Intercultural Forum,1, 2, 2008. International Association for Intercultural Communication Studies
http://structural-communication.com/Intercultural-Forum/ICF1(2)2008.html

International Corpus of Creative English, Japan: What Are Creative Characteristics of English Used by Japanese Learners?

FUJIWARA, Yasuhiro, *Chugoku Gakuen University*, Japan

Abstract

The purpose of this paper is to shed light on "creative" characteristics of Japanese learners/users' writing in the International Corpus of Creative English, Japan (for ICCE, see Hassal, 2006). In the past few decades, some learner written corpora have been compiled, of which the most well-known one is the International Corpus of Learner English (ICLE; Granger, 1998). By means of these learner corpora, a great deal of research has been done in the fields of corpus linguistics, applied linguistics, and second language acquisition (Granger, 1998; Granger, Hung, & Petch-Tyson, 2002).

With some similarities between the ICCE and the ICLE such as 'written,' and 'tertiary,' one striking difference is whether learners/users are given a topic to write about: while the participants in the ICLE project are given a title, those in the ICCE, with only the rigorous limitation concerning the number of the words, are allowed to freely write anything such as prose, fiction, poetry and so forth －the contents of the ICCE are entirely up to the authors. It therefore would be possible to assume that some creative characteristics are reflected in their lexical choices: what they create by "words" naturally shows "creative" features. More specifically, their lexical preference will be analyzed in terms of 1) frequencies and 2) correspondence analysis. This type of research can be done in the ICCE, China or Korea, probably yielding some similarities in Asian Englishes and differences in each nation.

1. INTRODUCTION

In recent times, corpus linguistics has been influential in almost all the fields of linguistics where the main interest lies in machine-readable performance data in several languages, especially in English. In other words, it has developed by examining language use in order to compensate for the theoretical lack of the so-called native speakers' intuition, which has been overemphasized by Chomskian school (McEnery & Wilson, 1996, p. 25). According to this history of its development, the mainstream in corpus linguistics has narrowly focused on an established native speaker variety of English such as British English or American English (e.g. British National Corpus).

However, the recognition of the recent widespread use of English by nonnative speakers of English as well as native speakers (Kachru, 1986; Graddol, 2001; Crystal, 2003) has been leading corpus linguistics to focus on nonnative varieties of English as well. This can be clearly recognized by the existence of the authoritative International Corpus of English (ICE) proposed by Greenbaum (1996). The ICE is the most well-known international corpus comprised of several English varieties, both written and spoken, in the "Inner Circle" where English is used as a native language (ENL; e.g. Britain) and in the "Outer Circle" where English is used as a second language or an official language (ESL; e.g. India) (for the detailed distinctions of all "Circles," see Kachru, 1986). The ICE makes it possible to comparatively investigate these varieties, yielding similarities and differences between them (see some works in Greenbaum, 1996).

This expansion of the research scope in corpus linguistics has been quite lately extended to the "Expanding Circle" as well, where English is used as a foreign language (EFL) or as an international language ―English as a means of communication between people of different nations" (Smith, 1976 and others in Hino, 2001, p. 39). An

interesting issue concerning corpora in this zone is that whilst the linguistic samples are collected from one similar, tertiary level of students, there are two ways of conceptualization of them: *learner* (as in International Corpus of Learner English (ICLE) or the Louvain International Database of Spoken English Interlanguage (LINDSEI)), and *user* (as in the Vienna-Oxford Corpus of International English: Seidlhofer, 2001, 2005)[1]. This discursive practice is probably due to the blur/debatable status of English in the Expanding Circle: it can be described as floating between a foreign language (i.e. learner) and an international language (i.e. user). In a similar vain, Hassal (2006), with his conceptualization of informants as "users," proposed the International Corpus of Creative English (ICCE), the data collected by an "Extremely Short Story Competition" (ESSC), where university students at the tertiary level compete with their own creative writing in English applying a 50 word constraint strictly. This practice was exported to Japan in 2006 by The Japanese Association for Asian Englishes as one of the commemorative events to celebrate its 10th anniversary (for the details on the Japanese version of the ESSC, see http://essc.fit.ac.jp/en/).

