A NEW HORIZON IN LEARNER CORPUS STUDIES: THE AIM OF THE ICNALE PROJECT

Shin'ichiro ISHIKAWA
Kobe University
iskwshin@gmail.com

ABSTRACT

Learner corpus study, which is the field where applied linguistics, second language acquisition, and language education can work together, is expected to contribute to analysis of learners' interlanguage use as well as to development of effective L2 teaching/ learning materials. Coping with the increasing needs, varied learner corpora have been compiled to date. However, there are not so many corpora which rigidly control writing conditions and are suitable for contrastive interlanguage analysis (CIA). In the current paper, we would like to introduce the background of the International Corpus Network of Asian Learners of English (ICNALE) project and show how we can use the ICNALE for a new type of learner corpus studies.

1. INTRODUCTION

1.1 Why do we need a learner corpus?

When we use the term "corpus," we usually refer to a database of language produced by native speakers. For example, the British National Corpus, which includes 100 million tokens of written and spoken English produced by English (or rather "British") native speakers, has been widely used as a reliable sample of authentic English. The importance of corpus studies has been acknowledged and many dictionaries and TESOL materials are currently based on the BNC or similar native speaker corpora.

In addition, increasing numbers of researchers have been interested in "learner corpus" as well as conventional native speaker corpus. Then, why do we need to have a corpus collecting learner English, which may include unnatural or awkward expressions and even errors?

Firstly, learner corpus is needed for studies of "interlanguage," a special type of language existing between one's first language (L1) and second language (L2). For example, when Japanese learners of English speak or write in English, their oral or written utterance is called an interlanguage. It is neither Japanese as their first language nor native-like English as their target language, but a unique linguistic form seen between them. In the field of second language acquisition (SLA), interlanguage has been widely studied. Although its research approach has been traditionally theory-oriented, learner corpus collecting authentic examples of learners' use of the target language is expected to drastically advance varied interlanguage studies.

Secondly, learner corpus is also beneficial for foreign language teaching. By comparing learner corpus with native speaker corpus, we can identify learners' typical errors and overused or underused patterns. Most of the recent EFL dictionaries compiled in UK utilize various findings obtained from learner corpus analysis for sophistication of lexical descriptions (Ishikawa, 2011).

1.2 Various Leaner Corpora

Various learner corpora have been compiled to date. The largest and most influential is the International Corpus of Learner English (ICLE) (Granger, et al. 2002; Granger, et al., 2009). Also, several corpora focus exclusively on Japanese learners of English (JLE). The Japanese EFL Learner Corpus (JEFLL) (Tono, 2007) includes essays written by Japanese junior high and high school students, the Nagoya Interlanguage Corpus of English (NICE) (Sugiura, 2007) includes those by Japanese college students, and the NICT JLE Corpus (Izumi, Uchimoto, & Isahara, 2004) collects transcribed data of speech utterances in the English oral proficiency interview (OPI) test.

Each of these corpora has its merits, but when using them for contrastive interlanguage analysis (Granger, 2002), namely, a comparison between native speakers (NS) and non-native speakers (NNS) or that

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 Press. Typographical errors were corrected in May 2012.

between NNSs with different L1s, we must be careful about how we interpret the obtained findings, since differences in the writing conditions might influence the language used in the essays (Sugiura, 2007). This suggests that a more controlled learner corpus is needed for a more reliable cross-interlanguage study.

2. THE FEATURES OF THE ICNALE

The International Corpus Network of Asian Learners of English (ICNALE) is a digitalized database which the author is currently compiling. The ICNALE is primarily designed as a reliable data-source for analysis of the interlanguages of Asian learners of English.

Compared with other learner corpora, the ICNALE has several unique features, including 1) sufficient corpus size, 2) focus on Asian learners of English, 3) collection of English native speakers' essays, 4) control of writing conditions, 5) inclusion of writers' bio data, 6) open accessibility, and 7) collection of add-on data.

