The ICNALE and Sophisticated Contrastive Interlanguage Analysis of Asian Learners of English

Shin'ichiro ISHIKAWA
Kobe University

Abstract

The International Corpus Network of Asian Learners of English (ICNALE) is a new learner corpus designed for a reliable contrastive interlanguage analysis of varied English learners in Asia. The ICNALE, in which writing conditions are controlled more strictly compared with other major learner corpora, allows researchers to examine the differences between writer groups in greater detail. The current paper outlines the features of the ICNALE and demonstrates how it can contribute to the sophistication of contrastive interlanguage analysis.

Keywords

ICNALE, Learner Corpus, Contrastive Interlanguage Analysis

I Introduction

1.1 The ICLE and Learner Corpus Studies

The history of learner corpus studies dates back to October 1990, when Professor Sylviane Granger began to collect writings by English learners who speak French as their first language (L1). The project was gradually enlarged, and the first version of the International Corpus of Learner English (ICLE) was released in 2002. This corpus contained 2.5 million words of essays written by learners with 11 different L1s: Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish, and Swedish. The second version was released in 2009, and with the new addition of the data of learners of L1 Chinese, Japanese, Norwegian, Turkish, and Tswana backgrounds, the size of the corpus has reached 3.7 million words. As the editors of the corpus hoped, the ICLE project clearly marked "a new stage in the evolution of EFL research" (Granger et al., 2002, p. 1).

Due to productive research in the field, the possibilities of learner corpus studies have come to be widely understood. Borin & Prütz (2004) mention that the learner corpus, though it is relatively new, has become "one of the most important resources for

studying interlanguage." Nesselhauf (2004) emphasizes that for language teaching, "it is not only essential to know what native speakers typically say, but also what the typical difficulties of the learners of a certain language, or rather of certain groups of learners of this language, are" and mentions the need for learner corpora in addition to native speaker corpora.

1.2 Contrastive Interlanguage Analysis

The contribution of the ICLE project to applied linguistics is not limited to the development of the corpus itself. As Hasselgård & Johansson (2011) write, a "special feature of the ICLE project is that a framework of learner corpus research has been developed alongside the corpus." The corpus development team has proposed various approaches and methodologies to examine the interlanguage of various groups and types of learners. Papers collected in Granger (Ed.) (1998) clearly showcase how learner data can be analyzed and how the findings can be applied to language education.

Especially, the analytical procedure called contrastive interlanguage analysis (CIA) has been established as a standard approach to examine learners' second language (L2) use. Granger (1998) introduces two types of CIA: One is the NL vs. IL contrast, namely, a comparison of native language and interlanguage to "uncover the features of non-nativeness of learner language." The other is the IL vs. IL contrast, namely, a comparison of different interlanguages to "gain a better insight into the nature of interlanguage." Gilquin et al. (2008) says that the former typically deals with misuse, overuse, and underuse, while the latter addresses developmental factors such as age and proficiency beyond the difference of L1s.

Although there are some critical views on NL vs. IL or L1 vs. L2 contrasts, it is clear that they are "extremely powerful heuristic techniques which help bring to light features of learner language which have not been focused on before, and which, once uncovered, can be analyzed from a strictly L2 perspective" (Granger, 2009).

CIA can be conducted independently or in combination with a traditional contrastive analysis (CA) between L_x and L_y. Granger (1996) proposes the Integrated Contrastive Model (ICM), in which CA and CIA are fruitfully connected by the viewpoints of prediction and diagnosis. Borin & Prütz (2004) schematize fives types of related studies: (1) IL_x vs. L2 (classical interlanguage analysis), (2) L2 vs. IL_y (CIA), (3) L1 and L2 (traditional CA), (4) L1 vs. IL_x (study of L2 influence on interlanguage), and (5) IL_x vs. IL_y (comparison of different groups of learners).

CIA studies have revealed many noteworthy facts about the interlanguages of various learners, but there is still room for further sophistication. For, many of the previous CIA studies do not pay sufficient attention to the possibility of internal variety inherent in each of the learner groups. Tono (2009) notes that previous learner corpus projects "largely ignore the educational contexts in each country and they assume that the findings could be applicable for advanced learners of English in general. This is

reasonable as long as the performance of advanced learners is relatively stable and less vulnerable to a specific learning environment in each country. In the case of younger or less advanced learners, however, observed data are heavily dependent upon the nature of input and interaction in the classroom." Ishikawa (2010) also emphasizes the need to consider differences of individual learners in terms of L2 proficiency level, motivation, and L2 learning history when discussing the features of some learner groups, which he calls a multi-layered CIA (MCIA).

1.3 Learner Corpora in Japan

The great success of the ICLE project in the development of a large international corpus and analytical methods for its use has led to world-wide growth in learner corpus studies. Especially in Asian countries, which are not wholly covered in the ICLE, many projects to compile a local learner corpus have been carried out.

In Japan, the National Institute of Information and Communications Technology (NICT) released the NICT Japanese Learners of English (NICT JLE) Corpus (Izumi et al., 2004). This corpus contains 960,000 words of transcribed speeches by 1,281 Japanese learners of English at various proficiency levels, which are recorded in the oral proficiency interviews (OPIs) conducted in the nation. The NICT JLE Corpus is known as the world's largest corpus of learners' speech.

Professor Yukio Tono released the Japanese EFL Learner (JEFLL) Corpus (Tono, 2007), which is a collection of 670,000 words of essays written by 12,000 Japanese students studying at junior or senior high schools. The JEFLL corpus is unique in that it focuses on novice learners, who have received little attention in the previous learner corpus studies.

Professor Masatoshi Sugiura released the Nagoya Interlanguage Corpus of English (NICE) (Sugiura et al., 2007), which contains 70,000 words of essays written by 217 Japanese college students and 120,000 words of essays written by 200 English native speakers. Although the NICE is a relatively small corpus, it is carefully designed on the basis of a critical analysis of major learner corpora previously created. Sakaue et al. (2008) reconsider the ICLE's data collection scheme and point out its limited controls on (i) the writers' proficiency, (ii) the number of topics, and (iii) the writing conditions, such as time or the use of dictionaries.

These three corpora have revealed many interesting facts about Japanese learners of English in comparison to English native speakers, but so far no corpora have enabled both an internal comparison of Japanese learners at different proficiency levels and an external comparison of Japanese learners with other Asian learners in different countries and areas.

II The International Corpus Network of Asian Learners of English

2.1 What is the ICNALE?

The International Corpus Network of Asian Learners of English (ICNALE) is a collection of 1.3 million words of essays written by 2,600 college students in 10 Asian countries and areas plus 200 English native speakers. It is one of the largest learner corpora publicly available and practically the sole learner corpus focusing on various Asian learners.

The ICNALE is designed as a reliable database for sophisticated international CIA as well as for studies of the World Englishes in Asia.

2.2 A Brief History of the Development of the ICNALE

The author released the Corpus of English Essays Written by Asian University Students (CEEAUS) in 2009, which includes approximately 170,000 words of essays written by Japanese learners, 20,000 words of essays by Chinese learners, and 40,000 words of essays by English native speakers. The CEEAUS is still available in the CD-ROM accompanying the book on statistical linguistics (Ishikawa et al., 2010).

Although the CEEAUS gained some attention from local scholars, it was clear that the size, variety of writers, and data control were far from satisfactory. For example, some students wrote two essays, while others wrote only one; Chinese students' proficiency was not investigated; and the nationalities of native speakers were not considered at all.

Therefore, the author reconsidered the entire data collection scheme and expanded the CEEAUS to cover a greater diversity of writers in Asia and to enable it to be used as a more reliable database for international contrastive studies.