The International Corpus of Creative English can be characterized as a corpus similar to the ICLE in the sense that both consist of written data from the tertiary students (its informants are traditionally regarded as "learners" in corpus linguistics), one noticeable difference is whether learners/users are given a topic to write about: while the participants in the ICLE project are given a title (e.g. '*Money is the root of all evil*'), those in the ICCE, with only the rigorous limitation concerning the number of the

[1] As noted above, the distinction between '*learner'* and '*user'* in referring to a student at the tertiary level has been more and more ambiguous, especially in the field of *World Englishes* (Jenkins, 2006). There are, however, two on-going corpus projects in which the notion of *user* does not include university students — it refers to professionals such as scholars as in the Corpus of English as a Lingua Franca in Academic Settings (CELFA: Mauranen, 2003), and journalists as in Japanese User Corpus of English (JUCE: Fujiwara, 2007).

words, are allowed to freely write anything such as prose, fiction, poetry and so forth —the contents of the ICCE are entirely up to the authors. It therefore would be possible to assume that some creative characteristics are reflected in their lexical choices: what they create by 'words' naturally shows 'creative' features. Furthermore, the Japanese version of the ICCE compiled by the ESSC 2006, consists of stories written not only by university students but also by senior and junior high school students in Japan. This variation can lead us to a thought-provoking question: *Is there any difference in creativity as to their growth?* In short, although this author is fully aware of the conceptualization problem of learners/users in this field, the ICCE must be deeply meaningful in investigating 'creative' characteristics in their writing. More specifically, their lexical preference will be analyzed by means of 1) frequencies and 2) correspondence analysis.

2. METHOD

2.1. Corpus

In this chapter, I will summarize the characteristics of the ICCE, Japan. As shown above, the most notable feature of this corpus is their 'creative' writing. Although there are some written corpora comprising of written works by several types of learners/users here and there, there is, to the best of my knowledge, no corpus compiled under the condition where they are allowed to write as they like.[2] Following Granger's corpus design criteria (1998, p. 8), the detailed information on the corpus is shown in Table 1 below.

[2] In Japan, there are several learner written corpora such as the Japanese EFL Learner Corpus (JEFLL) and the Corpus of Japanese English Learners (CJEL), of which all were compiled of the compositions written based on a given topic.

Table 1: ICCE-Japan corpus design

Language		**Learner/User**	
Medium	Written	Age	13 –25 (Ave. 18.46)
Genre	Essay in 50 words	Sex	Mixed
Topic	Free	Mother Tongue	Japanese
Technicality	low	Region	Japan
Task	Typing	Other Foreign Languages	Mostly none
Setting	No time limitation	Level	Beginner – Advanced
		Learning Context	EFL/EIL
		Practical Experience	Approximately 2.5-15 years

Table 2: The number of works, token, type, TTR, Std TTR in each subset of the ICCE-Japan

Corpus	N	Token	Type	TTR	Std TTR
JHS	35	1,760	565	32.14	35.70
SHS	105	5,071	1,207	23.87	40.60
UNI	293	14,705	2,358	16.08	40.04
Overall	433	21,536	2,956	13.76	39.97

Turning to each subset of the ICCE-JAPAN, Table 2 above shows the comparisons of the data produced by 1) junior high school students (hereafter, JHS), 2) senior high school students (SHS), and 3) university students (UNI) in terms of number of the story, token, type, type-token ratio, and standardized type-token ratio per 1000 words. Please note that due to the difference in the way of counting the number of the words by means of software used for this analysis, namely *Microsoft Word* and *WordSmith* 4.0, the number of tokens does not exactly reflect the number of the works.

2.2. Procedure

For the analysis of the ICCE-Japan, I firstly text-markup it in the style of Standard Generalized Markup Language (SGML). The other thing to note here is that this corpus is still not grammatically tagged, mainly because a numerous number of samples contain what we traditionally call "errors." All the automatic tagging programs available such as CRAWS or TOSCA Tagger are based on a so-called "standard" English. This therefore means the data from English learners are not properly tagged at a satisfactory level of accuracy. As mentioned above, however, I limit the discussion to their creative use of "lexis": some learners' grammatical errors of whole texts are not considered because they are not the subject of this paper. Below is the sample of the text.

Figure 1: Text sample markedup in the ICCE-JAPAN.

