A Corpus Study on English Language Use in Academic Writing by Japanese EFL Students: A Comparison of Their Use of Colligation and Collocation Patterns of Transitional Words with That of Native Speakers

Chizuko SUZUKI, Susan FUKUSHIMA,
Yumiko KINJO, & Shota YOSHIHARA

日本人大学生の英語使用に関するコーパス研究

－アカデミックライティングにおける接続詞の共起と連結パターンの英語母語話者との比較結果－

鈴木　千鶴子・福島　スーザン・金城　由美子・吉原　将太

Abstract

This paper presents part of a research project for providing Japanese university students in an EFL setting with an independent learning system to facilitate the writing of their graduation papers in English based on corpus studies. This study focuses on the comparative analyses of three corpora. The first is a Japanese A-Grade student paper corpus, the second contains the Japanese B-Grade papers and the third is a native speaker corpus, from the Michigan Corpus of Upper-Level Student Papers, (MICUSP) (The Regents of the University of Michigan, 2009). The authors are using these corpora to investigate the usage of ‘transitional word (s)’ and ‘modal adjuncts/adverbs’ as factors for determining text cohesion, in terms of their colligation and collocation patterns. Data analysis revealed the students’ tendencies as follows: 1) Regarding transitional words and phrases, i) overuse of basic common words, in contrast to underuse of complicated phrases; ii) a transitional word or phrase being used exclusively in a certain fixed position; 2) regarding modal adjuncts/adverbs, i) a modal adverb also being used exclusively in a fixed position; ii) modal adjuncts/adverbs rarely being used in collocation with modal auxiliaries; and 3) all of these tendencies are more obvious in the B-Grade (lower level) papers corpus than in the A-grade corpus, which suggests a developmental process of rhetorical discourse competence. The data in this study was used to develop a tutorial page in an e-learning system to help the students with their writing in English at the authors’ university in Japan.

Keywords: EFL learners’ corpus research, discourse analyses, cohesion, colligation and collocation, developmental process of rhetorical competence

Introduction

Studies into learner writing have covered a range of academic linguistic factors and second language learners, including academic vocabulary use (Paquot, 2010), Chen's (2006) study of conjunctive adverbials by high level Taiwanese writers, Granger and Tyson's (1996) comparison of first and second language writers' use of connectors in their essay writing in English, as well as Milton and Tsang's (1993) investigation of connectors in the writing of EFL students. These comparative studies indicate that first and second language writers employ different vocabulary and grammatical constructions in their writing.

The current project aimed at providing Japanese university students with an independent web-based online learning system to facilitate the writing of their graduation papers in English, specifically based on corpus linguistics studies. The scheme of the project is illustrated in Figure 1 below as a work schedule chart according to time sequence. This paper deals with the part encircled by a heavy gray line in the chart: Mainly comparative analyses of students' corpora with MICUSP to produce content materials for the tutorial part of the e-learning system.

