Z2) with the most significant factor loads were adopted for the horizontal and vertical axes.

Results and Discussions

RQ1 Differences in Dimension Scores

Learners and ENS

First, six dimension scores are compared between the essays of all Asian EFL learners and ENS.

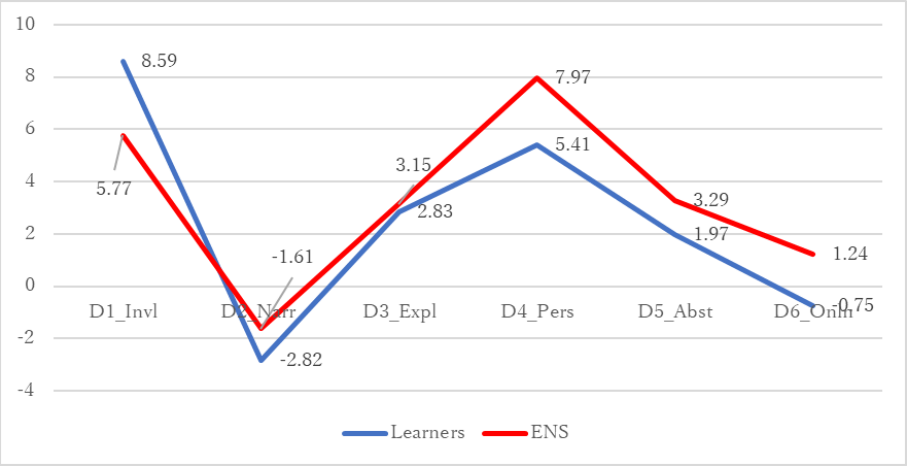

Figure 2. Dimension scores of the essays of Asian EFL learners and ENS.

Regarding D1 (Involved/Informational), a higher frequency of private verbs, that-deletion, contractions, present-tense verbs, and second-person pronouns leads to a higher involvement and higher frequency of nouns, longer words, prepositions, and adjectives leads to higher informationality. In Biber's analysis, telephone and face-to-face conversations present the highest scores (Scores: 35 to 40), while official documents show the lowest scores (Scores: -20 to -15). The scores of the ICNALE ENS essays and learner essays (total) are mainly similar to those of romantic fiction and prepared speeches (Scores 0 to 5) reported in Biber (1988, p. 128). Learners' essays are somewhat more involved than ENS essays.

Regarding D2 (Narrative/Non-narrative), a higher frequency of third-person pronouns and past-tense/past-aspect verbs leads to higher narrativity, while a higher frequency of present-tense verbs and attributive adjectives leads to non-narrativity. In Biber's analysis (Biber, 1988, p. 136), various fictions (romantic, mystery, science, general, and adventure) present the highest values (Scores: 5 to 7), while hobbies and broadcasts present the lowest values (Scores: -4 to -3). Learners' essays are more "here and now"-oriented and, therefore, less narrative when compared to ENS essays.

Regarding D3 (Explicit/Situation-dependent), a higher frequency of wh-relative clauses and pied piping constructions leads to higher explicitness or contextual independence, while a higher frequency of adverbs and place/time adverbials leads to contextual dependency. Official documents and professional letters present the highest values (Scores: 6 to 8), while telephone conversations and broadcasts present the lowest (Scores: -10 to -5). Learners' essays are hardly different from ENS essays in terms of D3. Both seem to be in a middle position between situation-independent and situation-dependent texts.

Regarding D4 (Overt expression of persuasion), a higher frequency of to-infinitives and prediction modals leads to more explicit stance marking. According to Biber (1988, p. 149), professional letters and editorials present the highest values (Scores: 3 to 4), while broadcasts present the lowest values (Scores: -5 to -4). Learners' essays are less stance-marked compared to L1 outputs of ENS.

Regarding D5 (Abstract/non-abstract), a higher frequency of conjuncts, passives (agentless or with by), and past participial clauses leads to higher abstractness. The highest values are assigned to academic prose and then to official documents (Scores: 4 to 6), while the lowest values are assigned to telephone conversations (Scores: -4 to -3). Learners' essays are less abstract, namely, more concrete, in comparison to ENS essays.

Finally, regarding D6 (On-line informational elaboration), a higher frequency of that-complementizers, demonstratives, and that-relative clauses on object position leads to higher time constraints in information development. Prepared speeches and interviews present the highest values (Scores: 3 to 3.5), while fiction (general, science, mystery, and adventure) offers the lowest values (Scores: -2 to -1.5). Learners' essays seem to be less time-constraint than ENS essays.

Thus, it could be concluded that learner essays are more involved than informational (D1), slightly less narrative (D2), less overtly expressive (D4), less abstract (D5), and less time-constraint (D6).

Learners at different proficiency levels

Next, dimension scores are compared between Asian EFL learners at four proficiency levels (A2, B1_1, B1_2, and B2+) and ENS as a reference.

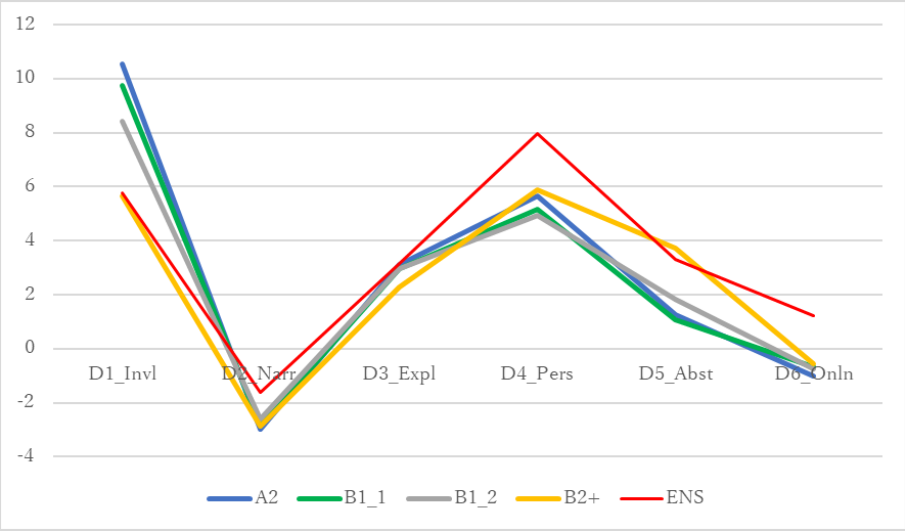

Figure 3. Dimension scores of the essay of learners at four proficiency levels and ENS.