<no> 06-153 </no> <title> Fire in your mouth </title> <name> A-406 </name>
<sex> M </sex> <age> 21 </age> <EL> 15y 0m </EL> <OE> 0 </OE>
<text> The words you speak are just like a smoldering fire. Words can be a fire that warms your heart. Words can be a blazing fire that burns you to death. Words can be a torch that lightens the way you walk. How to use the fire is up to you. </text>

Although this corpus was not morpho-syntactically tagged, there is one word manually annotated, "*like*." The reason for this post-hoc annotation is because '*like'* as a preposition (e.g. Clouds are *like* travelers) is, as I will describe below, not the scope of this paper.

Then, the markedup ICCE-JAPAN was processed, and the frequency data and the wordlist for each subcorpus were obtained by *WordSmith* 4.0. Since this research's focus is specifically on their unique, 'creative' features showed by their lexical choice, I then

subtracted from the wordlists some common words in all types of the corpora: some function words (i.e. articles, prepositions, and conjunctions), pronouns (e.g. I, you, s/he), *be* verbs (e.g. am, are, is), and demonstratives (e.g. this, that), in order to clearly designate creativity of each subset of the corpora. The obtained frequency data were finally normalized per 5000 words, and statistically analyzed by the correspondence analysis.

2.3. Research Limitations

Before turning to the results of the study, it is desirable to describe some research limitations. One is comparability between each subset of the corpus in a quantitative sense. As shown in Table 2, the difference in the size of each component should not be ignorable (JHS: N =1,760; SHS: N = 5,071; UNI: N = 14, 705). Considering the variance between each component, it seemed appropriate to set up the standard level as 5000 words, but it would be much more proper to wait for the time when the comparability is to be met by collecting the similar number of the works in each group. The other is the wordlists of the groups were not lemmatized due to the existence of error-like features, and inherent problems of the lemmatization. The ICCE-JAPAN including written English at the beginner level unsurprisingly has some misspelled words according to a "standard" of English, and it would not be practical to take into consideration all nonstandard features. In addition to that, lemmatization is, as Stubbs (2002) pointed out, not at all a simple procedure to take, for it is somewhat arbitrary to decide how you include some words under one lexeme. Consider the following words, *confuse, confused, confusing,* and *confusion*: some might include all these words into one lexeme, or divide them into two or three. Moreover, some forms, generally thought to be under the same lexeme, would behave so differently in a real text (for details, see

Stubbs, 2002). With these two reasons, the lemmatization of the wordlists was not done, but even with the lists of the word-forms, it would be possible to grasp some tentative tendency of their "creative" features.

3. RESULTS & INTERPRETATIONS

3.1. Frequencies

The first result, by focusing on frequencies, concerns their lexical preferences by each level of learners/users. The top twenty words by raw frequencies in the texts written by junior/senior high school students, and university students are shown in Table 3. This table would show their "first priority" in their life, and its transition as to their growth. In the first place, junior high school students are mainly interested in *friendship* as shown in the first rank word, "*friends*." Although *friend* and *friends* are located in the thirteenth rank in the SHS and in the eighteenth in the UNI, they have something else as their primary concern. Turning to the senior high school students, as the words, "*want*," "*will*," and "*dream*," suggest, they are mainly occupied with their *future*. Considering the educational system in Japan, senior high school days must be one of the significant "fork" in their life: before their graduation, they are supposed to decide the way of their life, whether they go on to a university/college or start to work and what field they would like to be specialized in. In the last place, as the following tokens, *like* as a verb and *love*, indicate, the main interest of university students are assumed to be "*love*." It might not be so surprising that as their age becomes over 18, they would think of "*love*" for their beloved such as their girl/boy friend, their family, and somebody important in their life.

It should be, however, noted that some future expressions such as "*will*," "*want*," and "*dream*" are used to refer to something else rather than their own future, or their

actual dreams at night, and "*love*" or "*like*" are of course used to mention their preferences for sports, foods, or something rather than somebody. Although this author is aware of the ignorance of the context, there would be, to a certain degree, some validity in assuming the order of their priority shown by their top 20 lexical preferences and its transition from junior high school students to university students.

Table 3: Top twenty words in JHS, SHS, and UNI.

