「特別講演」

The ICNALE Edited Essays: A Dataset for Analysis of L2 English Learner Essays Based on a New Integrative Viewpoint

Shin'ichiro ISHIKAWA

1. An Outline of the ICNALE Project

The International Corpus Network of Asian Learners of English (ICNALE) is a large-scale collection of English speeches and essays produced by college students (including some graduate students) from ten countries and regions in Asia and native English speakers. The ICNALE currently consists of four data modules: Spoken Monologue (Ishikawa, 2014), Spoken Dialogue (Ishikawa, 2018), Written Essays (Ishikawa, 2013), and Edited Essays.

Table 1: Structure of the ICNALE

Module	Released	Samples	Tokens	Contents
Spoken Monologue	2016	4,400	c500,000	one-minute monologues recorded on the answering phone
Spoken Dialogue	2020	---	---	approximately 30-to-40-minute utterances in OPI-like interviews
Written Essays	2013	5,600	c1,300,000	200-to-300-word essays
Edited Essays	2018	640	c150,000	learners' original essays and their edited versions

The ICNALE has seven key principles: (1) a focus on Asia, (2) consideration of linguistic modes, (3) condition control, (4) proficiency control, (5) learner background survey, (6) native-speakers' reference data collection, and (7) open distribution.

First, the ICNALE focuses exclusively on Asian learners. Paying attention to the diversity of English learners/users in the region, it collects data in both EFL areas (China, Indonesia, Japan, Korea, Taiwan, and Thailand) and ESL areas (Hong Kong, Pakistan, the Philippines, and Singapore).

Second, the ICNALE collects varied modes of learner English: spoken and written, and also monologue and dialogue.

Third, the ICNALE controls the conditions for speaking and writing as rigidly as possible. The number of topics (prompts) has been restricted to two ("It is important for college students to have a part-time job" and "Smoking should be completely banned at all the restaurants in the country"). Participants are required to speak or write about what they think of the topics and why they think so. An essay is required to be between 200 words and 300 words in length, and the time allotted for a monologue speech is 60 seconds. As a rule, samples that have not met these criteria are excluded.

Fourth, the ICNALE collects L2 proficiency data from all the participants; in fact, they are required to report their scores in English proficiency tests such as TOEFL, TOEIC, and IELTS and also take a receptive vocabulary size test (Nation & Beglar, 2007). Thus, based on their scores in English proficiency tests or vocabulary size tests, all participants are classified into four proficiency bands linked to the Common European Framework of Reference for Languages (CEFR) scale: A2, B1−1 (B1 lower), B1−2 (B1 upper), and B2+.

Fifth, the ICNALE collects a wide range of background information from the participants, including sex, age, a period of time spent learning English, any experiences of staying in English speaking countries, motivations to learn English, language skills they like to focus on, experiences of using L2 at schools, and so on.

Sixth, the ICNALE also includes native-speakers' L1 English production data. They are given the same prompts and required to speak or write in the same situations. Considering possible diversity within the category of "native speakers" (Leech, 1998), the ICNALE collects data from three groups of native speakers: college students, English teachers, and other adults.

Finally, the ICNALE data is made available to researchers around the world. Users can download the whole ICNALE dataset and conduct their own research. In addition, they can access ICNALE resources through the online query system, which is called "The ICNALE *Online*."

The ICNALE project was launched in 2007. After ten years of continuing efforts by a group of international researchers, it has become one of the largest learner corpora ever built; it is now utilized by researchers, teachers, and students around the world.

2. An Outline of the ICNALE Edited Essays

2.1 Aim

The ICNALE Written Essays, which includes 5,600 essay samples, has been widely used since its release. However, the included essays are neither error-tagged nor rated, and a corpus user cannot discuss what type of errors tend to occur in learner essays, how the errors should be corrected, and what degree of quality the essays possess. In order to make this kind of deeper analysis of learner essays possible, we have released a new ICNALE module, the ICNALE Edited Essays. This module aims to become a reliable dataset that will enable the analysis of L2 English learner essays based on a new integrative viewpoint.

2.2 Contents

The ICNALE Edited Essays includes learners' original essays, their fully edited versions, and rubric-based evaluation scores; these features enable users to analyze the quality of the learner essays using an integrative viewpoint.

The original essays were chosen at random from the ICNALE Written Essays. Excluding several cases where the number of original essays was not sufficient, we took 20 essays written by learners at each of the four different proficiency levels (A2, B1−1, B1−2, and B2+).