After a trial data collection and reflection on its scheme, essay data were collected in Japan, Hong Kong, Pakistan, Thailand, China, Taiwan, Korea, and Indonesia in 2011; and the native speakers' essays were also collected in the same year. Next, data were collected in the Philippines and Singapore in 2012.

In the data collection, all of the writers, including both non-native speakers and native speakers, were told to write essays under strictly controlled conditions; that is, they wrote about the same topic within the same amount of time, and they produced essays of the same length and using the same PC environments and references. Also, the writers' personal characteristics, L2 proficiency, L2 learning background, and experiences were investigated in as much detail as possible.

Thus, the ICNALE has collected data in 10 countries and areas of Asia in addition to the data of English native speakers. The data size has now reached approximately 1.3 million words. The 1.0 version of the ICNALE was released in December 2012 and the 2.0 version in January 2013.

As part of the ICNALE project, we also developed an online corpus query system,

called the ICNALE *Online*, in order to catch up with the recent trend in corpus studies characterized by use of "the fourth-generation" internet-based concordancers (McEnery & Hardie, 2012, p. 43). The first version of the ICNAE *Online* was released in 2010, the second version in 2011, and the final version in January 2013.

2.3 Key Features of the ICNALE

2.3.1 Focus on Asian Learners

Previous learner corpus studies have dealt mainly with European learners, paying relatively limited attention to Asian learners. However, recent economic, socio-cultural, and linguistic globalization has boosted the number of English learners in Asia, where the factors concerning learners of English, for example, overall L2 proficiency, opportunities to use English in everyday contexts, motivation to study English, or the needs for English in society, are essentially different from those in Europe.

Therefore, we attempted to collect a sufficient amount of data of Asia learners in the ICNALE project. Table 1 shows the countries and areas represented in the corpus and the amount of data collected in each of them. Tokens 1 and 2 represent the number of words in the download version and the online version (The ICNALE *Online*), respectively. In the online version, some of the punctuation and symbols in the original essays are deleted for speedy data queries.

Table 1 Countries and Areas Represented in the ICNALE

Type	Code	Countries	Writers	Essays	Tokens 1	Tokens 2
ENL	ENS	USA,UK,AUS, etc.	200	400	90,613	88,792
ESL	HKG	Hong Kong	100	200	47,505	46,111
	PAK	Pakistan	200	400	94,523	93,100
	PHL	The Philippines	200	400	99,463	96,586
	SIN	Singapore	200	400	99,267	96,733
	Total	---	700	1,400	340,758	332,530
EFL	CHN	China	400	800	202,725	194,613
	IDN	Indonesia	200	400	93,277	92,316
	JPN	Japan	400	800	179,042	176,537
	KOR	Korea	300	600	136,346	130,626
	THA	Thailand	400	800	181,120	176,936
	TWN	Taiwan	200	400	92,384	89,736
	Total	---	1,900	3,800	884,894	860,764
ALL	Total	---	2,800	5,600	1,316,265	1,282,086

When discussing Asian learners, we need to note that the social status of English is not necessarily homogeneous in the region. Thus, the ICNALE includes essays written

by native speakers using English as a native language (ENL), learners using English as a second language (ESL), and learners using English as a foreign language (EFL). These are in accordance with the inner circle, the outer circle, and the expanding circle in the typology of World English users proposed by Kachru (1985)(see Fig. 1).

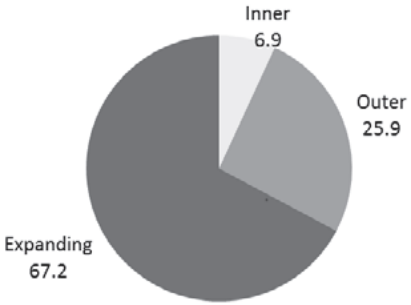


Fig. 1 Proportions of the three circles covered in the ICNALE (%)

This great variety in the data coverage gives a unique status to the ICNALE as a database for studying both the interlanguage of Asian learners of English and the World Englishes in Asia.

2.3.2 Control of Writing Conditions

There are two basic directions in the development of learners' essay corpora. One is to loosen the writing conditions, such as topics and time for writing, so as to collect as much of a variety of essays as possible. The other is to strictly control writing conditions so as to make the corpus data as homogeneous as possible. The former is suitable mainly for an exploratory analysis of various facets of the interlanguage of a particular writer group, while the other is appropriate for a contrastive study of different writer groups.

It can be problematic to conduct a comparison of different writer groups using the former type of corpus. For example, when we compare a timed essay written by a French learner about the importance of nature and an untimed essay written by a Chinese learner about his or her culture, it is extremely difficult to interpret the comparison results. As Ädel (2008) illustrates, we are prone to confusing the difference in writing conditions with that of writer groups.

The ICNALE is designed as a database primarily for the latter type of CIA. Therefore, factors that might influence the language are carefully controlled, and the same instructions are given to all the writers, including both native and non-native speakers. Fig. 2 shows the instruction sheet.

Do you agree or disagree with the following statements? Use reasons and specific details to support your opinion.
(Topic A) It is important for college students to have a part-time job.
(Topic B) Smoking should be completely banned at all the restaurants in the country.

Instructions
1. Clarify your opinions and show the reasons and some examples.
2. You can use 20 to 40 minutes for each essay. This means that you have 40 to 80 minutes to complete two essays. Do not finish too early or spend too much time.
3. You must use MS Word or a similar word processor.
4. Do not use dictionaries or other reference tools.
5. Do not plagiarize anyone else's essays.
6. The length of your single essay should be from 200 to 300 WORDS (not letters). Too short or too long essays cannot be accepted. You can check the length of your essay using the word count function of MS Word.
7. You must run spell check before completing your writing.

Fig. 2 Instruction sheet given to learners

Requiring writers to use a word processor has two benefits. First, the processes of data collection and processing are greatly facilitated. Second, use of the word processor seems to naturally urge learners to write longer. Pennington (2003) suggests that the "student writer working in a computer medium is led to write in a less self-conscious way and with greater engagement, thus writing with a freer mind and less 'rewriting anxiety.' As a result, the student's greater involvement may lead him or her to write for longer periods of time and produce longer texts."

As explicitly shown in the instruction above, the topic, time, length, and use of references are all standard, which makes the language collected in the corpus quite homogeneous. Thus, the ICNALE allows us to conduct a more sophisticated international contrastive analysis than the existing corpora do.

2.3.3 Control of L2 Proficiency

Another factor critically influencing the language is writers' L2 proficiency. Collecting the data of writers at the same specific level may be ideal, but doing so is extremely difficult for corpus developers, who usually prioritize the corpus size.

The ICLE is generally known as a collection of essays written by advanced learners, but, based on the manual evaluation of 20 randomly sampled essays from each of the different writer groups, the editors admit that "some of the *ICLE v2* subcorpora are rather in the higher intermediate range," although approximately 60% of the essays belong to the advanced level (Granger et al., 2009, pp. 11–12).

If collecting data exclusively from learners at a specific proficiency level is practically

impossible, we should investigate individual writers' L2 proficiency based on objective external criteria. In the ICNALE project, therefore, we firstly investigated writers' scores on the major English proficiency tests such as TOEIC, TOEFL, or IELTS as an objective measure of their proficiency levels. Also, for the purpose of showing different test scores on the same cline, we mapped test scores onto the proficiency bands defined in the Common European Framework of Reference (CEFR), as shown in Fig. 3.