	JHS		SHS		UNI	
N	Word	Freq	Word	Freq	Word	Freq
1	FRIENDS	20	SO	42	HAVE	118
2	VERY	18	HAVE	39	CAN	96
3	SO	17	DON'T	32	NOT	87
4	CAN	15	NOW	29	WANT	85
5	NOT	15	VERY	29	BECAUSE	84
6	DO	13	WANT	28	SO	79
7	HAVE	13	WILL	28	VERY	77
8	MUSIC	13	DO	24	DO	64
9	THINK	13	MANY	22	PEOPLE	63
10	PEOPLE	12	SCHOOL	22	TIME	63
11	WANT	11	WHAT	22	LIKE_verb	61
12	BECAUSE	8	DREAM	21	MANY	58
13	BECOME	8	FRIEND	20	LOVE	56
14	EVERYDAY	8	HAPPY	20	ALWAYS	54
15	JAPANESE	8	LIFE	20	ONE	54
16	JUGGLING	8	TIME	20	WORLD	54
17	WORLD	8	CAN	19	ALL	53
18	HAS	7	BECAUSE	18	FRIENDS	52
19	IMPORTANT	7	PEOPLE	18	NOW	48
20	LIFE	7	UP	17	DON'T	47

3.2. Correspondence Analysis

Table 4 below is the contracted cross table to demonstrate top 100 words according to the standardized frequency in total by 5000 words, with the raw and the standardized frequencies in all the three subsets of the corpus. In the same way, the cross table for words ranked from 101 to 200 was made. These two cross tables are subject to a multivariate statistical analysis called Correspondence Analysis to clarify the correlation of each educational level and normalized frequency of each word. This analysis can show the relative relationship of these two nominal variables by mapping the results to visually recognize the similarities and dissimilarities of them.

Table 4: Top 100 words according to frequency per 5000 words with raw frequency in each subset

			Overall		JHS		SHS		UNI	
Raw R	SD R	Word	Freq	SD Freq	Freq	per 5000 w	Freq	per 5000 w	Freq	per 5000 w
2	1	SO	138	115.16	17	48.57	42	39.62	79	26.96
1	2	HAVE	170	114.21	13	37.14	39	36.79	118	40.27
4	3	VERY	124	105.07	18	51.43	29	27.36	77	26.28
3	4	CAN	130	93.55	15	42.86	19	17.92	96	32.76
6	5	NOT	118	87.64	15	42.86	16	15.09	87	29.69
5	6	WANT	124	86.85	11	31.43	28	26.42	85	29.01
12	7	FRIENDS	82	84.32	20	57.14	10	9.43	52	17.75
8	8	DO	101	81.63	13	37.14	24	22.64	64	21.84
9	9	PEOPLE	93	72.77	12	34.29	18	16.98	63	21.50
7	10	BECAUSE	110	68.51	8	22.86	18	16.98	84	28.67
26	11	MUSIC	62	60.49	13	37.14	11	10.38	38	12.97
13	12	NOW	82	58.03	5	14.29	29	27.36	48	16.38
28	13	THINK	60	58.00	13	37.14	8	7.55	39	13.31
20	14	LIFE	71	53.88	7	20.00	20	18.87	44	15.02
17	15	WORLD	75	53.55	8	22.86	13	12.26	54	18.43
21	16	WHAT	71	52.57	6	17.14	22	20.75	43	14.68
16	17	LOVE	77	52.32	7	20.00	14	13.21	56	19.11
11	18	MANY	84	51.98	4	11.43	22	20.75	58	19.80
23	19	WILL	65	51.62	5	14.29	28	26.42	32	10.92
15	20	DON'T	80	49.09	1	2.86	32	30.19	47	16.04