Table 2: Number of samples in the ICNALE Edited Essays

		A2	B1−1	B1−2	B2+	Total
EFL	China	20	20	20	20	80
	Indonesia	20	20	20	NA	60
	Japan	20	20	20	20	80
	Korea	20	20	20	20	80
	Taiwan	20	20	20	20	80
	Thailand	20	20	20	NA	60
ESL	Hong Kong	NA	20	20	20	60
	Pakistan	NA	20	20	NA	40
	Philippines	NA	20	20	20	60
	Singapore	NA	NA	20	20	40
Total		120	180	200	140	640

2.3 Collection of essay evaluation data

Although there are varied approaches to essay evaluation, many of them involve using some kind of rating rubric, which helps raters to rate learner essays in a consistent and reliable manner. One of the most widely used rubrics in the field of TESOL is the ESL Composition Profile (Jacobs *et al.*, 1981), which uses five rating criteria: Content (CON), Organization (ORG), Vocabulary (VOC), Language use (LNU) , and Mechanics (MEC).

Therefore, we recruited five professional editors, all of whom are native English speakers with strong academic backgrounds and ample experience in editing academic papers for publication in major journals, and asked them to rate learner essays with reference to the ESL Composition Profile.

Table 3: Profiles of editors who participated in the ICNALE Edited Essays Project

Editors	Age	Sex	Degree	Years	L1 English
Editor A	28	Female	BA	3	Canadian
Editor B	32	Female	MS	5	Australian
Editor C	27	Female	BS	3	American
Editor D	38	Female	BS	10	British
Editor E	31	Female	PhD	2	Australian

Table 4: Scoring guide for the category of content

Score	Descriptors
10～12	EXCELLENT TO VERY GOOD: knowledgeable • substantive • thorough development of thesis • relevant to assigned topic
7～9	GOOD TO AVERAGE: some knowledge of the subject • adequate range • limited development of thesis • mostly relevant to topic but lacks detail
4～6	FAIR TO POOR: limited knowledge of the subject • little substance • inadequate development of topic
1～3	VERY POOR: does not show knowledge of the subject • non-substantive • not pertinent • OR not enough to evaluate

In the original rubric, different scores were assigned to the five categories. However, we asked editors to score all the categories using 1–12 points so that they could rate more easily and consistently. Next, we calculated two kinds of total scores: a simple sum and a sum reflecting the weights suggested in the original rubric.

2.4 Collection of editing data

After rating the essays, the five editors were asked to edit learner essays on MS Word in track change mode so that clarity of the essays is improved and they become fully intelligible. They were required to retain the original texts as much as possible. Making any additions, deletions, and changes in the original contents was prohibited.

The figures below show how a learner's original essay was edited by an editor.

Now many parents and teachers disagree that college students should have their own part-time jobs b~~.~~ ~~Because they are~~ believe that students will spend their study~~-~~ time on working. But I do not think so~~.~~. I think a part-time job is one of the most important things in college life. It will be the precious experience in your whole life for several reasons. Firstly~~,~~, a ~~P~~part-time job is a good way to fill

Fig. 1: Sample of the edited essays

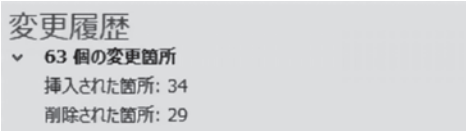


Fig. 2: Summary of changes added by an editor

As all the changes have been tracked in MS Word, corpus users can easily count how many words are added and deleted by an editor. In the sample above, 34 words were added, and 29 words were deleted. The number of edits, which is often called "an edit distance," is 63. Typically, the number of edits tends to decrease for good essays and increase for problematic essays.

2.5 Inter-editor variances

When a group of editors is in charge of scoring and editing, we need to be careful about possible inter-editor variances. Therefore, prior to the collection of evaluation and editing data, we conducted a small-scale calibration study in which we asked all five editors to rate and edit the same set of eight essays. These essays were written by learners who were at different proficiency levels and had different L1 backgrounds. Then, we confirmed the extent to which the average rating scores and the average numbers of edits could be maintained at stable levels across five editors.