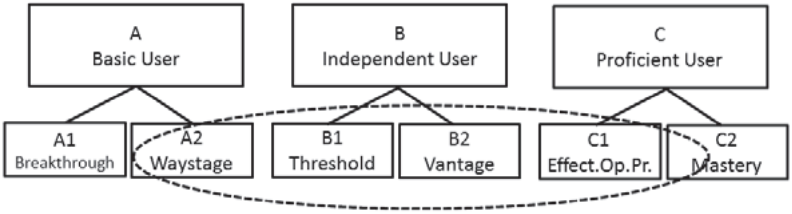


Fig. 3 Proficiency bands in the CEFR

Although the proficiency is classified into six levels, A1 (Breakthrough), A2 (Waystage), B1 (Threshold), B2 (Vantage), C1 (Effective Operational Proficiency), and C2 (Mastery) in the CEFR, we deleted the A1 level, merged B2, C1, and C2 into B2+, and subdivided B1 into B1_1 and B1_2 in order to represent Asian learners' variety of L2 proficiency in a more appropriate way. In the mapping of test scores, we followed the official conversions presented by test institutes such as ETS and Cambridge ESOL.

Many of learners in Korea and Japan have reported the TOEIC or TOEFL test scores, but this is not the case with learners in other countries and areas. Therefore, we required all the learners to take the English vocabulary size test (VST) (Nation & Beglar, 2007). The VST has been widely used in EFL education, and it is suggested that L2 vocabulary knowledge measured by the VST is robustly correlated with general L2 proficiency. As Meara & Milton (2003) and Milton (2010) state that it is appropriate to measure the vocabulary size of non-native speakers with a ceiling of 5,000 words, we used only the 50 test items up to the 5,000 word level. Although the original VST is in a pencil-and-paper format, we prepared an MS Excel version and integrated it into the essay submission sheet. This means that all the writers firstly had to take the VST and then write two essays. Fig. 4 shows an example question from the VST.

There seem to be no reliable conversions between vocabulary size and CEFR levels. For instance, based on the analysis of Greek and Hungarian EFL learners, Meara & Milton (2003) relate the size of 2500+ words to B1, 3250+ words to B2, 3750+ words to C1, and 4500+ words to C2. However, this conversion leads to a great overestimation of the proficiency of Asian writers. Therefore, based on the data of 268 learners who took both the VST and the TOEIC test, we conducted linear regression modeling and obtained a formula converting the VST score (0–50) to a TOEIC test score (10–990):

TOEIC = 10.495*VST+289 (R = .441). Fig.5 shows the relationship between VST score (horizontal axis) and TOEIC test score (vertical axis). Although the correlation index is not sufficiently strong, we judged that this conversion can be used with a certain degree of practical reliability.

Circle the letter a–d with the closest meaning to the key word in the question.
COMPOUND: They made a new compound.
a. agreement
b. thing made of two or more parts
c. group of people forming a business
d. guess based on past experience

Fig. 4 A sample question on the VST (4,000 word level)

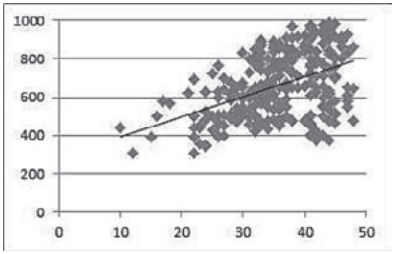


Fig. 5 Scatter plot (VST score and TOEIC test score)

Thus, those who had not taken any standard proficiency tests were also classified into four CEFR levels. Table 2 shows the ratio of writers in each proficiency band by country/area.

Table 2 Ratios of Writers at the Four Proficiency Levels (%)

	Area	A2	B1_1	B1_2	B2+
ESL	HKG	1.0	30.0	52.0	17.0
	PAK	9.0	45.5	44.0	1.5
	PHL	1.0	5.5	88.0	5.5
	SIN	0.0	0.0	67.0	33.0
EFL	CHN	12.5	58.0	26.3	3.3
	IND	16.0	41.0	41.5	1.5
	JPN	38.5	44.8	12.3	4.5
	KOR	25.0	20.3	29.3	25.3
	THA	29.8	44.8	25.0	0.5
	TWN	14.5	43.5	30.5	11.5

As easily expected, the average proficiency levels are generally higher in ESL countries in the outer circle than in EFL countries in the expanding circle. Fig. 6 shows the ordering of 10 writer groups by proficiency based on the accumulated ratios of B1_2 and B2+ writers.

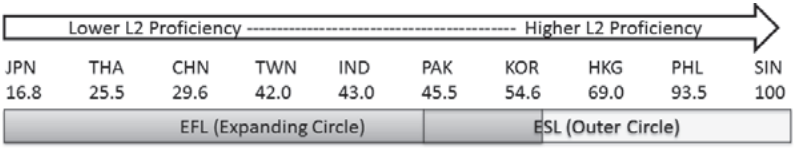


Fig. 6 The order of ten countries and areas based on the ratios of B1_2 and B2+ writers

It should be noted that there is a considerable degree of discrepancy even within the same circle. As shown in the figure, the expanding and outer circles overlap. If someone carelessly compares the entire data of Japanese learners and the entire data of Korean learners, what they contrast is not so much nationality and L1 as L2 proficiency. This clearly shows the importance of tallying the writers' L2 proficiency levels when comparing different writer groups. With the ICNALE, users can easily conduct a proficiency-adjusted comparison, for example, between Japanese learners and Korean learners at the same B1_2 level, which leads to greater sophistication of conventional CIA.

2.3.4 Survey of Writers' Background Information

In addition to major factors that influence the language of essays, such as essay topics, writing conditions, and writers' proficiencies, we also investigated minor factors such as writers' sex, age, academic major, motivations for learning, and L2 learning experiences. Consideration of these minor factors can also lead to a sophistication of conventional CIA.

Using an Excel-based questionnaire sheet (see Fig. 7), which also includes the vocabulary size test mentioned above, we collected three types of information on the writers' background: basic attributes, motivation in L2 Learning, and L2 learning experiences.

Concerning basic attributes, the writers' sex, age, grade, number of years studying English, college major, academic area (humanities, social sciences, science and technology, or life science) were surveyed.

Then, concerning motivation, learners were asked to report how they felt about the 12 statements on a scale from 1 (Strongly disagree) to 6 (Strongly agree): e.g., *I study English because I find pleasure when I understand the content sufficiently* or *I study English because I want to get a better job in the future*.

	A	B	C	D	E	F
1		**STEP2 Questionnaire**				
2		**This is an enquiry about why and how you have studied English.**				
3		**Fill out the pink cells with an appropriate number from 1 to 6.**				
4						
5	I study English because…					
6	1	I find pleasure when I understand the content sufficiently.			6	Strongly agree
7	2	I want to get a better job in future.			5	Agree
8	3	learning content is more important than being awarded high grades.			4	Somewhat agree
9	4	I want to be socially acknowledged.			3	Slightly agree
10	5	being awarded high grades is important for me.			2	Disagree
11	6	learning English is what we have to do anyway.			1	Strongly Disagree

Fig. 7 The Excel-based questionnaire sheet

Based on the accumulated points, learners' motivations were classified as either integrative or instrumental. The former is related to one's interests in the L2 and its culture and also to the desire to have interactions with or be integrated into its speech community, while the latter is related to the desire to achieve some practical goal (passing an exam, achieving a better grade, finding a better job, developing a job skill, winning a promotion, etc.) by using the L2 as an instrument. Integrative and instrumental motivations are roughly in accordance with intrinsic and extrinsic motivations. Some studies imply that integrative (intrinsic) motivation is often more effective than its counterpart in the long term. Although this kind of dichotomy is too simple and seems rather old-fashioned in the recent motivation studies, it still works in many of the countries and areas of Asia.