<table continued>

Table 4 continued

63	80	HEART	31	19.12	1	2.86	10	9.43	20	6.83
140	81	CANNOT	17	18.98	5	14.29	1	0.94	11	3.75
73	82	NEVER	27	18.35	1	2.86	11	10.38	15	5.12
287	83	FREE	9	18.17	6	17.14	0	0.00	3	1.02
75	84	HARD	26	18.12	2	5.71	7	6.60	17	5.80
120	85	DIFFERENT	19	17.75	4	11.43	2	1.89	13	4.44
81	86	HIGH	25	17.67	1	2.86	11	10.38	13	4.44
70	87	CAN'T	28	17.60	2	5.71	5	4.72	21	7.17
167	88	CAME	15	17.59	4	11.43	4	3.77	7	2.39
56	89	JAPAN	33	17.50	2	5.71	2	1.89	29	9.90
361	90	CULTURE	7	17.48	6	17.14	0	0.00	1	0.34
362	91	KANJI	7	17.48	6	17.14	0	0.00	1	0.34
106	92	EACH	22	17.46	3	8.57	4	3.77	15	5.12
50	93	SUMMER	35	17.36	0	0.00	9	8.49	26	8.87
132	94	AFTER	18	17.30	3	8.57	6	5.66	9	3.07
263	95	FAMOUS	10	17.20	5	14.29	2	1.89	3	1.02
61	96	EVERY	32	17.16	2	5.71	2	1.89	28	9.56
417	97	AKITA	6	17.14	6	17.14	0	0.00	0	0.00
418	98	ANGEL	6	17.14	6	17.14	0	0.00	0	0.00
113	99	EVEN	21	17.12	3	8.57	4	3.77	14	4.78
121	100	SOMETIMES	19	17.04	3	8.57	5	4.72	11	3.75

The results of Correspondence Analysis are shown in Figure 2 for top 100 words, and Figure 3 for words ranked from 101 to 200. The distributed variables in correspondence analysis are derived from total chi-square values, i.e. inertia. In these figures, Dimension 1 with explanation rates nearly 80 % (80. 03 % for Figure 2, and 78.83% for Figure 3) could be interpreted as their *growth*. The youngest group (i.e. jhs) was plotted on the left side of the figure, and the older groups (i.e. shs and uni) were on the right. In the both figures, the variables of university students (uni) were located in the first quadrant, those of senior high school students (shs) in the fourth quadrant, and those of junior high school students (jhs) in the third quadrant.

In the origin, the point of intersection where the axes meet, some common words used by all groups are clustered, of which the examples in Figure 1 are frequently-used verbs like *has*, *do*, and *went*, and intensifiers such as *so*, *much*, and *very*. The plotted words in the first, the third, and the fourth quadrants are more specific to each group.

Figure 2: Correspondence Analysis of word preference across educational backgrounds: 1-100