2.1 Corpus Size

Size does matter for reliability of corpus, which is regarded as a sample representing the linguistic population. We can estimate how the population is more precisely by observing a larger sample. Statistics proves that precision in estimation is in direct proportion to the square root of the sample size.

However, unlike in case of general corpora whose size has recently amounted to more than 100 million tokens, it is not easy to collect learner data extensively. Excluding the commercial database such as Cambridge Learner Corpus or Longman Learners' Corpus and the institutional ones such as Hong Kong University of Science & Technology Learner Corpus, the size of learner corpora has been relatively small.

The ICNALE has collected more than 1,000,000 tokens of essays written by a total of approximately 4,000 learners and it plans to be further expanded. Although the size of the ICNALE is smaller than that of the ICLE (3,700,000 tokens), the largest learner corpus ever built, it surpasses or competes with the sizes of other major learner corpora available to the public such as JEFLL (670,000 words), NICE (70,000 tokens), Giessen-Long Beach Chaplin Corpus (350,000 tokens), Janus Pannonius University Corpus (500,000 tokens), and Written Corpus of Learner English developed at Universidad Autonoma de Madrid (750,000 tokens).

2.2 Focus on Asian Learners

Most of the learner corpora ever built is largely Europeans-centered. Although the second version of the ICLE includes essays written by learners with as many as sixteen L1s (Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Turkish and Tswana), the coverage of Asian learners is rather limited.

Meanwhile, the ICNALE has collected data in eight countries and areas in Asia as well as data of NS. It covers all of the three "concentric circles" in Asia (Kachru, 1985), the Inner Circle where English is used as a native language (ENL), the Outer Circle where English is used as an official or a second language (ESL), and the Expanding Circle where English is learned as a foreign language (EFL).

Table 1 Data included in the ICNALE (As of August, 2011)

Circles	English Types	Countries	N of Essays	N of Tokens
Inner Circle	ENL	US, UK, Australia etc	546	124,862
Outer Circle	ESL	Hong Kong	200	47,235
		Pakistan	386	91,232
Expanding Circle	EFL	China	486	110,314
		Indonesia	408	94,608
		Japan	770	242,431
		Korea	404	89,168
		Taiwan	404	91,680
		Thailand	824	183,963

The ICNALE can be used not only for SLA-based interlanguage studies but also for socio-linguistic studies of the so-called "world Englishes."

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 Press. Typographical errors were corrected in May 2012.

2.3 Collection of NS Essays

Most of the existing learner corpora collect only learner data, which makes comparison between NS and NNS learners substantially difficult. Having no comparable NS data, some of the previous studies compare a leaner corpus with a large-scaled general corpus such as the British National Corpus (BNC) (Tono, 2007). However, learners' essays and materials included in general corpora, novels, newspapers, magazines, brochures, and letters, for instance, are qualitatively different and it can be problematic to interpret the results obtained from such a comparison.

The ICNALE includes essays written by NS in addition to those by NNS learners, which enables us to conduct a more robust NS/ NNS comparison. The table below shows the distribution of the countries of NS writers.

Table 2 NS writers' countries

NS Writers' Countries	N of Essays	Percentage
Australia	36	6.6
Canada	80	14.7
New Zealand	28	5.1
UK	68	12.5
USA	334	61.2

Data are taken not only from UK and USA but also from major countries in the Inner Circles. It is of note that Australia and New Zealand, which are Inner Circle countries in Asia, are included in the corpus. The ratio of the British English to the American English is controlled roughly to one to two.

While all of the NNS learners are college students, occupations of NS writers vary to some extent.

Table 3 Occupations of NS writers

NS Writers' Occupations	N of Essays	Percentage
Teachers	206	37.7
College Students	156	28.6
Business Persons	60	11.0
Translators	42	7.7
Others	82	15.0

Approximately forty percent of NS writers are teachers including professors, lecturers, and instructors, and thirty percent are college students. They are not necessarily TESOL or TEFL specialists. The NS module in the ICNALE intends to represent the language used not by specialists but by varied NS.