Finally, concerning experiences, learners are asked to report how well the 20 statements match their own experiences of learning English on a scale of 1 (Strongly disagree) to 6 (Strongly agree). A series of questionnaires intended to survey three basic elements in L2 learning: when (period), where (in or out of class), and what (skill type). They include, for instance, *In my secondary school days, I listened to English a lot in class*, *In my college days, I read English a lot outside class*, *So far in my life, I have been taught by English native speakers*, and *So far in my life, I have been taught essay writing*.

Comparing learners' essays based not on their L1s or nationalities but on their basic personal traits has rarely been attempted to date. However, Ishikawa & Ishikawa (forthcoming) reveal that learners with integrative motivation tend to use more complicated vocabulary.

2.3.5 Collection of Data of Native Speakers

In learner corpus studies, the comparison between an interlanguage (IL) and an L1, in other words, that between some particular non-native writer group and English native speakers, is widely conducted as one of the "extremely powerful heuristic techniques" (Granger, 2009). However, if two kinds of data are collected in different ways, the comparison becomes less reliable and powerful.

Therefore, the ICNALE makes a range of English native speakers write essays about

the same topics under the same writing conditions as learners, which also contributes to the sophistication of contrastive analysis.

The total number of English native speakers (ENSs) participating in the project is 200, which comprises (i) college students (100 writers) and (ii) non-college students (100 writers), most of whom are English teachers and instructors or professional business persons. The former are suitable for comparison with learners, all of whom are college students, while the latter are more, though not perfectly, suitable for investigation of a model of native speakers' essay writing. For, as Leech (1998) warns, "Native-speaking students do not necessarily provide models that everyone would want to imitate."

We paid attention to a balance in the nationalities of native speakers, as some Asian countries such as Japan, Korea, and the Philippines set American English (AmE), and other countries such as Pakistan and Singapore set British English (BrE), as a standard for their English education. The ICNALE covers both of these major types of English. Table 3 shows the percentages of nationalities of the native speakers.

Table 3 Percentages of Nationalities of English Native Speakers (%)

AmE/BrE	Country	Percent	
AmE	USA	57.0	57.0
BrE	UK	14.0	43.0
	Canada	14.0	
	Australia	8.5	
	New Zealand	6.5	

Inclusion of the data of English native speakers in Australia and New Zealand gives additional value to the ICNALE, for these are the two countries belonging to the inner circle in (greater) Asia. Thus, the ICNALE, which covers all of the inner, outer, and expanding circles in Asia, can be used as a database for studying not only interlanguages in Asia but also World Englishes in the region.

2.3.6 Dual Access

The ICNALE is publicly available under the creative common license in two forms: the download version and online version.

The download version is intended to be used mainly by professional researchers. Analysts can explore the entire corpus data freely based on their own interests and perspectives, using a concordance software or computer program for data queries. By referring to the individual writers' detailed background information, which is also downloadable as a spreadsheet, they can exclusively select the essays they want to focus, for instance, those written by Korean female learners, who major in engineering, study English mainly for instrumental purposes, have never been taught English essay writing, and whose TOEIC test score is between 700 and 750. This kind of pinpoint

selection of data is available only in the download version.

Meanwhile, the online version, which is called the ICNALE *Online*, is intended to be used by a wider range of users, especially teachers and learners who are not accustomed to using corpus queries.

The interface of the ICNALE *Online* is highly user-friendly, and users can freely set a retrieval setting. Fig. 8 shows a screenshot of the interface.

Fig. 8 The basic interface of the ICNALE *Online*

Users can make a quick selection of lemmatization (word form or lemma), case (insensitive or sensitive), parts of speech, writer groups, topic, and the number of results to be displayed.

When some writer groups are selected, an L2 proficiency selection window automatically pops up, and users can easily choose the proficiency level(s) they want to focus (see Fig. 9).

Fig. 9 Selection of L2 proficiency levels

The ICNALE *Online* currently offers four kinds of searches: concordance search, collocation search, wordlist search, and keyword search, all of which are common and well established techniques in corpus linguistics.

First, the concordance search enables analysts to see how a target word is used in the real textual context. They can browse the Key Word in Context (KWIC) concordance lines to observe the behavior of a target word (see Fig. 10). Users can designate one to three concurrent words as a target (e.g., "believe," "I believe," and "I believe that") and also specify other words occurring in a certain width of span.

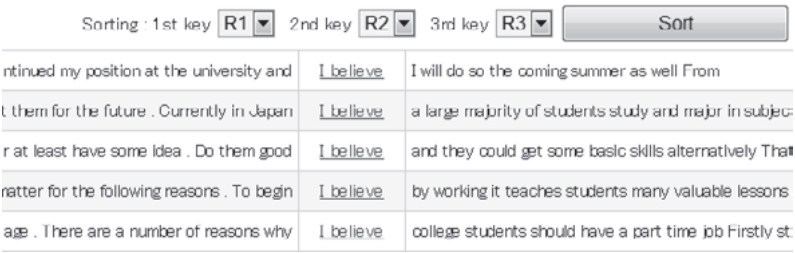


Fig. 10 Concordance search results ("I believe" used by native speakers)

Concordance lines can be freely sorted based on the collocate occurring in a particular position, and the search results are easily downloadable.

Next, the collocation search outputs a positional collocation frequency table, which shows what words collocate with the target word and how often they occur in particular positions in the sentence (see Fig. 11). By choosing a statistic (chi-squared score or *t*-score), analysts can find different types of "significant" collocations.

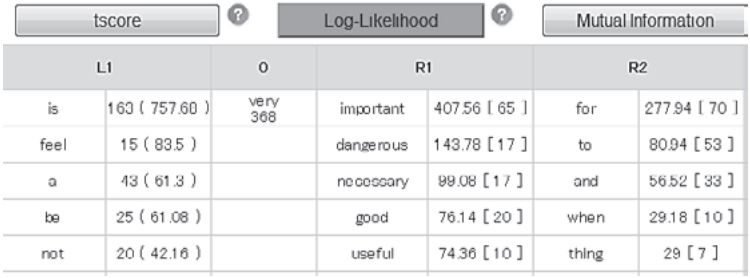


Fig. 11 Collocation search results (positional collocation table of "very" used by Chinese learners: Sorted based on the log-likelihood value)

Also, by clicking the [?] mark attached to each statistic, users can easily read the

guide about each statistical calculation, shown in Fig. 12.

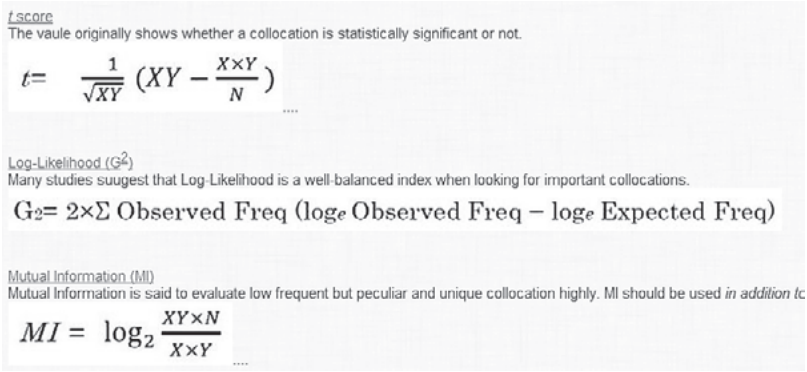


Fig. 12 Online guide for statistic adopted in the ICNALE *Online*

The wordlist search outputs the frequency lists of word forms or lemmas (see Fig. 13). In the word form list, forms such as "is," "am," and "are" are treated as different words, while in the lemma list, all of them are treated as "be." As trivial punctuation is deleted in the online version, the total number of tokens or types might be somewhat different from those of the download version.