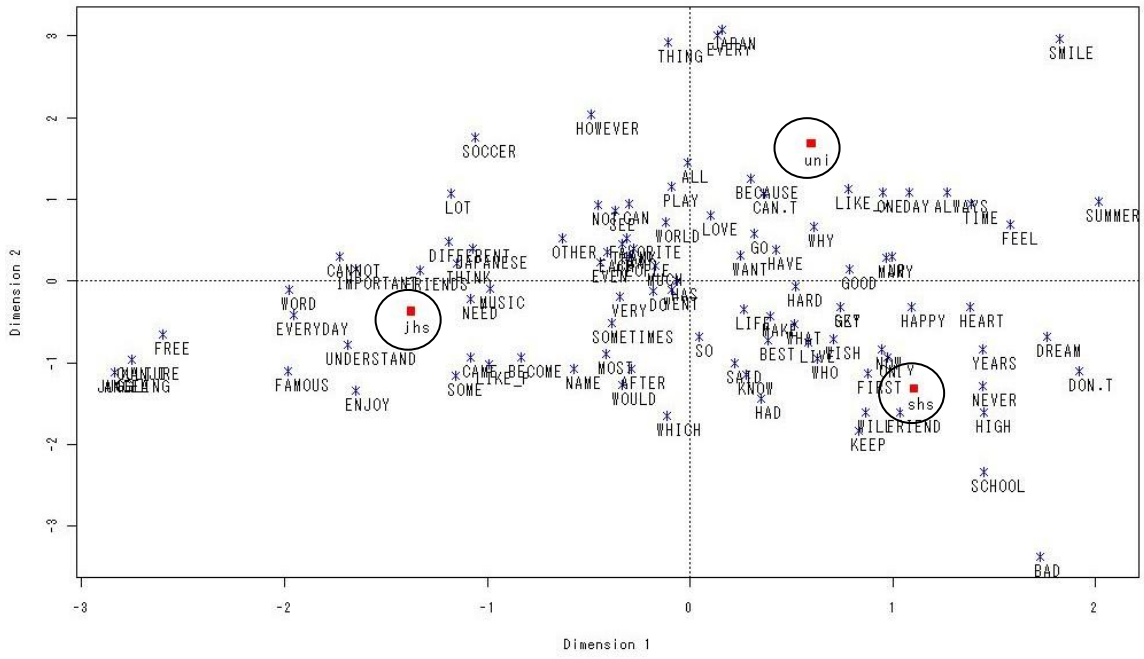
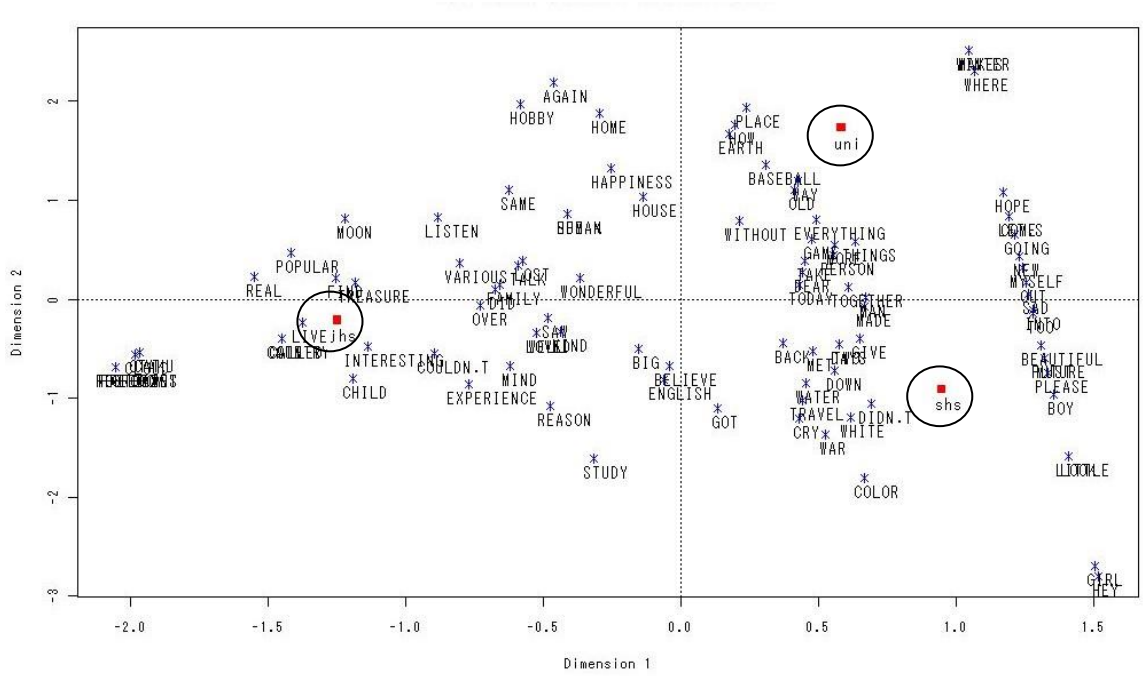


Figure 3: Correspondence Analysis of word preference across educational backgrounds: 101-200



These figures would be difficult to explain because there are so many variables of word-forms (N = 100 for each figure), and due to their complexity, these results might be interpreted in several ways. This mapping, however, visually shows each group's tendency to "imagine" in their "creative" mind. One possible interpretation might be that a creative feature of writing by JHS is something *fun*, seen from the following words: *enjoy*, *fine*, *interesting*, *friends*, *treasure*, and *music*. As for SHSs, their writing can be characterized not only as their *future* as shown in the use of *will*, *wish*, and *dream*, but also as *sorrow*, noted by some tokens, *bad*, *sad*, and *cry*. Added to this, it is really interesting to see that SHS have a stronger tendency to use *girl* and *boy*, and such interest might be, as the time goes by, turned into "*love*," the university students' primary concern. Additionally, we should not overlook that UNI might begin to have *a broad view of the world* as shown in their overused tokens such as *Japan*, and *earth*. *World* is plotted around the origin, for it is often used by secondary students as well as tertiary students. This is probably because this Extremely Short Story Competition in Japan was held in 2006, when the *World Cup* took place in Germany.

3.3. Summary

So far we have seen some "creative" characteristics in the 50-word essays written by Japanese learners/users of English by focusing on their lexical preferences. Because of the difference in size between each subcorpus of the ICCE-JAPAN, I am aware of that all the results should require much caution in interpreting. The given, available data and the analysis nonetheless allow me to argue the following points.