The orders based on D1 dimension scores are A2 (10.55) > B1_1 (9.75) > B1_2 (8.42) > ENS (5.77) > B2+ (5.66). This suggests that advanced learners come to produce less involved and more informational texts than novice learners. It should also be noted that B2+ learners produce more informational texts than ENS.

Next, the orders based on D2 dimension scores are A2 (-2.98) < B2+ (-2.86) < B1_1 (-2.82) < B1_2 (-2.61) < ENS (-1.61). Though several reverses in proficiency levels are observed, this largely suggests that advanced learners tend to produce relatively more narrative texts than novice learners. However, a considerably large gap remains between B2+ learners and ENS.

The orders based on D3 dimension scores are ENS (3.15) > A2 (3.12) > B1_1 (2.97) > B1_2 (2.95) > B2+ (2.30). Advanced learners come to produce less explicit and more situation-dependent texts than novice learners. Unlike in the cases of D1 and D2 dimension scores, however, the increase in proficiency seems to suggest deviation from the ENS reference rather than approaching it.

Regarding D4, the orders are ENS (7.97) > B2+ (5.89) > A2 (5.67) > B1_1 (5.15) > B1_2 (4.93). The relationship between proficiency levels and overt stance marking in writing is not simple. Learners seem to produce less overtly persuasive, namely, more objective texts from A2 to B1, but they then come to produce more overtly persuasive texts from B1 to B2+. Thus, among the four proficiency levels, B2+ learners are closest to the ENS reference.

The orders based on D5 dimension scores are B1_1 (1.08) < A2 (1.27) < B1_2 (1.81) < B2+ (3.71) < ENS (3.29). Though a slight alternation in the orders is observed, advanced learners seem to produce more abstract texts and get closer to the ENS writing style.

Finally, the orders based on D6 dimension scores are A2 (-1.0) < B1_2 (-0.73) < B1_1(-0.68) < B2+ (-0.57) < ENS (1.24). Though a slight alternation in the orders is observed, it seems to imply that advanced learners tend to develop the contents more speedily, though there remains a gap between B2+ learners and the ENS.

Thus, the comparisons of the dimension scores have shown that Asian EFL learners gradually produce less involved and more informational (D1), more narrative-oriented (D2), less explicit and more situation-dependent (D3), more abstract (D5), and more online-developmental (D6) texts as their proficiency levels go up. Also, their essays come closer to ENS essays in terms of D1, D2, D5, and D6.

Learners from different countries/regions

Finally, dimension scores are compared between Asian intermediate-level EFL learners from six countries/regions.

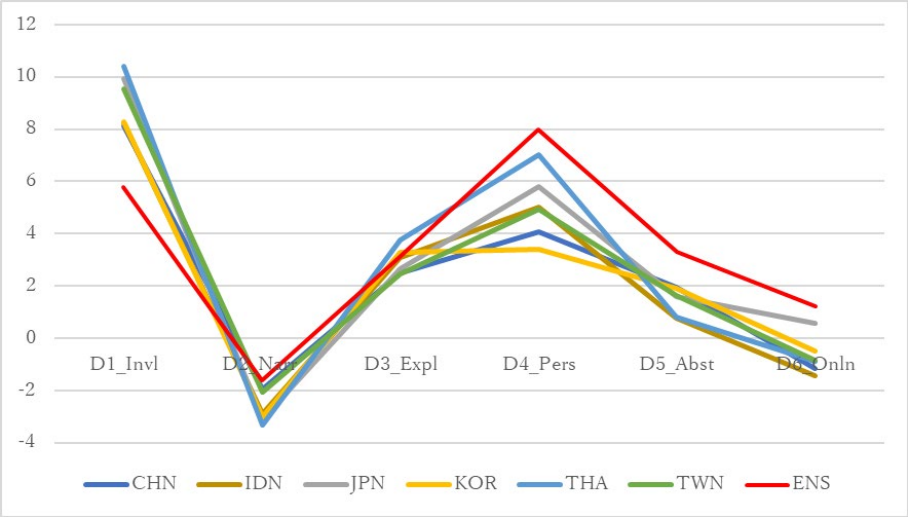

Figure 4. Dimension scores of the essays of intermediate learners from six countries/regions.

Regarding D1, Thai learners (10.4) produce more involved texts, while Chinese learners (8.12) produce more informational texts. Regarding D2, Chinese learners (-1.94) produce relatively more narrative-oriented texts, while Thai learners (-3.33) produce less narrative-like texts. Regarding D3, Thai learners (3.77) produce more explicit texts, while Chinese/Taiwanese learners (2.48/2.47) produce more situation-dependent texts. Then, as regards D4, Thai learners (7.03) produce more overtly persuasive texts, while Korean learners (3.42) produce more neutral texts. Regarding D5, Chinese/Korean learners (1.95/1.90) produce more abstract texts, while Indonesian learners (0.77) produce rather concrete texts. Finally, regarding D6, Japanese learners (0.58) produce more online-developing texts, while Indonesian learners (-1.42) tend to produce texts without such speedy developments.

The region-based comparisons illuminate that even among the EFL learners at the same proficiency levels, a considerably large gap exists in each of the six dimensions. They also suggest that learners from China and Thailand may tend to produce somewhat deviant texts when compared to other Asian EFL learners.