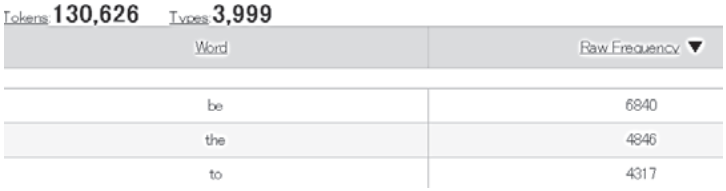


Fig. 13 Wordlist search (lemmas most frequently used by Korean learners)

Finally, the keyword search makes it possible to compare two different texts, for example, essays by native speakers and those by Japanese learners, and to specify "keywords" that occur statistically more in the target text than in the reference text. A keyword search is a powerful analytical technique to investigate the keywords overused or underused by a particular learner group when compared to native speakers or a different learner group. Fig. 14 shows an example of how keyword search results are presented.

Word	Statistic
the	185.62
smoker	130
thus	93.66
his	87.58

Fig. 14 Keyword search results (Overused words for Singaporean learners in comparison to English native speakers)

By utilizing one of these functions or some in combination, users, even if they are quite new to corpus queries, can conduct a standard analysis of learners' interlanguage and obtain the appropriate outputs.

III Sophistication of CIA with the ICNALE: A Case Study

3.1 New Perspectives for CIA

As summarized in Section 1.2, CIA, which is conducted between native speakers and learners or between learner groups with different L1s, helps us to deepen our understanding of learners' interlanguage. The difference between writer groups is shown most characteristically in the words overused by one or each group. Thus, many studies in the field have attempted to identify overused words as a kind of keyword characterizing a certain writer group.

However, the identification of keywords for a particular writer group is not as easy as it seems, and we need to be careful at least in three ways.

First, we should discriminate the essential and meaningful gap between writer groups from the technical gap in the writing conditions. Based on an analysis of learner data, Altenberg (1997) reported that Swedish learners overuse a more involved, namely, spoken-oriented style in essay writing, a finding that was also supported by Petch-Tyson (1998). However, a careful reexamination of the data proved that "this is primarily due to task setting (time available) and intertextuality (access to secondary sources)" not to the difference between native speakers and particular learner groups, and Altenberg's (1997) finding actually represents the fact that "learners exhibit more involvement in timed than in untimed essays, but less if they have access to other texts" (Ädel, 2008).

Second, we should consider the possibility of internal variety within a writer group. If we say something is a keyword that characterizes a particular writer group, it is expected to apply to all or at least most of the learners belonging to that writer group. This illuminates the importance of comparing learners at different proficiency levels in the target group. Only the features applicable to a wide range of learners in the group,

not those applicable to a limited part of them, are qualified to be a keyword for that writer group.

Third, we need to examine the possibility of external or international universality in learners' interlanguage. Many studies have presented various lists of overused words as something characterizing different writer groups, but there is always the possibility that these words are also characteristic of other writer groups. It is highly dangerous to attribute the difference obtained in a comparison between a particular learner group and native speakers directly to that learner group without investigating other learner groups.

As outlined previously, the ICNALE is one of the few learner corpora enabling us to appropriately control all of these three problematic factors and to conduct a more sophisticated CIA.

3.2 Research Design

Here, we aim to exemplify how conventional keyword extraction procedures can be further sophisticated by using the ICNALE. Our research aim is to identify "true keywords" that characterize Japanese learners of English. We will define true keywords as a set of words that are not influenced by writing conditions and are overused by Japanese learners at all proficiency levels in comparison to native speakers and at the same time not overused by any other groups of writers in Asia.

The data used for the study is the ICNALE Version 2.1 released in February 2013. The concordancer used for corpus analysis is AntConc 3.2.5w. The degree of overuse is measured by a log likelihood ratio (LL), and the words whose LL values are higher than 10 are regarded as significantly overused words. Although the ICNALE holds two kinds of datasets of ENSs, we will use only the ENS1 module, which collects essays written by college students in the Inner Circle. Topic-dependent words such as "part-time," "job," "smoking," and "restaurants" are manually excluded from the analysis, even if their LL values are sufficiently high.

Our research questions (RQs) are (1) Which words are overused by Japanese learners in general in comparison to English native speakers?, (2) How do Japanese learners at different L2 proficiency levels use them?, and (3) How do different Asian learners use them? By examining these RQs, we aim to identify true keywords for Japanese learners.

Our analysis focuses on the top 10 most significantly overused words. Concerning RQ3, only learners at the B1_2 level will be examined, which eliminates the influence of the gap in proficiency. In order to investigate the relationships among different learner groups as well as those between learner groups and overused lexical items, we will conduct a correspondence analysis, which is one of the data visualization methods for cross-tabular data that has been increasingly adopted in recent corpus studies (Ishikawa et al., 2010).

3.3 Findings and Discussion

3.3.1 RQ1 Words Overused by Japanese Learners

A comparison between the entire dataset of Japanese learners and that of English native speakers revealed as many as 162 overused words whose LL values are higher than 10. Table 4 lists the top 30 among them.

Table 4 Overused words

Word	Freq	LL	Word	Freq	LL	Word	Freq	LL
we	2416	410.3	restaurant	921	85.5	think	1811	50.9
smoke	1883	175.3	but	1380	83.3	reasons	411	49.6
completely	547	147.4	must	336	78.1	example	364	49.4
money	1354	133.7	seat	165	74.0	job	2100	48.7
smoking	3393	102.8	n't	1304	65.1	important	853	47.1
agree	621	102.7	so	1582	58.5	useful	102	45.8
society	381	98.4	earn	243	56.9	course	189	43.5
people	1981	94.9	dishes	126	56.5	get	581	43.5
seats	224	90.9	eating	211	55.2	smell	310	43.5
smoker	483	89.5	reason	321	51.2	passive	151	43.4

As expected, it was shown that more than a few of them are directly or indirectly dependent on the topics of part-time jobs for college students and not smoking at restaurants. By manually checking each word in its original context, we finally selected the top 10 topic-independent overused words: "we," "agree," "people," "but," "must," "n't," "so," "reason," "think," and "example." ("Reasons" was excluded because its singular form "reason" was included, and "example" was added instead.)

A list of these distinctly overused words seems to suggest that Japanese learners have a characteristic tendency to overuse (i) indefinite personal nouns or pronouns ("we," "people"), (ii) thought-related verbs ("agree," "think"), (iii) words concerning paragraph structuring ("but," "so," "reason[s]," "first," "second" [NB: "first" and "second" are used in collocation with "reason"]), (iv) a modal verb of obligation ("must"), (v) contraction ("n't" >not), and (vi) a unit of the phraseology ([for] "example"). This is largely in accordance with the intuition of those engaged in English education in Japan.

3.3.2 RQ2 Use of the Top 10 Words Overused by Japanese Learners at L2 Proficiency Levels

Then, are these words really overused by Japanese learners at various proficiency levels? Table 5 shows the LL values of these words obtained by the comparison of Japanese learners at A2, B1_1, B1_2, and B2 levels to native speakers.