Table 5: Average scores for eight samples

	Rating Scores (/12)						Number of Edits
	CON	ORG	VOC	LNU	MEC	Average	
Editor A	7.88	7.88	7.25	7.25	8.38	7.73	59.63
Editor B	9.88	10.38	8.38	7.63	8.25	8.90	49.50
Editor C	9.13	8.88	8.25	8.00	8.50	8.55	41.88
Editor D	8.38	8.13	7.88	7.25	8.00	7.93	48.50
Editor E	7.50	8.00	8.25	7.75	8.38	7.98	40.00
Average	8.55	8.65	8.00	7.58	8.30	8.22	47.90
Dif	2.38	2.50	1.13	0.75	0.50	1.18	19.63
Dif / Average (%)	27.78	28.90	14.06	9.90	6.02	14.30	40.97

The averages of the five category scores ranged between 7.73 (Editor A) and 8.90 (Editor B), and the gap was 1.18, which is equivalent to 14.30% of their average (8.22) and 9.79% of the full score (12). We could say that the variance is generally small in terms of essay evaluation. Meanwhile, the averages of the numbers of edits (*i.e.,* additions + deletions) ranged between 40.00 (Editor E) and 59.63 (Editor A), and the gap was 19.63, which amounts to 40.97% of their average (47.90). The variance was larger in terms of essay editing, though excluding Editor A, who tended to underrate learner essays and make more edits to them, the gap shrank to 21.13%.

This suggests that the rubric-based rating was more stable than editing, which required editors to define "intelligibility" by their own standards. The ICNALE Edited Essays includes detailed data of this calibration study and thus enables users to interpret the results of the data analysis in a more cautious and in-depth way.

2.6 Online query system

Users can download the entire dataset of the ICNALE Edited Essays and analyze it using any concordancer of their choice. In addition, they can access the data through the ICNALE *Online* interface, which currently offers two kinds of searches: KWIC Search and Keyword Search.

2.6.1 KWIC Search

KWIC Search enables users to retrieve the concordance lines including the target word(s). The figures below illustrate how users can examine the use of the term "go," which occurs in the original essays written by Chinese learners.

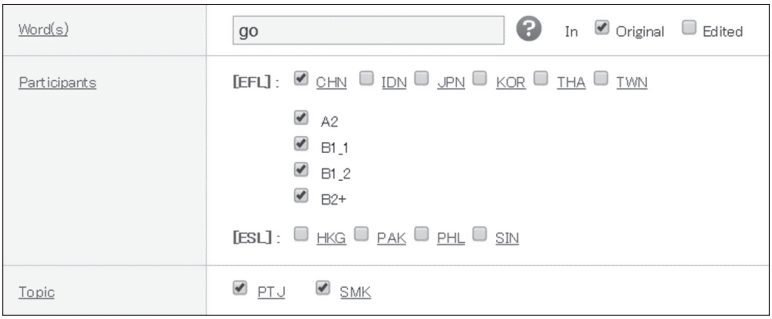


Fig. 3: Settings for the KWIC Search

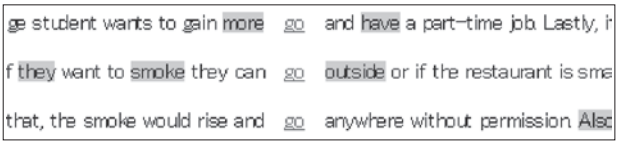


Fig. 4: Results of the KWIC Search

By entering a word or an expression that they would like to analyze, and choosing the version of the essays (original or edited), learners' countries/areas, their L2 proficiency levels, and the topic of the essays ("a part-time job" or "non-smoking"), users can obtain a list of concordance lines including the target word(s).

In the KWIC concordance, some words appear shaded; this shows that the words or phrases including them are changed in the edited essays. Fig. 4 shows that words such as "more," "have," "they," "smoke," "outside," and "Also" or expressions including them are to be altered in the edited versions.

When any word in the KWIC concordance is clicked, a new window pops up, and users are able to examine an original essay and its edited version at the same time.

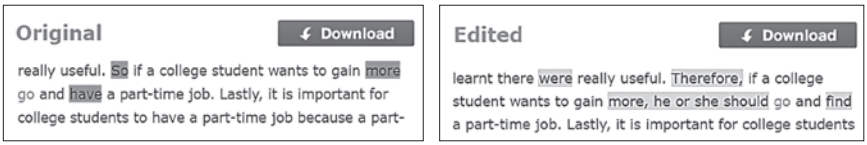


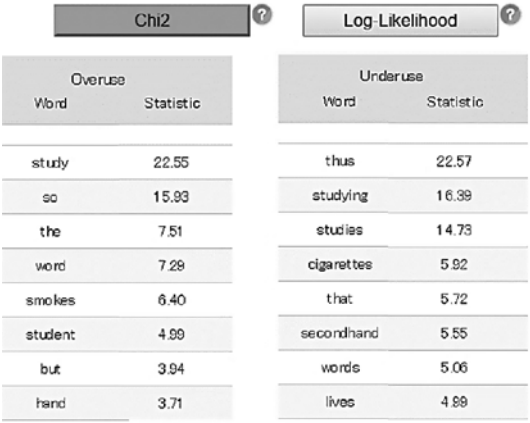
Fig: 5 Side-by-side comparison of the original/edited essays

Thus, users can easily analyze the kinds of problems that exist in learners' original essays and the correction of these problems by professional editors.