RQ2 Key Lexicogrammatical Features

Table 3 shows the lexicogrammatical features that each of the 24 learner groups overused or underused by more than 50% in comparison to the ENS writers, which are regarded as positive and negative key features characterizing each learner group. Features are listed in the decreasing order of the frequency gap. For A2 Chinese learners, for instance, SPP2 (213.73% of the ENS frequency) and [SPAU] (37.7% of the ENS frequency) are the most salient overused and underused items.

TABLE 3
Overused/Underused Lexicogrammatical Features

Learners	Positive Key Features Overused (>150%)	Negative Key Features Underused (<50%)
CHN_A2	SPP2, POMD, PERIOD, TIME, EMPH	[SPAU], DEMP, DEMO, CAUS
CHN_B11	SPP2, [PROD], POMD, PERIOD, NEMD	THVC, DEMP, CAUS,
CHN_B12	SPP2, POMD, [PROD], PERIOD	CAUS, THVC
CHN_B2	CONJ, COMMA, NEMD, [PROD], TIME,	CAUS, AMP, DEMP, DEMO
IDN_A2	CAUS, POMD	THVC, DEMP, [PEAS], [SPAU]
IDN_B11	CAUS, POMD, [PROD], TPP3, PERIOD,	[PEAS], THVC, [SPAU], VBN
IDN_B12	SPP2, [PROD], CAUS, POMD	THVC, [PEAS], WDT, [SPAU]
IDN_B2	SPP2, TIME, COMMA, NEMD	THVC, [PROD], [SPAU], [THATD], EMPH,
JPN_A2	NEMD, FPP1, [THATD], PERIOD	[SPAU], [PEAS], VBN,
JPN_B11	NEMD, FPP1, PERIOD,	[SPAU], QUPR, VBN,
JPN_B12	FPP1, PERIOD, [PROD], [THATD], CONJ, NEMD	[SPAU], [PEAS], VBN, PRMD
JPN_B2	[PROD], NEMD, [THATD], CONJ, PERIOD, EX, POMD, COMMA	[PEAS], VBN, [SPAU], THVC, PHC, PRMD
KOR_A2	SPP2, [PROD], CAUS, [THATD], PERIOD, AMP	THVC, [PEAS], WDT, [SPAU], TPP3,
KOR_B11	SPP2, PERIOD, CAUS,	[SPAU], WDT, THVC, [PEAS]
KOR_B12	SPP2, PERIOD, CONJ,	[SPAU], [PEAS], WDT,
KOR_B2	CONJ, SPP2, [PROD], PERIOD, POMD,	[PEAS], [SPAU], THVC,
THA_A2	SPP2, NEMD, CAUS, [PROD], PERIOD,	[SPAU], THVC, [PEAS], WDT, DEMP, VBN
THA_B11	SPP2, [PROD], NEMD, CAUS, QUPR, POMD, PERIOD,	[SPAU], [PEAS], VBN, THVC
THA_B12	SPP2, [PROD], NEMD, CAUS, POMD, QUPR	[SPAU], [PEAS], VBN, THVC
THA_B2	SPP2, [PROD], NEMD, DEMO, XX0, PERIOD, CAUS,	THVC, TPP3, [CONT], FPP1, [PEAS], VBG
TWN_A2	SPP2, [PROD], QUPR, POMD, AMP, [THATD]	THVC, [PEAS], GER, WDT, VBN, [SPAU]
TWN_B11	SPP2, [CONT], POMD, QUPR, [THATD],	[PEAS], [SPAU], VBN,
TWN_B12	SPP2,	[PEAS], AMP, THVC, DEMP
TWN_B2	CONJ, SPP2, CD, EX	DEMP, AMP, [SPAU], COND, [PEAS],

The total number of overused and underused words reaches 212 in tokens and 31 in types. However, several features are commonly overused or underused by plural learner groups.

Table 4 presents the features overused or underused by three or more learner groups, which might represent commonly "fossilized" features in L2 English essays of various Asian EFL learners.

TABLE 4
Overused/Underused Features by More than Three Learner Groups

Overused		Underused	
Feature tag	Freq	Feature tag	Freq
SPP2: second-person pronouns (you)	17	[SPAU]: split auxiliaries	19
PERIOD*: sentence-final punctuations	15	[PEAS]: perfect aspect	18
[PROD]: pro-verb (do)	14	THVC: *That* verb complements	16
POMD: possibility modals	12	VBN*: past participle verb	10
NEMD: necessity modals	11	DEMP: demonstrative pronouns	7
CAUS: causative adverbial subordinators	9	WDT*: *Wh*-determiners	6
[THATD]: subordinators *that* deletion	6	CAUS: causative adverbial subordinators	4
CONJ: conjuncts	6	AMP: amplifiers	3
QUPR*: quantifier pronouns	4		
COMMA*: commas	3		
FPP1: first person pronouns	3		
TIME: time adverbials	3		

Underused features, which are the features that ENS writers overuse in comparison to learners, are often closely related to overused items. A closer look at Table 4 would identify several patterns characterizing Asian EFL learner essays.

Conversationality (Overuse of SPP2/FPP1)

SPP2 is the overused feature for 17 of 24 learner groups. The overuse of SPP2 by a variety of learners and the overuse of FPP1 by Japanese learners suggests that Asian learners' essays are essentially subjective and interactive, as in oral conversations. They prefer speaking to readers ("you") and also talking about themselves ("I"/"we"). Meanwhile, ENS writers, especially teachers and others, tend to discuss the topic more detachedly.