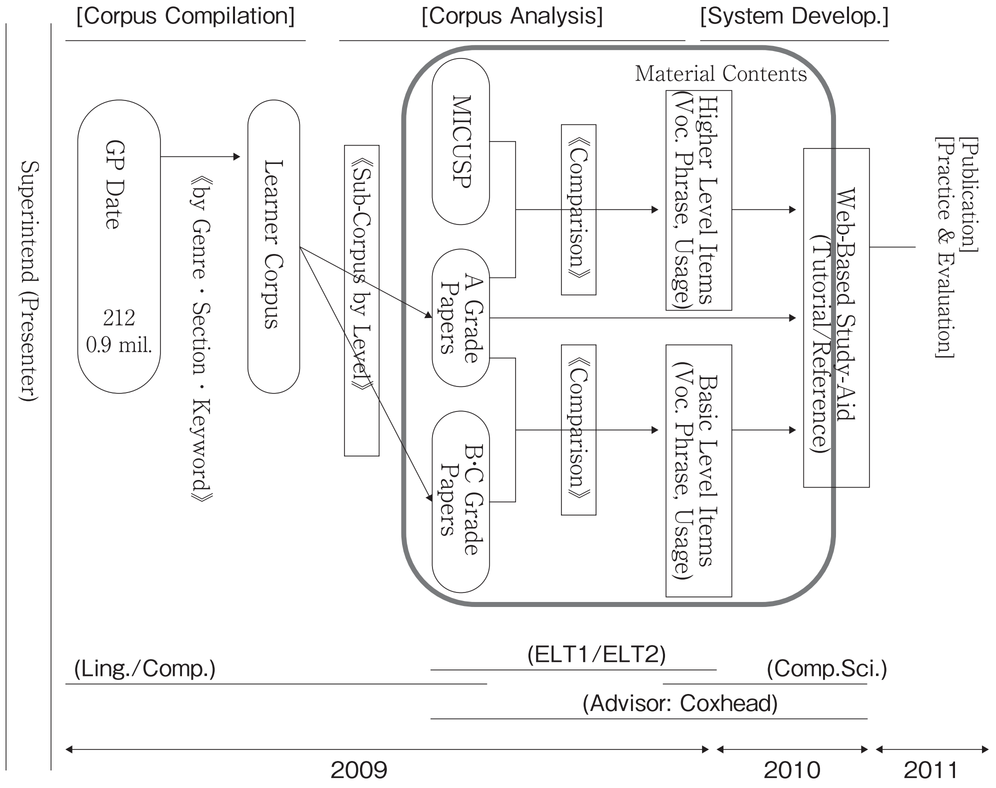

Figure 1. The entire project scheme.

In the project the following studies had already been carried out:

1) Corpus compilation of the students' graduation papers: JC-GP (Junshin Corpus of Graduation Papers) 2005-2009, approx. 900,000 tokens (See Table 1 below.)
2) Overviewing analyses of JC-GP (Suzuki et al., 2009)
3) Evaluating analyses of JC-GP for reference materials (Kinjo et al., 2010)
4) Corpus classification into 2 sub-corpora by grade, A and B (JC-GP_A & JC-GP_B), then into 4 sub-corpuses by section: Abstract, Introduction, Discussion & Conclusion; into 10 disciplines: Arts, Computer-Based Online Learning Materials, Culture, Education, Global Issues, History, Language Studies, Literature, New Technology, & Sociology
5) Developing a sentence search concordance system of JC-GP by loading the JC-GP_A corpus data (Fukushima et al., 2010)
6) Analyzing JC-GP_A compared with native speakers' corpus, MICUSP in terms of transitions (Suzuki et al., 2011)

These earlier studies based on these corpora led the present authors to several conclusions. The students' vocabulary knowledge was proved to have reached a required level for writing graduation papers in terms of size and range through analyzing the JC-GP compared with West's (1953) GSL (General Service List) revised by Paul Nation and available on his website in the Range and Vocabulary Profile program (go to http://www.victoria.ac.nz/lals/staff/paul-nation.aspx). The coverage of the GSL over the JC-GP corpus is 82.83%. The authors also carried out an analysis of the vocabulary used by the writers as in the Academic Word List (AWL) (Coxhead, 2000). The coverage of the AWL is 5.51%. This coverage figure is roughly half that of the AWL over the corpus of academic writing gathered by Coxhead (2000). A summary of AWL coverage figures over a range of corpora can be found in Coxhead (2011). Table 2 shows, however, the students' written English remains apparently distanced from native likeness. Connectors or transitional words and some modal adjuncts/adverbs were, in particular, pointed out to be necessary for deeper examination, in terms of colligation and collocation in order to elucidate the students' rhetorical discourse competence regarding cohesion.