2.4 Control of Writing Conditions

Control of writing conditions is crucial when compiling learner corpora, because there are no reasonable ways to interpret the findings obtained from comparison of essays written in different conditions. By comparing a short essay which Japanese learners write with the help of a dictionary about smoking, and a much longer one which Chinese learners write without the help of a dictionary about leisure and hobby, for instance, we may be able to find out several seemingly interesting differences. However, no one can say which of the essay length, writers' L1, dictionary use, or topic causes the differences.

In the ICNALE project, we have prepared a detailed data collection guideline beforehand and made all the learners rigidly follow it. Thus, parameters which might influence linguistic production of learners are controlled as carefully as possible.

We permit a certain level of freedom in time and length so that all the writers can complete their essays, meaning no fragmentary or incomplete essays are included in the corpus. It should be noted that the degree of freedom in the ICNALE is much smaller than that in major learner corpora. In the ICLE, for instance, the time is not controlled ("untimed") and the length of an essay is regulated from 500 to 1000 tokens.

Table 4 Writing conditions

Conditions	Details
Time	20 to 40 minutes
Length	200 to 300 tokens
Dictionary Use	No
Spell Checker Use	Yes
Writing Mode	On the computers
Number of Topics	2

The ICNALE limits the number of topics to two: "a part time job" and "smoking" and the number of each type of essay is controlled to be the same. The below is a direction given to writers.

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

Concerning the topic variety in corpus, there are two fundamentally different approaches. One is the approach to try to include as many topics as possible. A greater variety of topics assures greater linguistic variety, which makes it possible to study wider range of vocabulary and grammar. Thus, the number of the topics in the ICLE, for instance, amounts to more than nine hundred (Sugiura, 2007). The other is the approach to try to control topic variety as strictly as possible. For, as mentioned before, a greater variety of topics leads to weakening the validity of cross-interlanguage comparisons. Both of these two approaches have their pros and cons, and the ICNALE, which is intended as a database for contrastive studies, adopts the latter approach.

2.5 Bio Data Collection

We often attribute linguistic features seen in a group of essays to their writers' nationality or L1 background. However, there are much more parameters potentially concerning them. Especially, factors such as age, proficiency, and years of studying English should be considered appropriately in data analysis.

Therefore, the ICNALE collects bio data from all the NNS writers, including country, sex, age, name of school, grade, academic major, years of studying English, and English proficiency.

Proficiency data is collected in two ways. First, if NNS writers have taken standard proficiency tests such as TOEIC®, TOEFL®, IEALTS®, and CET (National English proficiency test administered in mainland China), they are required to report the scores. Also, all of them are required to answer a monolingual version of the vocabulary size test developed by Nation and Beglar (2007) before writing essays. The test, which is originally a pen and paper one, is digitalized for the project. The test covers 14,000 words in total, but we use only the top 5,000 words, considering the proficiency levels of average learners in the area. Test takers choose an English word semantically matching the meaning of the target word shown in the sample sentence.

> see: They <u>saw</u> it.
> a. cut b. waited for c. looked at d. started

Prior to the project, we made five hundred Japanese college students at varied proficiency levels take both of the vocabulary size test and the TOEIC® test, which consists of listening and reading sections and is widely administered in Japan, Korea, and many Asian countries. Then, we obtained a score conversion formula based on the linear regression modeling, making it possible for us to estimate the scores in the TOEIC® test for all the NNS writers even if they have not taken the test. This is how all the learners are classified into four proficiency bands: Upper (>=700 in the TOEIC® test), Semi-upper (>=600), Middle (>=500), and Lower (<500).

In addition, we have experimentally conducted a questionnaire survey on approximately 300 Japanese learners of English. The questionnaire consists of 30 personality questions, 12 motivation questions, and 22 learning background questions.

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 Press. Typographical errors were corrected in May 2012.

2.6 Open Access

Maekawa (2011) emphasizes that only the database open to the public can be called a corpus. However, it is not easy to obtain a written agreement from each of the learners, who are usually students, and to clear the copyright restraints. This is why many of the existing learner corpora are not open to the public at all or require users to pay the license fee, which prevents wider use of them. According to the list of "Learner Corpora around the World," which is maintained by the Center for English Corpus Linguistics at University Catholique de Louvain, only 20 to 30 percent of more than one hundred learner corpora are freely available online or in the form of CD-ROM.