(4) [You SPP2] have a part time job when you < SPP2> are in college that means you <SPP2> make money by yourself <SPP2>. (CHN)

(5) I <FPP1> think it is better to stop doing a part time job. But now we <FPP1> are college students. We <FPP1> have to manage our <FPP1> time wisely. (IDN)

(6) For some students, the extra money will assist them in obtaining needed books and materials for study. (ENS2)

Mixed modality (Overuse of POMD/NEMD)

POMD and NEMD are overused by 11-12 of 24 learner groups. Though the overuse of POMD is partly because learners often talk about (in)ability of the act (e.g., "college students can save money" [JPN]; "majority of students can't accumulate much wealth" [CHN]), they also use many hedge-like POMD, which marks politeness and causes their claims to sound indirect and circumlocutionary. Meanwhile, learners also overuse NEMD, which, contrary to POMD, makes their claims directive, imperative, and even aggressive. A strange mixture of politeness and impoliteness seen in many learners' essays may reflect their limited understanding of modality and stance, which ENS writers seem to control more carefully and effectively by combining POMD, NEMD, and PRMD (predictive modals).

(7) In my opinion, the answer (i.e., to the choice between work and study) can <POMD> be various… (CHN)

(8) So you must <NEMD> be careful not to waste money and time. You should <NEMD> always keep in your mind that time is finite… (JPN)

(9) … students should <NEMD> learn about financial responsibility which would <PRMD> empower them to have more control over their own lives…These may <POMD> seem like basic things but many students may <POMD> not have learned these things at home and might <POMD> not been taught at their college. (ENS2)

Fragmentality (Overuse of PERIOD/COMMA/CONJ and underuse of DEMP)

PERIOD is overused by 15 learner groups, and COMMA is by three. Overusing these punctuation marks suggests that learners' essays often consist of shorter fragmental sentences or phrases. This also explains why six learner groups overuse CONJ. Learners divide the text into smaller portions and recombine them with simple conjunctions such as "for," "however," "therefore," "moreover," "thus," "furthermore," "rather," and "otherwise," and also composite conjunction phrases such as "on the other hand," "for

example," and "as a result." Fragmentality of learner texts also explains why they underuse DEMP, such as "this" and "those." Too many occurrences of logical connectors and a lack of demonstrative pronouns suggest that learners' texts are less logically sophisticated and less cohesive in comparison to ENS texts.

(10) …(working) also can increase their chances to practice in society. However <CONJ>, some people object this opinion. They claim that part-time jobs will waste student's time and affect their studying. Therefore <CONJ>, they just need to work hard… (CHN)

(11) I think it is important for university students to experience before getting a job. For <CONJ> example, if students who are finished university get a job and have to hardly work, they can not accurately work. (KOR)

(12) there will be students that need to work because of monetary reasons and if that <DEMP> is the case, then… (ENS3)

Simplified verb phrases (Overuse of PROD and underuse of PEAS/VBN)

PROD is overused by 14 learner groups, while PEAS and VBN are underused by 18 and 10 groups, which exemplifies that learners tend to choose simple, versatile verbs such as "do" and use verbs in the simple present tense. A lack of a lexical and temporal variety of verb phrases is one of the salient characteristics of Asian EFL learner essays.

(13) There are some places that students can do <PROD> part time job while they are free.. (THA)

(14) Generally, we usually do <PROD> part time job from the evening to the late at night (KOR)

(15) I have <PEAS> had a part time job only in the last semester of college. I have <PEAS> been <VBN> working at the recreation center weight room… (ENS1)

Structural simplification (Overuse of THATD and underuse of THVC/SPAU)

As mentioned above, learners' text is simple and colloquial, which is also suggested in their overuse of THATD. Even when ENS writers use that-complementizers (THVC), some learners skip them. Also, learners avoid syntactically complex structures such as split auxiliaries (SPAU) by adverb insertions.

(16) I think <THATD> it (i.e., a part time job) is important for us… (KOR)

(17) Many students graduate, get jobs, and find that <THVC> they do not know how to do something for eight hours a day… (ENS3)

(18) … students learn skills that will <SPAU> essentially shape their future… (ENS2)

The analyses above have exemplified many essential gaps between learner essays and ENS essays, some of which were revealed by comparing the frequency of lexicogrammatical features rather than individual words or abstract constructions.

RQ3 Primary Factors for Clustering

Hierarchical clustering analysis produced the tree diagram shown in Figure 5.

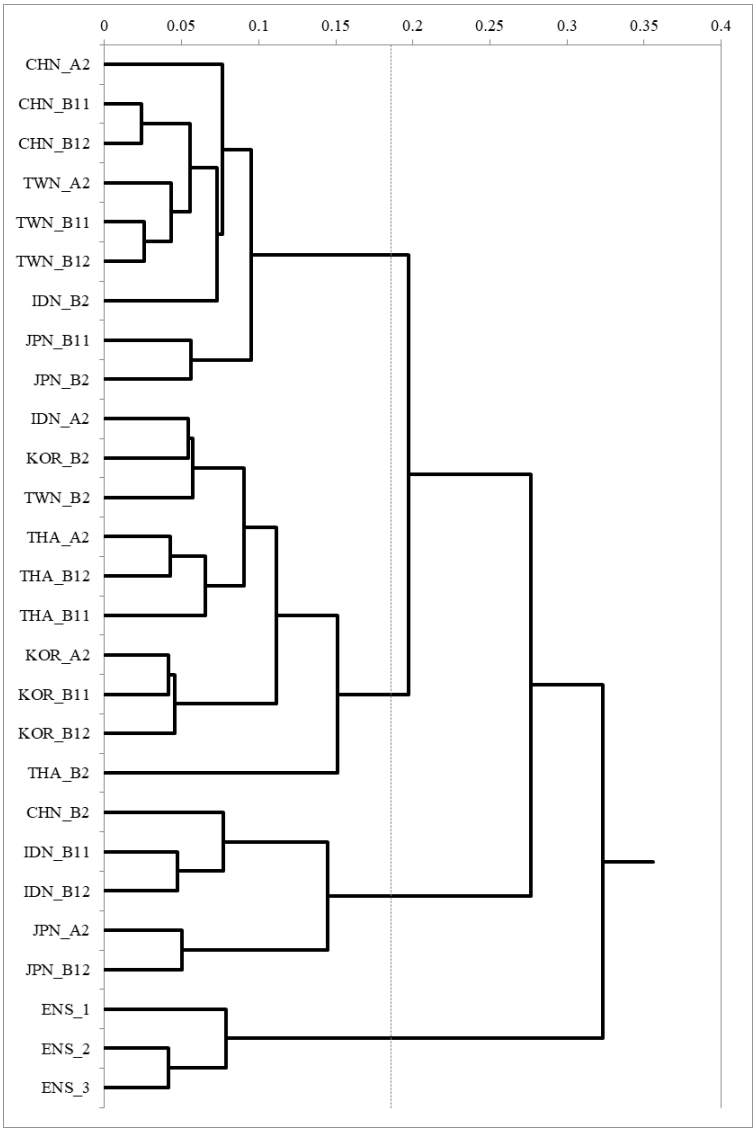

Figure 5. Tree diagram based on the cluster analysis.