Table 5 LL Values of the 10 Words for Japanese Learners at Different Proficiency Levels

Words	A2	B1_1	B1_2	B2+
we	375.2	371.8	96.0	85.1
agree	91.0	80.7	49.1	27.7
people	57.6	89.2	48.2	27.8
but	68.8	77.7	24.7	8.4
must	71.9	72.4	22.4	11.0
n't	38.9	59.4	45.0	11.0
so	45.3	50.3	31.3	3.3
reason	57.0	29.4	37.9	2.2
think	61.3	33.1	10.1	6.6
example	31.5	49.7	27.2	7.2
Avg.	89.9	91.4	39.2	19.0

What should be noted here is that the B2+ learners do not significantly overuse lexes such as "but," "so," "reason," "think," and "example" anymore, and thus, the number of words whose LL scores are higher than the threshold in all the proficiency levels is limited to just five words: "we," "agree," "people," "must," and "n't," which can be candidates for true keywords for Japanese learners.

The quotations below are excerpts from essays by Japanese learners at the B1_2 level. Each contains many of the key overused words, which are shown in bold italics.

> I ***agree*** with the statement that it is important for college students to have a part-time job. Of course, I have a part-time job. A part-time job ***must*** be a valuable experience.... Some ***people*** who did not have a part-time job when they are college students will not know the difficulty until they get a job.... So, ***we*** should experience many things and get a correct view, the right way to think and so on....
>
> (JPN_016_PTJ)

> I ***agree*** that smoking should be completely banned at all the restaurants in the country. There are two reasons, health problem and manner. First, smoking makes health problem. Smoking is***n't*** good not only for a smoker's health but also for other ***people***'s health... However, if a person who does***n't*** do smoking becomes in such situation, he will want to claim about it. Of course, ***we*** ca***n't*** know all such disease is caused by smoking ***people***... Smoking makes a lot of ***people*** feel bad, so not to smoke in public space is good manner. But, do ***we*** have to ban smoking? ...
>
> (JPN_010_SMK)

Another fact of note is that the overall degree of overuse, as shown in the average LL values in the right column of Table 5, clearly decreases in proportion to the increase in

L2 proficiency.

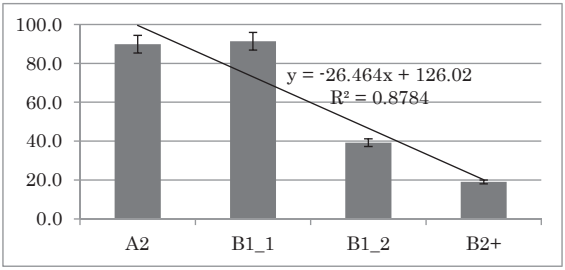


Fig. 15 Decrease of the average LL values according to increase in L2 proficiency

Although the difference between A2 and B1_1 is not distinct, the average LL value consistently decreases from B1_1 to B2+. A high *R* square value of the regression model shown in the figure suggests that a consistent trend of change is observed among the four proficiency levels. An important finding is that a striking overuse of a particular set of lexes, even though it may characterize novice Japanese learners, is not necessarily characteristic of Japanese learners in general.

3.3.3 RQ3 Use of the Top 10 Words Overused by Asian Learners

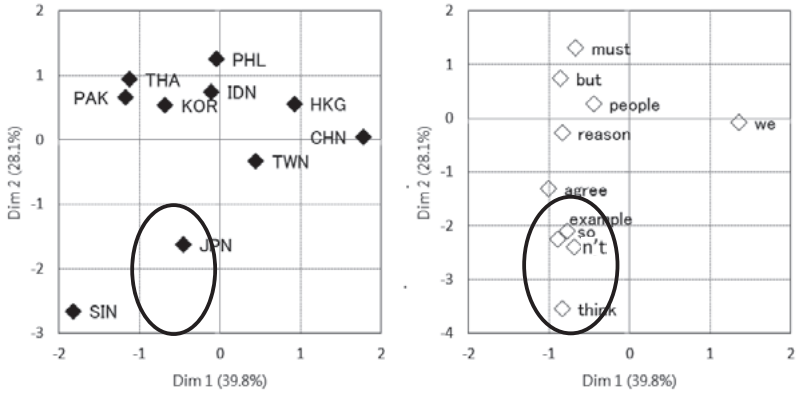
Finally, we will examine how Asian learners of English in the outer and expanding circles, who seem to have a lot in common with Japanese learners, use the 10 words identified in Section 3.3.1. Table 6 shows the LL values of the 10 words obtained in the comparisons of the learners in 10 countries and areas of Asia to native speakers.

Table 6 LL Values of the 10 Words for Asian EFL Learners at the B1_2 Level

Word	Expanding Circle (EFL)						Outer Circle (ESL)				Avg.
	JPN	CHN	IDN	KOR	THA	TWN	HKG	PAK	PHL	SIN	
we	96.0	281.4	82.1	29.7	8.4	54.1	7.7	2.4	39.5	0.0	60.1
agree	49.1	0.0	14.4	16.4	19.8	15.9	1.1	0.0	0.0	11.6	12.8
people	48.2	31.8	36.3	30.0	25.1	21.8	0.0	44.4	15.4	0.0	25.3
but	24.7	14.2	9.1	39.0	17.4	6.6	0.0	44.8	23.2	0.0	17.9
must	22.4	16.6	89.9	33.2	82.8	6.4	4.1	8.0	39.8	0.0	30.3
n't	45.0	3.4	0.0	0.6	8.0	3.2	0.0	0.0	0.0	0.0	6.0
so	31.3	2.3	0.1	0.7	0.0	0.0	0.0	8.8	0.0	0.0	4.3
reason	37.9	0.2	29.1	17.1	14.0	7.6	0.1	5.7	3.6	1.1	11.6
think	10.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
example	27.2	0.0	5.4	3.7	1.3	1.3	0.001	0.0	0.0	0.0	3.9
Avg.	39.2	50.0	33.3	18.9	22.1	14.6	2.6	19.0	24.3	6.3	23.0

The analysis revealed that some of the words overused by Japanese EFL learners are also overused by various Asian learners, while many others are overused only by Japanese learners. The words whose LL values are higher than 10 only in JPN are the four words "n't," "so," "think," and "example."

The scatter plots in Figs. 16–17, generated by correspondence analysis, illustrate the relations among different writer groups and between writer groups and the overused words.



Figs. 16–17 Scatter Plots of Item 1 (left, Fig. 16) and Item 2 (right, Fig. 17)

The figures above also support the finding that JPN is characterized most distinctively by a lexical cluster of "example," "so," "n't," and "think" in the third quadrant.

The quotation below is an excerpt from an essay including all of these Japanese-specific overused words, which are shown in bold italics.

> I agree with this idea. Today in many public places, smoking is banned. For ***example***, in a station smoking places are separated by box. Most people are beginning to pay attention to smoking. ***So***, smoking in a restaurant is not permitted. There are three reasons for this. First, smoking is very harmful not only for themselves but also for other people around them. It is true that people have freedom of smoking, but I ***think*** they should refrain it in public places... Second, people in the restaurant ca***n't*** enjoy their meal by smoking...
>
> (JPN_048_SMK)

Considering the results of the previous analysis, that the words overused by Japanese EFL learners are limited to the five words "we," "agree," "people," "must," and "n't" (see Fig. 18), we can safely conclude that true keywords for Japanese learners, namely, the

words not influenced by writing conditions, overused by Japanese learners at all the proficiency levels, and not overused by any other groups of writers in Asia, are limited astonishingly to only one word, "n't," the contracted form of "not."