The ICNALE is freely available in two ways, online or download. The online version, which is called The ICNALE Online, offers several search functions as illustrated in the next section. The ICNALE is currently open to the public under the Creative Common License (Attribution-NonCommercial-NoDerivs 3.0 Unported). Users may copy, distribute and transmit the data, while they may not use it for commercial purposes, alter, transform, and build upon it. Users also have to attribute the corpus to its project leader.

2.7 Add-on Data

The Japanese learner module is one of the cores in the ICNALE and we have experimentally prepared several add-on data modules so that it can be investigated from a wider viewpoint.

Table 5 Add-on modules

Add-on Modules	Details	N of Tokens
Proof-writing module	Essays proofread and corrected by three professional NS proof writers	c 50,000
Rewriting module with an explicit corpus-based instruction	Essays revised by learners themselves, who are shown beforehand a list of words overused and underused by Japanese learners	c 5,500
Rewriting module with an implicit model-based instruction	Essays revised by learners themselves, who are shown beforehand a model essay written by NS	c 5,500
Japanese essays module	Essays written by Japanese learners in Japanese about the two same topics	c 60,000

By comparing an original learner essay module and a proof-writing module, we can investigate which words or expressions are corrected by NS, and by comparing an original essay module and two kinds of rewriting modules, we can see which kind of instruction is more effective for better rewriting by learners (Ishikawa, forthcoming). Also, by comparing the essays written in English and those in Japanese, we can objectively discuss the degree of L1 influence on learners' use of the target language (Ishikawa, 2010).

3. THE ICNALE ONLINE

So that a corpus can be widely used for varied research purposes, it is crucial to make it free online and to offer user-friendly search interface. The ICNALE Online (http://language.sakura.ne.jp/s/icnale) currently offers four search functions. Users can easily investigate the interlanguages of Asian learners even if they do not have corpus concordance software.

3.1 KWIC Search

With the KWIC (Key Word in Context) Search, users can obtain a list of context lines ("concordance lines") including the NODE or the target word, which allows them to see how a word is used in learners' interlanguage.

In order to get a list of concordance lines, users need to enter some word as a search target in the NODE box. Also, users can specify the target in more detailed ways: 1) as a NODE word (e.g. "go") or a context consisting of a NODE word and its collocates (e.g. "will" + "go" + "to"), 2) as a word form (e.g. "go" for "go") or a lemma ("go" for "go," "goes," "went," "gone," "going"), 3) with case insensitive ("go" for "go," "Go," "GO"…) or case sensitive ("go" for "go"), 4) as a word of all the POS (parts-of speech) types or a word used as a specific POS such as nouns, verbs, adjectives, and so on.

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 Press. Typographical errors were corrected in May 2012.

Fig. 1 KWIC Search interface

Then, users need to choose one or more countries of the writers. When possible, users can specify the writers' proficiency levels from four bands (Upper, Semi-upper, Middle, and Lower). And finally, users need to choose a topic. In a default setting, both of the two topics are preset.

The figure below shows a part of the KWIC Search output. It presents concordance lines including "play" as a NODE (Word unit=a lemma; Case=insensitive; POS= no specification) used by Japanese learners of English. Users can observe concordance lines online or download them as a spreadsheet data for a further analysis.