Then, which of a country/region and proficiency is a primary factor for the clustering? If proficiency is a more salient factor than a country/region, writer groups at the same proficiency level should come together despite the difference in countries/regions. Such a tendency, however, was not observed. Rather, the diagram shows that many learners are clustered according to the countries/regions rather than proficiency. For instance, three of the four proficiency groups come together quite early for learners from China, Taiwan, Thailand, and Korea.

However, this overall tendency does not wholly negate the influence of a proficiency variable, which is supported by two facts from the diagram above.

First, Japanese learners are divided sharply into the less advanced (A2/B1) and advanced (B1/B2+) groups, which are positioned entirely apart. Proficiency may have a more decisive effect for Japanese learners than the others.

Second, among the four proficiency groups, B2+ learners are often placed apart from the others. As already mentioned, B2+ Japanese learners are apart from novice Japanese learners. B2+ Thai learners are also apart from the other Thai learners. Furthermore, B2+ Korean learners are close not so much to the other Koreans as to Indonesian and Taiwanese learners; B2+ Indonesian learners are close to Chinese and

Taiwanese learners; B2+ Chinese learners are close to Indonesian learners; and B2+ Taiwanese learners are close to Indonesian and Korean learners. These facts may suggest that for many Asian learners, an essential change in writing style occurs not between each of the four levels but between only A2/B1 and B2+; in other words, learners acquire a new writing style only after they reach B2+ level. This trend might be interpreted as independence from factors such as L1 and the type of English education administered in each of the countries/regions. Learners at A2 and B1 levels write largely on the basis of their L1 style and what they have learned at school. While, after reaching B2+, they begin to try a new style. This hypothesis partly explains the peculiar relationship between Chinese and Taiwanese learners sharing the same L1 background: A2/B1 learners belong to the same cluster, while only B2+ learners are excluded. However, independence from L1 and education effect, if any, does not directly mean the acquisition of an ENS-like writing style. In the diagram above, none of the B2+ learners are included in the ENS cluster. B2+ learners seem to be still at the stage of acquiring a new writing style through trial and error.

These findings suggest a complex picture of the primary factor for the clustering. A country/region, which also entails L1 and the type of English education, has a more decisive effect on novice and intermediate learners. However, such a regional effect gradually decreases as learners' proficiency levels go up and reach B2+ level, when they begin to be independent from the L1 and education effect and grope for a new writing style by themselves.

RQ4 Writer/Feature Classification

Finally, the result of the correspondence analysis is shown in Figure 6. The area around the origin (<±0.1) is shaded, which can be interpreted as a neutral area showing no particular lexicogrammatical features. SPP2 (Z1:-2.7, Z2: 0.2) is presented separately, as it seems to be a kind of outlier in the current feature set.