2.6.2 Keyword Search

Keyword Search enables users to identify words that occur at a statistically higher rate in one of the two texts to be compared.

Fig. 6: Settings for the Keyword Search



Overuse		Underuse	
Word	Statistic	Word	Statistic
study	22.55	thus	22.57
so	15.93	studying	16.39
the	7.51	studies	14.73
word	7.29	cigarettes	5.92
smokes	6.40	that	5.72
student	4.99	secondhand	5.55
but	3.94	words	5.06
hand	3.71	lives	4.89

Fig. 7: Results of the Keyword Search

By choosing the target (original essays) and reference (edited essays) datasets, which need to be for the same learner group, users can obtain a list of words that were overused in the original essays (namely, words to be deleted in the edited essays) and words that were underused in the original essays (namely, words added in the edited essays). Users can choose Chi-square value or Log-likelihood value as a statistical method for the keyness calculation.

Fig. 7 suggests that learners often use "so" in an erroneous way and that it should be replaced by "thus," for example. Thus, users can easily see the kinds of errors and inappropriate vocabulary that existed in learners' original essays and how they were corrected.

3. A Case Study: Multivariate Analysis of Chinese Learners' Essays

3.1 Aim and RQ

The ICNALE Edited Essays makes it possible to discuss the quality of learner essays based on a new integrative viewpoint. Using a part of its data, Ishikawa (In press) discussed the quality of Japanese learners' essays, paying attention to the rubric-based rating scores, the number of edits made by editors, and the average keyness values showing the degree of inappropriate vocabulary use by learners in comparison to native English speakers.

In the current study, then, we will discuss the features of Chinese learners' essays by considering essay-related parameters and writer-related parameters in an integrative way. Our research questions are as follows:

RQ1 To what extent are essay evaluation scores and the number of edits correlated?

RQ2 How are essay-related and writer-related parameters clustered?

3.2 Data and method

We analyzed 80 essays written by Chinese learners, which had been included in the ICNALE Edited Essays. Considering RQ1, we paid attention to the strength of the correlation between the sum of the five category-based evaluation scores and the number of edits made by editors.

Next, considering RQ2, we paid attention to the wide variety of essay-related and writer-related parameters shown below:

Table 6: Parameters used for the analysis

Essay-related parameters
Essay topics (part-time job [PTJ] or non-smoking [SMK]), Essay length (number of words per essay [LEN] ranging between 200−300 words), Number of edits (inverse of the number of additions and deletions [EDT]), Essay evaluation scores (five category-based scores ranging between 1−12: Content [CON], Organization [ORG], Vocabulary use [VOC], Language Use [LNU], and Mechanics [MEC], as well as their simple sum [TTL])
Writer-related parameters
Sex (female [FEM] or male [MAL]), Age (ranging between 18−21 years old [AGE]), Major (humanities [HUM], social sciences [SCS], or science and technology [SCT]), vocabulary size (scores in the receptive vocabulary size test (Nation & Beglar, 2007), ranging between 0−50 [VST]), Strength of motivation (integrative motivation [INT] and instrumental motivation [INS] ranging between 0−6 and both types [MOT] ranging between 0−12), Amount of L2 use (at primary schools [PRM], secondary schools [SEC], and colleges [COL], as well as in classes [INC] and out of classes [OTC] ranging between 1−6), Former L2 teaching (frequency of instruction by native English speaking teachers [NST] and specific instructions focusing on pronunciations [PRN], presentations [PRS], and essay writing [ESW], the type of L2 skills that learners like to focus on: (listening [LNS], reading [RDS], speaking [SPS], and writing [WRS], ranging between 1−6).

Discussing the number of edits, we used the inverse number, with the assumption that good essays would require fewer edits. Writer-related data, except for the vocabulary size, was obtained from the questionnaire survey.