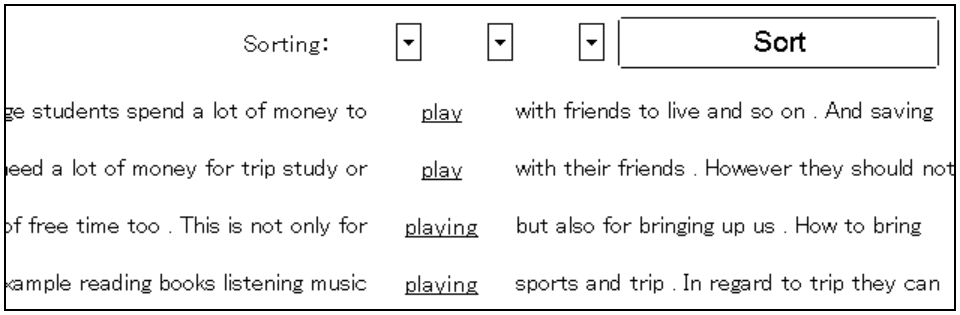

Fig. 2 Results of the KWIC Search

All the lines can be freely sorted by specifying up to three sorting keys (-3 [ie. the third word at the left of the NODE], -2, -1, +1, +2, or +3).

3.2 Collocation Search

With the Collocation Search, users can obtain a list of collocates or words frequently co-occurring the target word (NODE), which allows them to see how a word is phraseologically used in learners' interlanguage.

In order to get a list of collocates, users need to enter some word as a search target in the NODE box. Users can also specify the detailed search conditions both for the NODE word (word unit, case, and POS) and its collocates (word unit and case).

The figures below are positional collocation tables based on two kinds of statistics, which are a part of the Collocation Search outputs. The NODE is "think," the writers are Chinese learners, and both topics are chosen for analysis. The setting for the NODE is Word unit= a word form, Case= insensitive, and POS= no specification. That for the collocate is Word unit= a word form and Case= insensitive.

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 Press. Typographical errors were corrected in May 2012.

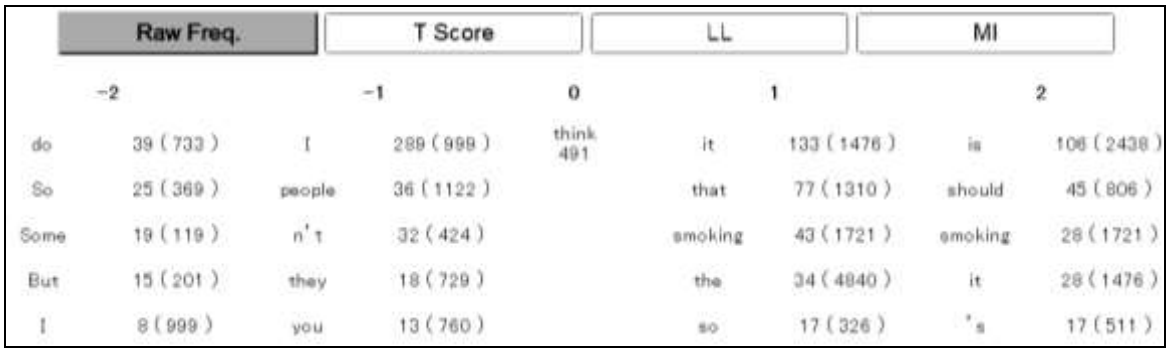

Fig. 3 Results of the Collocation Search (raw frequency)

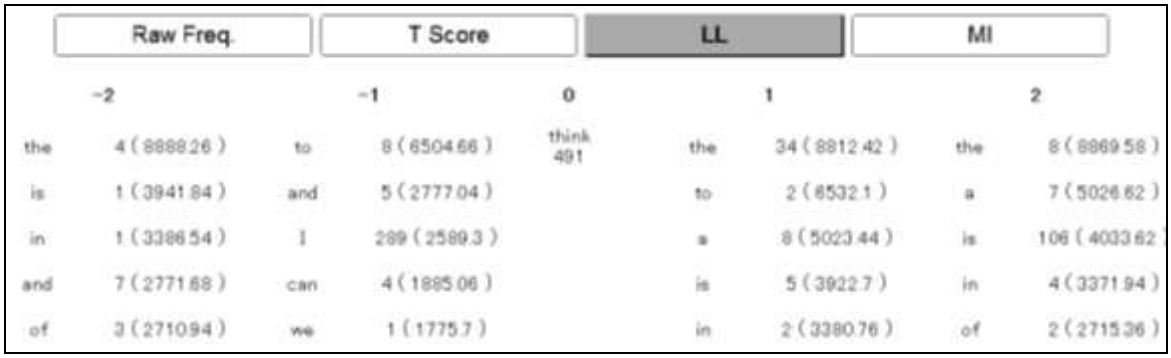

Fig. 4 Results of the Collocation Search (LL score)