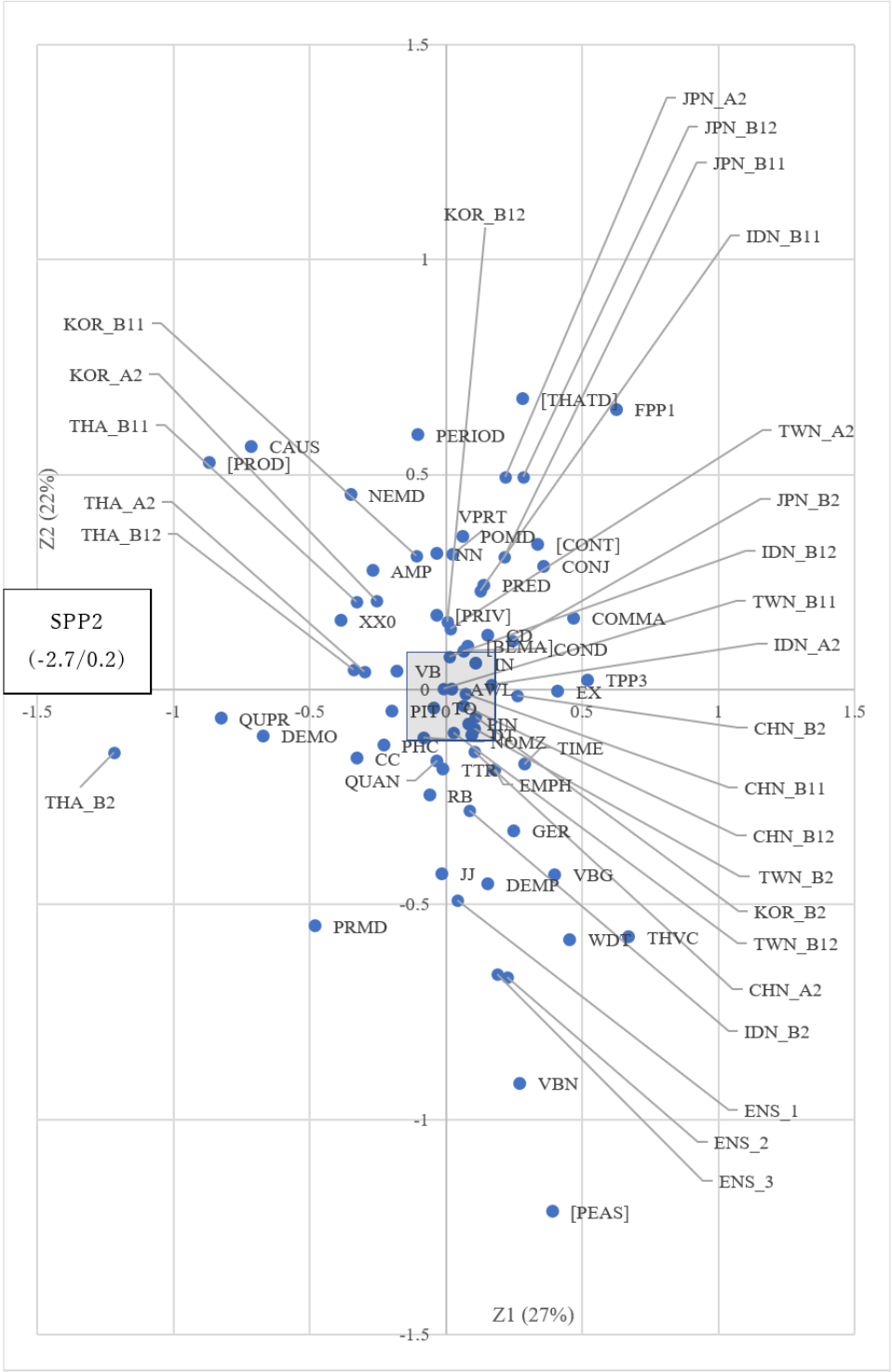

Figure 6. Scatter plot based on the correspondence analysis.

Items positioned close to the origin are usually interpreted as neutral. When excluding the items positioned in the square around the origin (Z1: -.009 to 0.09, Z2: -0.09 to 0.09), the remaining items are classified into four quadrants as shown in Table 5, which summarizes core writing styles seen in the essays of Asian EFL learners as well as ENS.

TABLE 5
Features Summarizing Core Writing Styles for EFL Learners and ENS

Quad.	Key features	Writer-groups	Estimated text types
1 (+/+)	THATD, FPP1, CONT, PRED, COMMA, CD	JPN_A2/B12/B2 IDN_B11	Colloquial/personal
2 (-/+)	PERIOD, CAUS, PROD, NEMD, AMP, XX0, SPP2	KOR_A2/B11 THA_B11	Interactive/persuasive
3 (-/-)	DEMO, PHC, CC, PRMD	THA_B2	Static/descriptive
4 (+/-)	TIME, EMPH, GER, VBG, DEMP, THVC, WDT, VBN, PEAS, SAPU	CHN_A2 ENS2/3	Dynamic/reflective

Quadrant 1 represents a colloquial/personal style, which is characterized by the use of commas, that-deletion, contractions, first-person pronouns, cardinals, and predictive adjectives. Quadrant 2 represents an interactive and persuasive style, which is characterized by the use of second-person pronouns, periods, causative adverbial subordinators, pro-verb do, necessity modals, amplifiers, and analytic negation. Quadrant 3 represents a a static and detailed description style, which is characterized by the demonstratives, phrasal coordination (A and B), coordinating conjunctions, and predictive modals ("will," "shall," etc.). Then, Quadrant 4 represents dynamic and reflective description style covering the time change, which is characterized by a variety of verb forms (gerund, V-ing, V-pp, perfect), time-related expressions, and formal complex structures (emphatics, demonstrative pronouns, that-complementizers, split auxiliaries, wh-determiners).

Conclusion

Major Findings

By combining learner corpus data with two theoretical frameworks—CIA and MDA—, this study scrutinized the lexicogrammatical features of the L2 English essays produced by Asian college students with various country/region backgrounds and at different L2 proficiency levels.

Regarding RQ1 (Differences in Dimension Scores), this study focused on the possible gaps between (a) learners and ENS, (b) learners at different L2 proficiency levels, and (c) learners from different countries/regions. The data analyses showed that (a') in comparison to ENS essays, learner essays are more involved rather than informational (D1), slightly less narrative (D2), less overtly expressive (D4), less abstract (D5), and less time-constraint (D6); (b') as learners' proficiency levels go up, they come to produce less involved and more informational (D1), more narrative-oriented (D2), less explicit and more situation-dependent (D3), more abstract (D5), and more online-developmental (D6) texts and their essays get closer to ENS essays in terms of D1, D2, D5, and D6; and (c') learners from China and Thailand may tend to produce somewhat deviant texts when compared to the other Asian EFL learners.

Regarding RQ2 (Key Lexicogrammatical Features), this study identified a bunch of statistical key lexico-grammatical tags, which suggested five principal orientations characterizing learner essays: (a) conversationality (Overuse of SPP2/FPP1), (b) mixed modality (Overuse of POMD/NEMD), (c) fragmentality (Overuse of PERIOD/COMMA/CONJ and underuse of DEMP), (d) simplified verb phrases (Overuse of PROD and underuse of PEAS/VBN), and (e) structural simplification (Overuse of THATD and underuse of THVC/SPAU).

As regards RQ3 (Primary Factors for Clustering), this study conducted a cluster analysis to clarify which of the regional and proficiency backgrounds influence the lexicogrammatical features of learner essays more saliently. The analysis strongly suggested that, especially for novice and intermediate learners, the regional backgrounds, which are closely related to learners' L1 and the type of English education administered in each area, can be a more decisive factor. However, regional effects might lessen untill

learners' proficiency levels reach B2, when they begin to be independent from the L1 and education effect and grope for a new writing style by themselves.