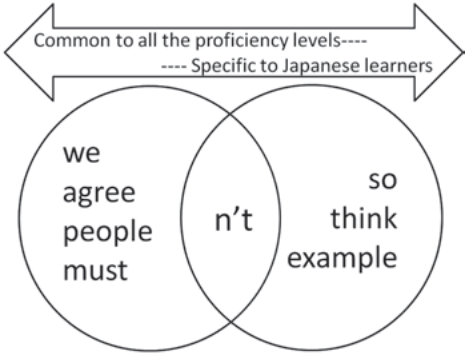
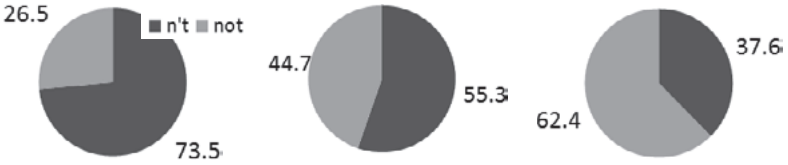


Fig. 18 Filtering the overused words

Although a learner corpus, if it is carefully designed, enables us to identify true keywords for a particular writer group, it is not necessarily easy to explain why the learners overuse that specific vocabulary. One's interlanguage seems to be influenced by a composite of various linguistic, personal, educational, social, and cultural factors.

In this case, however, the fact that only Japanese learners, even those at a relatively advanced level, overuse "n't" in written essays, may be partly due to the editing principle of English textbooks used in Japan. Figures below show the percentages of "n't" and "not" appearing in the 12 junior high school textbooks (Fig. 19), the four senior high school textbooks (Fig. 20), and two sample corpora (FROWN/FLOB) which collect two million words of American and British written English in 1990s (Fig. 21).



Figs. 19–21 Percentages of "n't" and "not" in junior high school textbooks (left, Fig. 19), senior high school textbooks (middle, Fig. 20), and FROWN/FLOB (right, Fig. 21)

As the focus of English education is becoming increasingly communication-oriented, colloquial vocabulary and usages, including frequent use of contractions, have become more popular in textbooks than they were previously, which might influence the

interlanguage of Japanese learners. What matters pedagogically is that Japanese learners' overuse of "I think," "so," and "but" are often discussed and largely well known to those engaged in English teaching in the country, while the overuse of contractions seems not to have received appropriate attention to date. In this sense, our finding is of some pedagogical note.

IV Conclusion

In this paper, the author discussed aspects of learner corpus studies with special attention given to CIA and introduced the key features of the ICNALE.

Also, through a case study aimed at identifying true keywords for Japanese learners of English, we showed how the ICNALE can contribute to the sophistication of conventional CIA. Our success in discriminating true keywords from pseudo-keywords suggests the advantage of using a controlled international learner corpus for discussions on learners' interlanguage.

CIA, if it is conducted with an appropriate database and a reliable procedure, can shed new light on our understanding of learners' L2 use in diverse social and cultural contexts. It might also help us explore new directions in designing a more effective L2 teaching curriculum optimized for different writer groups, as Hasselgård & Johansson (2011) conclude that revealing "features of learner language, or interlanguage, ... can potentially lead to improved language teaching as well as insights into the process of language learning."

Although the ICNALE was only recently released, the number of studies using its data has been gradually increasing, and aspects of the interlanguage of Japanese and other Asian learners have been examined from a variety of perspectives: modal verbs (Chen, 2013), prepositions (Matsushita, 2012), *-ly* adverbs (Ishikawa, 2010a), linking adverbials (Ishikawa, 2011a), speech-act verbs (Inoue, 2011), use of different parts of speech (Inoue, 2012), NS/NNS gaps (Ishikawa, 2010b), phraseology use (2011b), learner corpus-based dictionary making (Ishikawa, 2011c), the effect of rewrite (Ishikawa, 2012a), proficiency marker identification (Ishikawa, 2012c), and learners' interlanguage as World Englishes (Ishikawa, 2012d). Also, the ICNALE was used for compiling a new corpus-based Japanese-English dictionary (Kishino, 2013) that offers detailed information about Japanese learners' typical over/underuse of words.

However, it is doubtless that there remain many things to be done for further improvement of the ICNALE. We have already set about several new projects to collect (i) essays written in learners' L1, (ii) essays corrected by professional proofreaders, (iii) essays corrected by learners themselves (Ishikawa, 2012a), and (iv) speech about the same topics produced by learners. By incorporating these new data modules into the current essay modules, we aim to make the ICNALE a more valuable and reliable database for sophisticated CIA.

Acknowledgements

The ICNALE project is supported by Grants-in-Aid for Scientific Research by Japan Society of the Promotion of Science (No. 22320104).

The author would like to express gratitude to the ICNALE project advisers, Masao Aikawa, Ichiro Akano, Tetsuya Enokizono, Kazuaki Goto, Hideo Masuda, Masamichi Mochizuki, Yasumi Murata, and Hiroshi Shimatani.

The author also appreciates the assistance in data collection by Katsuki Mayumi, Fang Li, and Lu Yuanwen (China); Leonardi Lucky Kurniawan (Indonesia); Yuka Ishikawa (Japan); John Milton (Hong Kong); Sook Kyung Jung, Oryang Kwon, and Masahiro Hori (Korea); Asim Mahmood (Pakistan); Karen L. Gabinete (The Philippines); Vincent Ooi (Singapore); Siaw-Fong Chung (Taiwan), Sonthida Keyuravong and Punjaporn Pojanapunya (Thailand).

References

Ädel, A. (2008). Involvement features in writing: Do time and interaction trump register awareness? In G. Gilquin, S. Papp, & M. Díez-Bedmar (Eds.). *Linking up contrastive and learner corpus research* (pp. 35-53). Amsterdam, The Netherlands: Rodopi.

Altenberg, B. (1997). Exploring the Swedish component of the international corpus of learner English. In B. Lewandowska-Tomaszcyk & P. J. Melia (Eds.). *Proceedings of PALC'97: Practical applications in language corpora* (pp. 19-132). Lódz, Poland: Lódz University Press.

Borin, L., & Prütz, K. (2004). New wine in old skins? A corpus investigation of L1 syntactic transfer in learner language. In G. Aston, S. Bernardini, & D. Stewart (Eds.). *Corpora and language learners* (pp. 67-87). Amsterdam, The Netherlands: John Benjamins.

Gilquin, G., Papp, S., & Díez-Bedmar, M. (2008). Introduction. In G. Gilquin, S. Papp, & M. Díez-Bedmar (Eds.). Linking up contrastive and learner corpus research (pp. vii-xi). Amsterdam, The Netherlands: Rodopi.

Granger, S. (1996). From CA to CIA and back: An integrated 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, Lund 4-5 March 1994* (pp. 37-51). Lund: Lund University Press.

Granger, S. (Ed.). (1998). *Learner English on computer*. London, UK: Longman.

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.). *Learner English on computer* (pp. 3-18). London, UK:

Longman.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In K. Aijmer (Ed.). Corpora and language teaching (pp. 13-32). Amsterdam, The Netherlands: John Benjamins.

Gilquin, G., Papp, S., & Díez-Bedmar, M. B. (Eds.). (2008). *Linking up contrastive and learner corpus research*. Amsterdam, The Netherlands: Rodopi.

Granger, S., Dagneaux, E., & Meunier, F. (2002). (Eds.) *International corpus of learner English: Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). *International corpus of learner English*. (Version 2). Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.

Granger, S., Hung, J., & Petch-Tyson, S. (Eds.). (2002). Computer learner corpora, second language acquisition and foreign Language teaching. Amsterdam, The Netherlands: Benjamins.

Hasselgård, H., & Johansson, S. (2011). Learner corpora and contrastive interlanguage analysis. In F. Meunier, S. de Cock, G. Gilquin, & M. Paquot (Eds.). A taste for corpora: In honour of Sylviane Granger (pp. 33-61). Amsterdam, The Netherlands: John Benjamins.