Table 1　Profile of the Students Paper Corpus: JC-GP (WordSmith 5.0)

Year	Papers no.	Word Type	Token	TTR	Stand. TTR
2001	42	7,528	150,991	6.49	34.97
2002	42	7,031	183,508	5.18	34.28
2003	41	8,027	183,724	5.74	36.15
2004	39	7,861	180,047	5.70	33.81
2005	48	9,782	245,386	5.19	34.89
Overall	212	19,800	943,656	2.70	34.82

Table 2 Coverage of JC-GP Compared to GWL and AWL

WORD LIST	TOKENS/%	TYPES/%	FAMILIES
one	720078 / 77.49	3460 / 14.01	998
two	49580 / 5.34	2428 / 9.83	922
three (AWL)	51184 / 5.51	2013 / 8.15	562
Not in the lists	108441 / 11.67	16790 / 68.00	?????
Total	929283	24691	2483

The students' two sub-corpora JC-GP_B and JC-GP_A were investigated and compared with MICUSP as a native speaker students' corpus, focusing on the use of transitional words and phrases, and that of modal adjuncts/adverbs. This study, accordingly, investigated the following research questions:

1. What kinds of transitions were overused/underused in each corpus? [frequency of transitions]
2. Where were the transitions used, in the initial, medial or final position of a sentence in each corpus? [colligation patterns of transitions]
3. Where were some modal adjuncts/adverbs used, in the initial, medial or final position of a sentence in each corpus? [colligation patterns of modal adjuncts/adverbs]
4. How often were the modal adjuncts/adverbs used in collocation with modal auxiliaries in each corpus? [modal adjuncts/adverbs' collocation with auxiliaries]

Methods

The data analyzed consisted of the following three corpora:

1) JC-GP_B: 137 lower grade papers covering all the ten disciplines. The papers were evaluated lower by two or more raters out of three university teachers, including at least one native speaker teacher.
2) JC-GP_A: 65 higher grade papers evaluated higher in the same way as 1) by two or more raters out of three raters, also covering the ten disciplines.
3) MICUSP: 68 English native speaker students' papers covering five disciplines: Education, Economics, Linguistics, Literature, and Sociology, which largely overlap with the most common disciplines of JC-GP.

The corpora 1) and 2) were compiled by the present authors, and the corpus 3) was collected by the present authors from the MICUSP Simple Interface (BETA version) (MICUSP, 2009). The profiles of the three corpora are as listed below in Table 3.

To analyze the JC-GP_A corpus, a Sentence Search System of Graduation Papers developed by

Table 3　Profiles of Three Corpora Analyzed (WordSmith 5.0)

text file	JC-GP_B	JC-GP_A	MICUSP
file size	4,320,043	2,056,314	741,593
tokens (running words) in text	714,570	336,230	122,185
tokens used for word list	694,217	326,023	120,061
types (distinct words)	21,510	13,581	9,665
type/token ratio (TTR)	3.10	4.17	8.05
standardized TTR	35.00	35.04	39.12
standardized TTR std.dev.	64.86	63.51	59.72
standardized TTR basis	1,000	1,000	1,000
mean word length (in characters)	4.81	4.89	4.93
word length std. dev.	2.61	2.73	2.77
sentences	40,697	17,852	4,965
mean (in words)	17.06	18.26	24.18
std. dev.	8.74	9.19	14.00

one of the present authors, which was built into the Junshin Online Academia on the website: http://www.n-junshin.ac.jp/GradPaper/ (Fukushima et al., 2010), was used. Moreover, AntConc 3.2.3, developed by Laurence Anthony, which is obtainable at http://www.antlab.sci.waseda.ac.jp/antconc_index.html, WordSmith 5.0, as well as the search system of MS Word 2010 were used for analyzing JC-GP_A and MICUSP.

As regard to transitions, six words (‘because’, ‘however’, ‘therefore’, ‘(al)though’, ‘hence’, & ‘then’) and four phrases (‘on the other hand’, ‘it is clear that’, ‘at the end of’, & ‘the fact that’) were examined. In the case of the modal adjuncts/adverbs, the following 24 adverbs (including a phrase) classified into four categories according to the strength of certainty, were sampled for examination since they occurred at least once in any of the three corpora (D. Suzuki, 2011; Huddleston et al., 2002).