Finally, as regards RQ4 (Writer/Feature Classification), this study conducted a correspondence analysis to summarize the complex relationships between various text/learner-related variables. The analysis suggested four archetypal essay types for Asian learners: (a) learners from Japan and Indonesia tend to produce colloquial and personal essays, which are characterized by the overuse of commas, that-deletion, contractions, first-person pronouns, cardinals, and predictive adjectives; (b) learners from Korea and Thailand tend to produce interative and persuasive essays, which are characterized by the overuse of second-person pronouns, periods, causative adverbial subordinators, pro-verb do, necessity modals, amplifiers, and analytic negation; (c) advanced-level Thai learners tend to produce static and descriptive essays, which are marked by the overuse of the demonstratives, phrasal coordination, coordinating conjunctions, and predictive modals; and (d) ENS (and partly novice Chinese learners) tend to produce dynamic and reflective essays, which are marked by the overuse of a variety of verb forms (gerund, V-ing, V-pp, perfect), time-related expressions, and formal complex structures (emphatics, demonstrative pronouns, that-complementizer, split auxiliaries, wh-determiners). However, it should also be noted that such a connection between countries/regions and essay types often becomes ambiguous when learners' proficiency levels sufficiently increase.

A CIA/MDA integrative approach adopted in this study enabled a more detailed observation of the features of learner texts than many of the previous CIA studies that focused only on individual words. It also enabled an analysis of learner texts from a broader perspective by discussing the lexis and grammar in combination.

Limitations and Future Plans

The current analysis, however, may have several limitations, mainly in the methodology and interpretation of the findings. Four of these need to be mentioned in particular.

The first one concerns the ambiguity of the construct of a country/region. In learner corpus studies, especially in Europe, researchers tend to discuss "L1," not a "country/region." When considering the fact that (i) many of the European languages are lexically and structurally close to English; (ii) many of the learners generally have a high level of English proficiency; (iii) plural languages are officially used in some countries (e.g., Belgium, Switzerland, and Finland); and (iv) many are plurilingual speakers, it seems to be quite reasonable for researchers to focus on L1 effect. Meanwhile, many of these features do not apply to Asia, where social, psychological, and cultural factors such as teaching methods, teaching materials, entrance exam systems, expectations of parents, teachers, and others, motivations to learn English, and cultural restrictions on explicit self-expression, are likely to play a much more important role than L1. Therefore, the ICNALE project has put a greater emphasis on a country/region rather than on L1. However, what causes the regional differences and how L1 influences them need to be analyzed more carefully.

The second one concerns the generalizability of the findings. This study analyzed highly homogeneous texts (i.e., 200–300 words essays written about the topic of a part-time job for college students). This made it possible to control the possible effects of various parameters, such as the essay length and topics, and reliably discuss the gaps between ENS and learners, learners at different proficiency levels, and learners from different countries/regions. Meanwhile, whether the findings from a limited type of essay data can be applied to any type of learner essay is not clear. This suggests the need to analyze a greater variety of learner essays to clarify what part of the findings applies to Asian learners' essay writing in general and what is limited to the current samples.

The third one concerns the accuracy of the data processing with MAT. After reliability tests based on two L1 English corpora (LOB and Brown), Nini (2015b, 2019) concludes that "MAT is largely successful in replicating Biber's (1988) analysis" and "MAT can be used to assign Biber's (1988) Dimension scores to texts" and "to categorize a text for its text type, as proposed by Biber (1989)." However, whether MAT

appropriately processes learner texts that can be essentially different from L1 English has not been fully proven. This suggests the need for manual checking of the results of automatic tagging with MAT.

Finally, the fourth one concerns the appropriateness of using ENS data as a yardstick of comparison. This study considered the possible internal variety of an ENS yardstick by using three kinds of ENS datasets (students, teachers, and others). However, it did not attempt to use non-ENS references. A yardstick could theoretically include English varieties in the Outer Circle (Kachru, 1991) and English as a lingua franca or ELF (Kirkpatrick, 2012). Granger carefully replaced the word "native" with "reference" when she revised her CIA model in 2015. Gilquin (2022) also mentioned the risk in seeing ENS outputs as "one norm to rule them all" and recommended adopting a much wider range of corpus-based referential yardsticks. This may suggest reconsidering the traditional dichotomy between learners and ENS. The author once proposed to use high-rated learner outputs as a new reference in CIA studies (Ishikawa, 2023b). Such alternative reference varieties would also be worth considering when applying a CIA/MDA approach to learner outputs.

Although several things remain to be reconsidered in future studies, a series of findings from this analysis would still be significant to learner corpus research in general and English language teaching. A combination of CIA and MDA is a powerful analytical method in that it enables SLA researchers to discuss the development process in L2 writing in greater detail, and it also allows teachers to clarify the gaps between standard L1 essays and students' L2 essays and offer more appropriate and informative feedback to the students.

Acknowledgments

This study was supported by Grants-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (20H01282 and 23H00641). A part of this paper was orally presented at the 21st Asia TEFL International Conference held in Daejeon, Korea, from August 17 to August 20, 2023. The author appreciates the valuable comments from the participants. The author also expresses sincere gratitude to the two anonymous reviewers for their insightful comments and suggestions, which greatly improved this paper.

The Author

Shin'ichiro (Shin) Ishikawa is Professor of applied linguistics at Kobe University, Japan. His research interests cover the fields of applied linguistics, corpus linguistics, statistical linguistics, and foreign language teaching. He is a leader of the ICNALE corpus project.

School of Languages and Communication
Kobe University,
1-2-1, Tsurukabuto, Nada-ku, Kobe, Japan
Tel: 81-78-881-1212
Email address: iskwshin@gmail.com

References

Altenberg, B., & Tapper, M. (1998). The use of adverbial connectors in advanced Swedish learners' written English. In S. Granger (Ed.), *Learner English on computer* (pp. 80–93). Longman.
Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.
Biber, D. (1989). A typology of English texts. *Linguistics*, *27*(1), 3–43.
Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.
Gao, X. (2023). A comparable corpus-based study of phrasal verbs in academic writing by English and Chinese scholars across disciplines. *Corpora*, *18*(2), 219–244.
Gilquin, G. (2022). One norm to rule them all? Corpus-derived norms in learner corpus research and foreign

language teaching. *Language Teaching*, *55*(1), 87–99. https://doi.org/10.1017/S0261444821000094

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

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, *1*(1), 7–24.