We conducted a hierarchical cluster analysis in order to observe the interrelation between the variables. The distance was defined as the square root of (2−2*r*), and the Ward method was adopted for calculation of distance.

3.3 Results

3.3.1 RQ1 Strength of correlation

The table below presents Pearson's *r* values between the essay evaluation scores and the number of edits.

It was found that the number of edits showed a middle-level correlation (0.391−0.492) with any of the five category-based evaluation scores as well as their total score.

Table 7: Correlations between the essay evaluation scores and the number of edits

	EDT	CON	ORG	VOC	LNU	MEC	TTL
EDT	1.000						
CON	0.391	1.000					
ORG	0.436	0.786	1.000				
VOC	0.470	0.655	0.684	1.000			
LNU	0.443	0.644	0.684	0.742	1.000		
MEC	0.448	0.602	0.615	0.550	0.635	1.000	
TTL	0.492	0.904	0.896	0.841	0.865	0.716	1.000

This result seems to corroborate our common belief that good essays receive fewer edits and that bad essays receive more edits. However, it is important to note that the correlation *r* was not as high as was generally expected, which suggests that some of the Chinese learners wrote good but problematic essays, and others wrote poor but not so problematic essays. These findings show that grammatical and lexical correctness may not always guarantee the quality of learner essays.

3.3.2 RQ2 Clustering of essay-related and writer-related variables

By conducting a cluster analysis, we obtained a tree diagram, which is provided below. By setting a cutting point of 2, 33 variables were classified into five clusters (A, B, C, D, and E).

First, we would like to pay attention to Cluster E, which demonstrates that the essay evaluation scores and the number of edits are closely related. It also suggests that the relationship between evaluation and editing becomes clearer in essays about a social topic (non-smoking) rather than in essays about a familiar personal topic (a part-time job). Thus, we may infer that the essay topic may influence the essay's evaluation and its editing.

Next, Cluster B shows that the male students, many of whom have majored in science and technology, tend to have more experiences of being taught pronunciation, presentation, and essay writing, and they usually know more words in the target language. The experience of learning how to write good essays and the wider vocabulary size are expected to contribute directly to the quality of essays. However, as suggested by the fact that Clusters B and E do not merge until the distance reaches 2.5, such a connection is not necessarily confirmed by the current data.