Inoue, S. (2011). How speech-act verbs should be described: A study based on NS and NNS corpus. In K. Akasu & S. Uchida (Eds.), *Asialex 2011 proceedings: Lexicography, theoretical and practical perspectives* (pp. 170-179). Tokyo: The Asian Association for Lexicography.

Inoue, S. (2012). Gakushusha no hinshi shiyo noryoku no kento NS/NNs kopasu ni motozuku keiryo bunseki. *The Institute of Statistical Mathematics Cooperative Research Report*, 277, 53-62. [How Learners Use English Parts of Speech: A Quantitative Analysis Based on NS and NNS Corpus].

Ishikawa, S. (2010a). Nihonjin eigo gakushusha no –ly fukushi shiyo: Gakushusha kopasu CEEAUS ni motozuku keiryoteki kosatsu. *Journal of the Chubu English Language Education Society*, *39*, 181-188. [Use of –ly adverbs by Japanese learners of English: A quantitative analysis of the learners' corpus, CEEAUS].

Ishikawa, S. (2010b). Nihonjin eigo gakushusha ni yoru chukan gengo no goi unyou. In H. Kishimoto (Ed.). Kotoba no taisho (pp. 217-231). Tokyo: Kuroshio Shuppan. [Lexical use in the interlanguage by Japanese learners of English]

Ishikawa, S. (2011a). A corpus-based study on Asian Learners' use of English linking adverbials. *Themes in Science and Technology Education*, *3*(1-2), 139-157.

Ishikawa, S. (2011b). Phraseology overused and underused by Japanese learners of English. In K. Yagi et al. (Eds.). *Phraseology, corpus linguistics and lexicography* (pp. 83-94). Nishinomiya, Japan: Kwansei Gakuin University Press.

Ishikawa, S. (2011c). Learner corpus and lexicography: "Help-boxes" in EFL dictionaries

for Asian learners: A study on the international corpus network of Asian learners of English. In K. Akasu & S. Uchida (Eds.), *Asialex 2011 proceedings: Lexicography, theoretical and practical perspectives* (pp. 190-199). Tokyo: The Asian Association for Lexicography.

Ishikawa, S. (2011d). 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, UK: University of Strathclyde Press.

Ishikawa, S. (2012a). Writing, rewriting, proof writing: Gakushusha kopasu ni motozuku shutei koka no keiryoteki kenkyu. *Journal of the Chubu English Language Education Society*, *41*, 245-256. [Writing, Rewriting, Proof Writing: Learner corpus-based study on the effect of revisions].

Ishikawa, S. (2012b). *Beshikku kopasu gengogaku*. Tokyo, Japan: Hitsuji Shobo. [Corpus linguistics: Basics].

Ishikawa, S. (2012c). L2 Learners' use of English words and phraseologies: Corpus-based identification of lexical proficiency markers. In J. Szerszunowicz, B. Nowowiejski, K. Yagi, & T. Kanzaki (Eds.) *Research on phraseology in Europe and Asia: Focal issues of phraseological studies*, *1*, 389-410. Białystok, Poland: University of Białystok Publishing House.

Ishikawa, S. (2012d). Ajia ken doshinen ni okeru eigo goi shiyo. In A. Inoue & T. Kanzaki (Eds.). *21 seiki eigo kenkyu no shoso* (pp. 450-464). Tokyo, Japan: Kaitakusha. [English vocabulary use in concentric circles in Asia.]

Ishikawa, S., Maeda, T., & Yamasaki, M. (2010). *Gengo kenkyu no tame no tokei nyumon*. Tokyo, Japan: Kuroshio Shuppan. [An introduction of statistics for language studies].

Ishikawa, S., & Ishikawa, Y. (forthcoming). How writers' personal attributes influence their L2 use: A study based on the ICNALE. Paper presented at Learner Corpora 2013: Compiling and using learner corpora to teach and assess productive and interactive skills in foreign languages at university level (May 16-17, 2013, University of Padova, Italy).

Izumi, E., Utimoto, K., Isahara, H. (2004). Nihonjin 1200 nin no eigo supikingu kopasu. Tokyo, Japan: Alc. [Corpus of English speeches by 12,000 Japanese people].

Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: The English language in the outer circle. In R. Quirk. H. Widdowson (Eds.), English in the world: Teaching and learning the language and literatures (pp. 11-30). Cambridge, UK: Cambridge University Press.

Kaminishi, T. (2012). Nihonjin eigo gakushusha to korokeshon shiyo: Tokei shuho ni yoru kajoshiyo oyobi kasho shiyo korokeshon no tokutei. *The Institute of Statistical Mathematics Cooperative Research Report*, 277, 87-100. [Japanese learners of English and collocation use: Statistics-based identification of collocations overused

and underused by them].

Kishino, E. (Ed.). (2010). *Wisdom Japanese-English dictionary*. 3rd Ed. Tokyo, Japan: Sanseido.

Leech, G. (1998). Preface. In S. Granger (Ed.). *Learner English on computer* (pp. xiv-xx). London, UK: Longman.

Meara, P., & Milton, J. (2003). *X_Lex, the Swansea levels test*. Newbury, UK: Express Publishing.

McEnery, T., & Hardie, A. (2012). Corpus linguistics. Cambridge, UK: Cambridge University Press.

Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. In I. Bartning, M. Martin, & I. Vedder (Eds.). *Second language acquisition and testing in Europe* (pp. 211-232). Online: Eurosla.

Nation, P., & Beglar, D. (2007). A vocabulary size test. The Language Teacher, 31(7), 9-13.

Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In J. McH. Sinclair (Ed.). How to use corpora in language teaching (pp.125-152). Amsterdam, The Netherlands: John Benjamins.

Pennington, M. C. (2003). The impact of the computer in second language writing. In B. Kroll (Ed.). Exploring the dynamics of second language writing (pp. 287-310). Cambridge, UK: Cambridge University Press.

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

Sakaue, T., Sugiura, M., & Narita, M. (2008). Gakushusha kopasu NICE no kochiku. In M. Sugiura (Ed.). *Eigo gakushusha no korokeshon chishiki ni kansuru kisoteki kenkyu* (pp. 1-14). Nagoya, Japan: Nagoya University. [Compilation of a learner corpus, NICE].

Sugiura, M. (Ed.). (2007). *Eigo gakushusha no kolokeshon chisiki ni kansuru kisoteki kenkyu*. Nagoya, Japan: Nagoya University. [A fundamental study on English L2 learners' collocation knowledge].

Sugiura, M., Narita, M., Ishida, T., Sakaue, T., Murao, R., & Muraki, K. (2007). A discriminant analysis of non-native speakers and native speakers of English---NICE: Learner corpus 2.0 to come. In M. Davies, P. Rayson, S. Hunston, P. Danielsson (Eds.). *Proceedings of the Corpus Linguistics Conference, CL2007: University of Birmingham, UK, 27-30 July 2007* (pp. 1-17 of Article #216). Birmingham, UK: University of Birmingham.

Tono, Y. (2004). Multiple comparison of IL, L1 and TL corpora: The case of L2 acquisition of verb subcategorization patterns by Japanese learners of English. In G. Aston, S. Bernardini, & D. Stewart (Eds.). *Corpora and language learners* (pp. 45-66). Amsterdam, The Netherlands: John Benjamins.

Tono, Y. (Ed.). (2007). *Nihonjin chukosei ichimannin no eigo kopasu*. Tokyo,

Japan: Shogakukan. [Corpus of English essays by 10,000 Japanese junior and senior high school students].

Tono, Y. (2009). Integrating learner corpus analysis into a probabilistic model of second language acquisition. In P. Baker (Ed.). *Contemporary corpus linguistics* (pp. 184-203). London, UK: Continuum.