Granger, S., Dagneaux, E., & Meunier, F. (2002). *International corpus of learner English*. Press universitaires de Louvain.

Halliday, M.A.K. (1991). Corpus studies and probabilistic grammar. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics: Studies in honour of Jan Svartvik* (pp. 30–40). Longman.

Ishikawa, S. (2023a). *The ICNALE guide: An introduction to a learner corpus study on Asian learners' L2 English*. Routledge.

Ishikawa, S. (2023b). A new yardstick of comparison for contrastive interlanguage analysis: A study on the ICNALE Global Rating Archives. In U. Widiati, M. Hidayati, N. Suryati, Suharyadi, A. N. Wulyani, I. L. Damayanti, N. A. Drajati, S. Karmina, E. L. Zen, L. Hakim, & Prihantoro (Eds.), *Proceedings of the 20th AsiaTEFL-68th TEFLIN-5th iNELTAL Conference* (ASIATEFL 2022) (pp. 607–619). Atlantis Press. https://doi.org/10.2991/978-2-38476-054-1_52

Kachru, B. B. (1991). Liberation linguistics and the Quirk concern. *English Today*, *7*(1), 3–13. https://doi.org/10.1017/S026607840000523X

Kirkpatrick, A. (2012). English as an Asian lingua franca: The 'lingua franca approach' and implications for language education policy. *Journal of English as a Lingua Franca*, *1*(1), 121–139. https://doi.org/10.1515/jelf-2012-0006

Kwon, H. (2022). English learner corpora and research in Korea. *Corpora*, *17*, 5–22. https://doi.org/10.3366/cor.2022.0244

Leedham, M. (2015). *Chinese students' writing in English: Implications from a corpus-driven study*. Routledge.

Lin, Y-C., & Chung, S-F. (2022). A corpus-based study of native speakers' and Taiwanese EFL learners' use of the adverb just. *Corpora*, *17*, 99–117. https://doi.org/10.3366/cor.2022.0249

Lorenz, G. (1998). Overstatement in advanced learners' writing: Stylistic aspects of adjective intensification. In S. Granger (Ed.), *Learner English on computer* (pp. 53–66). Longman.

Lu, Y. (2017). *A corpus study of collocation in Chinese learner English*. Routledge.

Nini, A. (2015a). Multidimensional analysis tagger (v. 1.3) [Computer software]. http://sites.google.com/site/multidimensionaltagge

Nini, A. (2015b). Multidimensional analysis tagger (v. 1.3) manual. https://drive.google.com/file/d/11BAw-Dl5FDLTJFXqixc2uIgQQtU0g0no/view?pli=1

Nini, A. (2019). The multidimensional analysis tagger. In T. B. Sardinha & M. V. Pinto (Eds.), *Multidimensional analysis: Research methods and current issues* (pp. 67–94). Bloomsbury Academic.

Whitty, L., Parkinson, J., & Pham, H. T. P. (2022). *Can* and *could* in academic writing: A corpus-driven comparison of English L1 and Vietnamese EFL students. *The Journal of Asia TEFL*, *19*(1), 93–108. http://dx.doi.org/10.18823/asiatefl.2022.19.1.6.93

Yoo, I. W., & Shin, Y. K. (2022). English lexical bundles in a learner corpus of argumentative essays written by Korean university students. *Corpora*, *17*, 23–42. https://doi.org/10.3366/cor.2022.0245

Zhi, Y. (2022). Investigating MAKE patterns in learner English writing: What can frequency tell? *The Journal of Asia TEFL*, *19*(2), 577–591. http://dx.doi.org/10.18823/asiatefl.2022.19.2.11.577

(Received October 28, 2023; Revised February 13, 2024; Accepted March 10, 2024)

The Journal of Asia TEFL
Vol. 21, No. 1, Spring 2024, 119-138

Appendix

Code	Meaning
AMP	amplifier
AWL	word length (in letters)
BEMA	*be* as main verb
CAUS	causative adverbial subordinators (because)
CC	coordination conjunctions (and, but, or/nor, for, etc.)
CD	cardinal number
COMMA	comma
COND	conditional adverbial subordinators (if/unless)
CONJ	conjuncts (however, therefore, in addition, for example, as a result, etc.)
CONT	contractions
DEMO	demonstratives (this, that, these, those +N)
DEMP	demonstrative pronouns
DT	determiners
EMPH	emphatics (just, really, most, so, do+V, a lot, etc.)
EX	existential there
FPP1	first person pronouns
GER	gerunds (V-ing(s) nominal forms of 10+ letters)
IN	subordinators
JJ	adjectives
NEMD	necessity modals (ought, should, must)
NN	total other nouns
NOMZ	nominalisations (-tion/ment/ness/ity)
PEAS	perfect aspect
PERIOD	periods
PHC	phrasal coordination (and)
PIN	prepositions
PIT	pronoun it
POMD	possibility modals (can, may, might, could)
PRED	predicative adjectives
PRIV	private verbs (conclude, expect, feel, hope, know, mean, suggest, think, etc.)
PRMD	predictive modals (will, would, shall)
PROD	pro-verb do
QUAN	quantifiers (each, all, every, many, some, any, etc.)
QUPR	quantifier pronouns (everyone, someone, everything, something, etc.)
RB	adverbs
SPAU	split auxiliaries (they are objectively shown that...)
SPP2	second-person pronouns
THATD	subordinator that deletion (e.g., I think [φ that] ...)
THVC	that verb complements (e.g., I think that...)
TIME	time adverbials (again, early, late, now, today, etc.)
TO	infinitives
TPP3	third person pronouns
TTR	type-token ratio
VB	verb, base form
VBG	present participial form of a verb
VBN	past participial form of a verb
VPRT	present tense
WDT	wh-determiner (e.g., which book...)
XX0	analytic negation (not/n't)