The figure 3 shows that at the -1 (1 word left) position, "I" occurs most often and "people" follows, while at the R1 (1) position, "it" occurs most often and "that" follows. This suggests that Chinese learners tend to use the word "think" frequently in the collocational structure of "I [people] think (that) it …." Thus, collocation Search allows users to study the target word not independently but in a lexical network.

Users can choose four kinds of statistics as a criterion for collocate detection. Raw Frequency shows the unadjusted frequency of collocations. Both of T Score and LL Score, which are based on the *t*-value and log-likelihood value respectively, evaluate whether a collocation is statistically significant or not. Meanwhile, MI (Mutual Information) Score evaluates the strength of collocation. Degree of dependence on frequency is usually the highest in raw frequency, second in T Score, third in LL Score, and the lowest in MI Score. It is generally said that LL Score is the most balanced index for descriptive observation of a word's collocational pattern and MI Score is suitable for identifying less frequent but strange and unique collocations. See Ishikawa (2008) for the detail of calculation of each of the statistics.

3.3 Word List

With the Word List function, users can obtain a frequency list of words used by different groups of writers. Frequency is presented in two ways: raw frequency and per million word (PMW) adjusted frequency. The latter is more suitable when comparing two corpora of different sizes. Also, by adjusting the setting, users can obtain both of the word-form list and the lemma list.

The table below lists top five case-insensitive word forms used by five writer groups. The frequency is shown in the PMW for a mutual comparison.

Table 6 High frequent words

NS		China		Japan		Korea		Thailand	
the	39112	the	45960	the	32426	the	35327	the	44982
to	38262	to	34511	to	32178	to	33386	to	32882
and	28352	a	27385	I	25966	is	27544	and	24744
a	25237	in	24031	is	24852	a	25682	in	22336
of	21150	is	22227	a	22893	and	21375	a	20341

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 Press. Typographical errors were corrected in May 2012.

It is generally said that high-frequent word types, most of which are functional, are highly identical in spite of the difference in text types or writer groups, but the data shows that there exists a discrepancy even in the top five frequent words. For instance, "I" is used frequently only by Japanese learners, while "the" and "to" by all the learners, which will gives us a hint to analyze the characteristics of the interlanguage of each learner group.

3.4 Keywords Search

With the Keywords Search, users can obtain a list of words occurring significantly more often or less often in one of the two corpora compared. By analyzing keywords, which can be summation of the features of a given text or a corpus, we can discuss how learners use the vocabulary in their own unique way.

In order to obtain a list of keywords, users need to specify the target corpus and the reference corpus, for instance, Japanese writers as the target and native speakers as the reference, or Chinese writers as the target and Korean writers as the reference. Users can also specify the topic, word unit (word form or lemma), and case (insensitive or sensitive). These specifications should usually be the same for the two corpora.

The table below is the list of top 5 keywords in the interlanguages of Chinese, Japanese, and Korean learners compared to native speakers. The settings are Word unit= word form, Case= insensitive, and Topic= part time job.

Table 7 Keywords in the interlanguage of Chinese, Japanese, and Korean learners (Based on Chi2)

China				Japan				Korea			
Overuse		Underuse		Overuse		Underuse		Overuse		Underuse	
we	594	would	183	we	622	would	280	you	484	their	144
our	318	that	141	money	299	the	225	part	259	them	104
part-time	303	work	116	think	258	believe	190	money	219	that	101
can	253	I	108	I	219	financial	166	job	193	would	88
you	213	they	102	job	192	studies	158	tuition	109	they	66

Each of the three writer groups over/underuses words in a different way. For instance, Japanese learners overuse "we" and "I," Korean learners overuse "you," while Chinese learners overuse both of "we" and "you," but underuse "I," which suggests that Asian learners often tend to overuse the first or the second person pronouns and the interlanguage of Japanese learners is more writer-oriented, while that of Korean learners is rather reader-oriented. In contrast, there are some keywords common to all the writer groups. For instance, all of them underuse "would," suggesting that Asian learners in general are poor at using epistemic modal verbs. By comparing the keywords in varied interlanguages of Asian learners, we can distinguish tendencies specific to a particular country from those common to many countries in Asia.