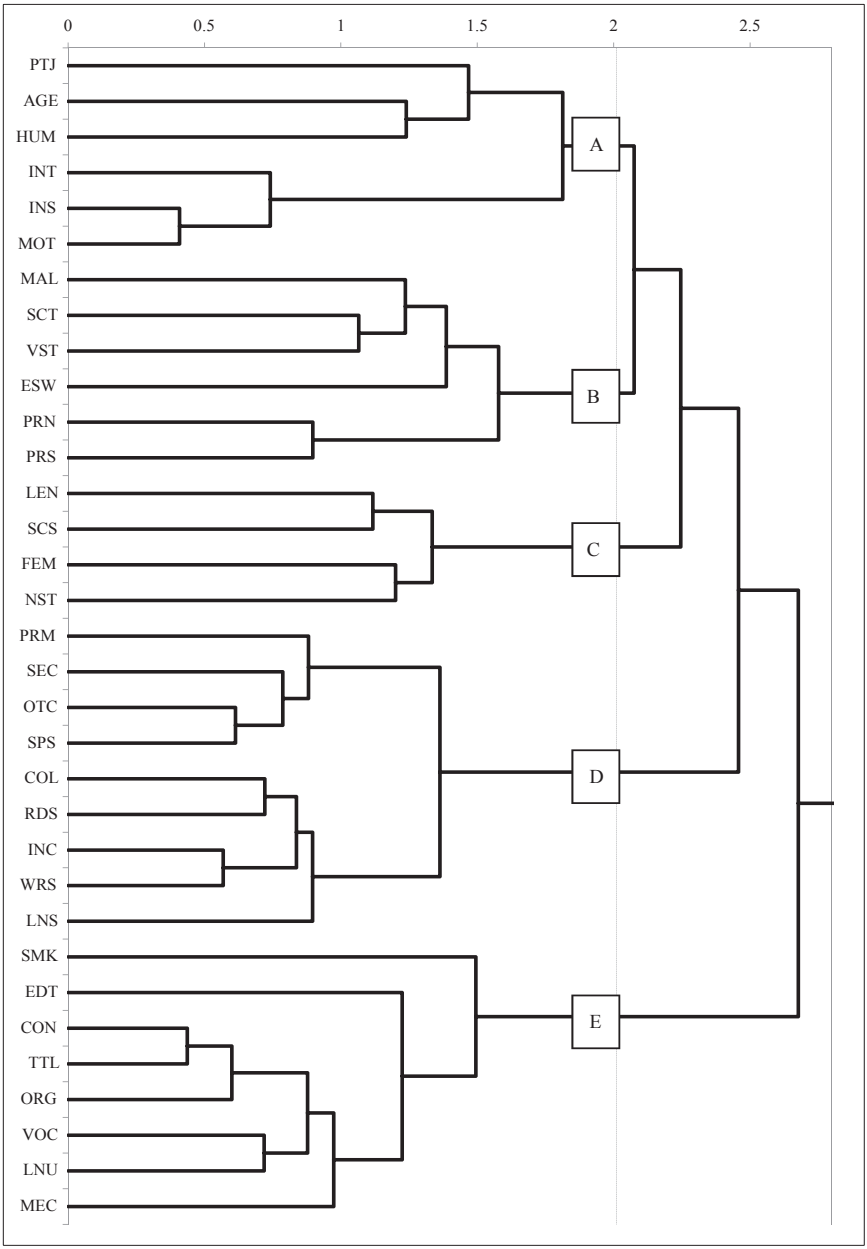


Fig. 8: A tree diagram based on the cluster analysis of the 33 variables

Table 8: The variables in each cluster

Cluster A	Cluster B	Cluster C	Cluster D	Cluster E
PTJ	MAL	LEN	PRM	SMK
AGE	SCT	SCS	SEC	EDT
HUM	VST	FEM	OTC	CON
INT	ESW	NST	SPS	TTL
INS	PRN		COL	ORG
MOT	PRS		RDS	VOC
			INC	LNU
			WRS	MEC
			LNS	

Other clusters also suggest interesting facts about the Chinese writers and their L2 essays. Cluster A shows that humanities students, including some English-major students, often have a higher L2 learning motivation and that they perform better when writing about a personal topic (a part-time job). Cluster C reveals that female students, many of whom majored in social sciences, tend to have more experiences of being taught by native English-speaking teachers, and they tend to write somewhat longer essays. Finally, Cluster D shows that learners' former experience of L2 use and the type of skills that they like to focus on are related to some extent. It is noteworthy that which skill learners have focused on in L2 learning mighr influence how much they have actually used L2 in various situations.

4. Summary

This paper introduced the outline of the ICNALE project and explained the aim, design, and contents of the ICNALE Edited Essays as its newest addition.

It then illustrated how the ICNALE Edited Essays could be used for the analysis of learner essays based on a new integrative viewpoint. Our case study, though quite preliminary, has proven that the essay evaluation scores and the number of edits show a middle-level correlation and that they can be influenced by the essay topic and also by a variety of essay-related and writer-related parameters. The author hopes that the ICNALE Edited Essays will contribute to spreading a new data-based analysis of the quality of learner essays, which is based on an integrative observation of texts and text producers.

Bibliography

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System*, *29* (3), 371−383.

Brooks, G. (2012). Assessment and academic writing: A look at the use of rubrics in the second language writing classroom. *Humanities Review* (Kwansei Gakuin University), *17*, 227−240.

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

Ishikawa, S. (2012). *Beshikku kopasu gengogaku*. Tokyo: Hitsuji Shobo. [A Basic Guide to Corpus Linguistics].

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world, 1* (pp. 91−118). Kobe, Japan: Kobe University.

Ishikawa, S. (2014). Design of the ICNALE-Spoken: A new database for multi-modal contrastive interlanguage analysis. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world, 2*（pp. 63−76). Kobe, Japan: Kobe University.

Ishikawa, S. (2018). The ICNALE Spoken Dialogue no sekkei: Taiwa ni okeru L2 koto sanshutsu kenkyu no tame ni. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world, 3*（pp. 3−21). Kobe, Japan: Kobe University. [Design of the ICNALE Spoken Dialogue: For studies of L2 oral production in dialogues]

Ishikawa, S. (In press). Comparison of three kinds of alternative essay-rating methods to the ESL Composition Profile: An approach to lessen teachers' workloads in evaluating learners' L2 English essays.

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.

Leech, G. (1998). Preface. In S. Granger (Ed.), *Learner English on computer* (pp. xiv-xx). Harlow, England: Addison Wesley Longman.

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

（石川慎一郎　神戸大学　Email: iskwshin@gmail.com）