1) In analyzing frequencies with the specific focus on individual groups, we might be able to say that as they become older from junior high school students to university students, their primary concern will go through the transition from *friendship*

through *future* to *love*.

2) Considering the relative relationship between each level of the students and their lexical use by means of Correspondence Analysis, it would be, though tentative, possible to state that if they are allowed to write freely, a) junior high school students might write down something *fun*, predominantly with *friends*. b) Senior high school students would be rather interested in the *future* events, perhaps, their own future, and some *negative* incidents. c) In the case of university students, they would be grown up enough to consider *love*, and *global* issues.

All the above issues are the summary of the results and interpretations of this paper.

4. CONCLUSION

The main aim of this paper is to shed light on "creative" characteristics of Japanese learners/users' writing in the International Corpus of Creative English, Japan. Although this research has, to some extent, revealed their creative features in their essays, much should be left to be done. In order to compensate for the limitations of this study, and to further investigate "creative" characteristics in Asian Englishes, we need future research in the following directions:

1) The comparability between each subcorpus must be guaranteed, with the sufficient number of the words in corpus linguistics. Also, if a similar corpus in some countries, especially Asian countries such as China or Korea, is compiled, there would be a literally "breakthrough" in investigating "creative" characteristics in Asian Englishes— the comparisons between the ICCE-JAPAN, and, say, the ICCE-China must be fascinating by seeing some similarities and differences in each nation.

2) These results should be followed by qualitative analysis. As McEnery and Wilson (1996, p. 77) argue, 'both qualitative and quantitative analyses have something to contribute to corpus study' since 'qualitative analysis can provide greater richness and precision, whereas quantitative analysis can provide statistically reliable and generalizable results.' In line with their argument, it is especially important to focus on a context where some tokens focused like "love" are used.

With these lines of future research, it would be possible to deeply understand human creativity, accompanied with cultural differences, and universal humanity.

References

Crystal, D. (2003). *English as a global language* (2nd ed.). Cambridge: Cambridge University Press.

Extremely Short Story Competition 2007 (n.d.). Retrieved October 15, 2007, from http://essc.fit.ac.jp/en/

Fujiwara, Y. (2007). 'Compiling a Japanese user corpus of English.' *English Corpus Studies, 14*, 55-64.

Graddol, D. (2001). "English in the future." In A. Burns., and C. Coffins. (Ed.), *Analysing English in a global context: A reader* (pp. 36-37). New York: Routledge.

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: Longman.

Granger, S., Hung, J., & Petch-Tyson, S. (2002). *Computer learner corpora, second language acquisition and foreign language teaching*. Language Learning and Language Teaching, Vol. 6.

Greenbaum, S. (1996). *Comparing English worldwide: The international corpus of English.* New York: Oxford University Press.

Hassal, P. (2006). "Developing an international corpus of creative English." *World Englishes, 25*, 131-151.

Hino, N. (2001). "Organizing EIL studies: Toward a paradigm." *Asian Englishes, 4,* 34-65.

Jenkins, J. (2006). "Current perspectives on teaching world Englishes and English as a lingua franca." *TESOL Quarterly, 40,* 157-181.

Kachru, B.B. (1986). *The alchemy of English: The spread, functions and models of non-native English in the world.* Oxford: Pergamon Press.

Mauranen, A. (2003). "The corpus of English as lingua franca in academic settings." *TESOL Quarterly, 37,* 513-527.

McEnery, A. M., and Wilson, A. (1996). *Corpus linguistics.* Edinburgh: Edinburgh University Press.

Smith, L. (1983). *Readings in English as an international language.* Oxford: Pergamon Press.

Seidlhofer, B. (2001). "Closing a conceptual gap: The case for a description of English as a lingua franca." *International Journal of Applied Linguistics, 11,* 135-158.

Seidlhofer, B. (2005). "Standard future or half-baked quackery? In C. Gnutzmann and F. Intemman (Ed.). *The globalisation of English and the English language Classroom* (pp. 159-173). Tubingen, Germany: Narr.

Stubbs, M. (2002). *Words and phrases: Corpus studies of lexical semantics.* Blackwell.