As shown in the figure below, in the default setting, keywords are identified based on the chi-squared score (Chi2), but users can switch it to the log-likelihood.

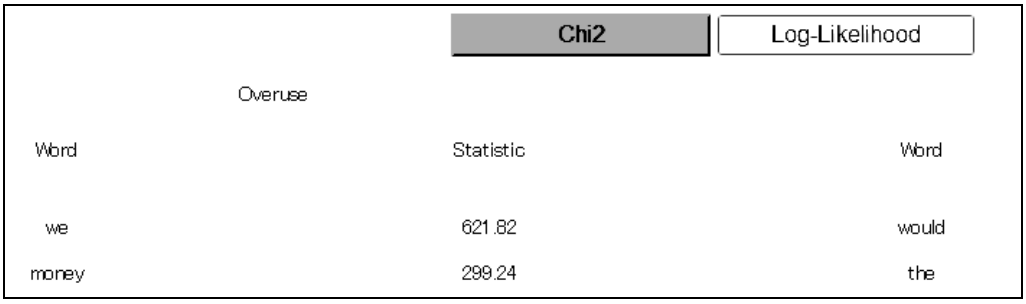

Fig. 5 Results of the Keywords Search

As mentioned before, use of the Log-Likelihood Score is often recommended especially when comparing the datasets whose sizes vary to a large extent.

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 Press. Typographical errors were corrected in May 2012.

4. SUMMARY

In the current paper, we have outlined the aim, principles, and design of the ICNALE, a newly compiled international learner corpus focusing on Asian learners of English. The ICNALE is unique especially in that it controls topics and writing conditions rigidly, making it a reliable database for various kinds of contrastive interlanguage analyses.

The ICNALE can be used for varied research purposes including investigation of a development of the interlanguage of a particular learner group, cross-interlanguage studies of different writer groups, which will shed a light on the uniqueness and commonality present in interlanguages of Asian learners. Also, as the ICNALE includes all the data of the inner, outer, and expanding circles in Asia, it can also be utilized for socio-linguistic studies concerning the so-called "Englishes" in the area.

Acknowledgements

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

References

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

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

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

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

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

Ishikawa, S. (2008). *Eigo kopasu to gengo kyouiku*. Tokyo: Taishukan shoten. [English corpus and language education].

Ishikawa, S. (2010). 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. (2011).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. (forthcoming). Writing, rewriting, proof writing: Gakushusha kopasu ni motozuku shutei kouka no keiryou teki kenkyu. [Writing, rewriting, proof writing: Learner corpus-based study on the effect of revisions]

Izumi, E., Uchimoto, K., & Isahara, H. (Eds.). (2004). *Nihonjin 1200 nin no eigo supiking kopasu*. Tokyo: Alc. [Speaking corpus of 1200 Japanese learners].

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.

O'Keefe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge, UK: Cambridge University Press.

Maekawa, K. (2011). Tokutei ryouiki kenkyu: "nihongo kopasu" to "gendai nihongo kakikotoba kinko kopasu." *The Proceedings of the Lecture Meeting Commemorative of the Completion of the Balanced Corpus of Contemporary Written Japanese*. 1-10. [Priority-area research: "Japanese corpus" and "Balanced corpus of contemporary written Japanese"]

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

Tono, Y. (2007). (Ed.). *Nihonjin chukosei 1 man nin no eigo kopasu: JEFLL corpus*. Tokyo: Shogakukan. [JEFLL Corpus: English corpus of 10,000 Japanese junior and senior high school students]