(a) no doubt, undoubtedly, certainly, surely, definitely, clearly, necessarily, obviously, plainly, truly, unquestionably, assuredly

(b) presumably, doubtless, seemingly, apparently, evidently

(c) arguably, likely, probably

(d) conceivably, maybe, perhaps, possibly

Results and Discussion

Regarding the first and second research questions about the transitions, the analyses of the students’ corpora compared to MICUSP revealed the following two remarkable results. Firstly, as for the frequency, basic common transitional words like ‘because’, ‘however’, ‘therefore’, and ‘then’ were overused more by the lower grade papers corpus, JC-GP_B, as indicated by the bar graph in Figure 2.

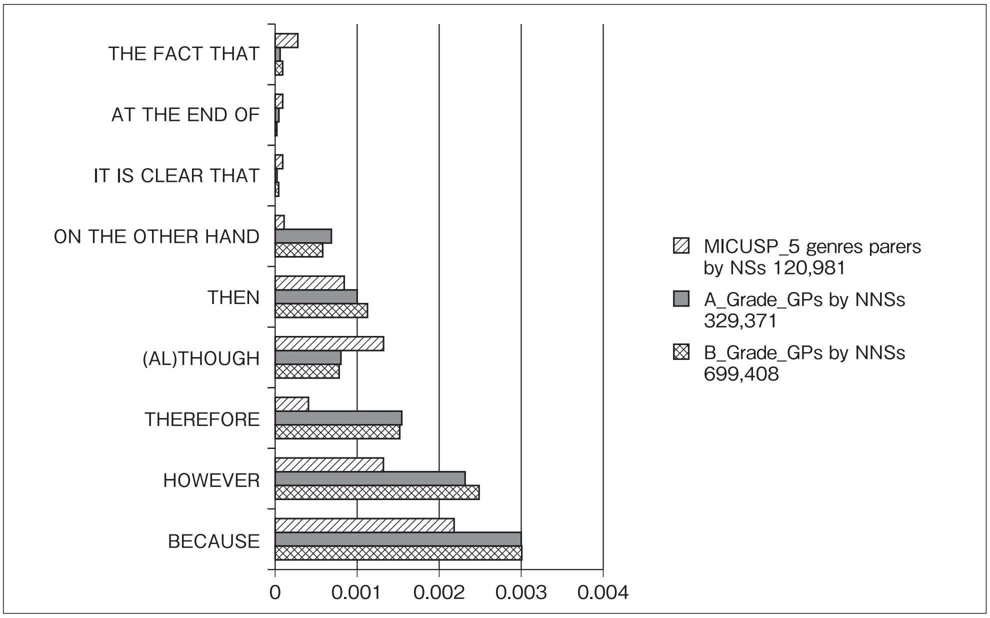

Figure 2. Comparative analysis results of transitions frequency rate (%).

Secondly, the students tend to use a transitional word or phrase exclusively in a certain fixed location such as in the initial position of a sentence, somewhere in the middle of the sentence, or in the final position, while the native speaker students in MICUSP use it in more varied positions. For example, ‘therefore’ is used by the students as “Therefore, it is necessary to make clearer the overall vision and system of bilingual education to assist the Hispanics in America.” and “Therefore, expansion of employment and earning opportunities are required to promote economic growth.”; whereas it is used by the native speakers as “The grammar rules do not come naturally to the students yet and therefore they commit the mistakes again in their informal writing.”, “Something may be relevant, in which case, *may* is expressing epistemic uncertainty on the part of the speaker and is therefore a hedging device.” or “Therefore, poverty *is* necessary for American society to work, just as Wright argues.” The results showing this kind of tendency in colligation patterns by the students and the native speakers are demonstrated by the graphs in Figure 3 below.

To address the third research question concerning the modal adjuncts/adverbs, the students’ peculiarity in terms of their colligation was observed again as seen in the following three graphs of average ratio, sampling two categories of adjuncts/adverbs, (a) and (c), with the total ratio of all the adjuncts/adverbs. The students clearly tend to use a modal adjunct/adverb in the sentence’s initial

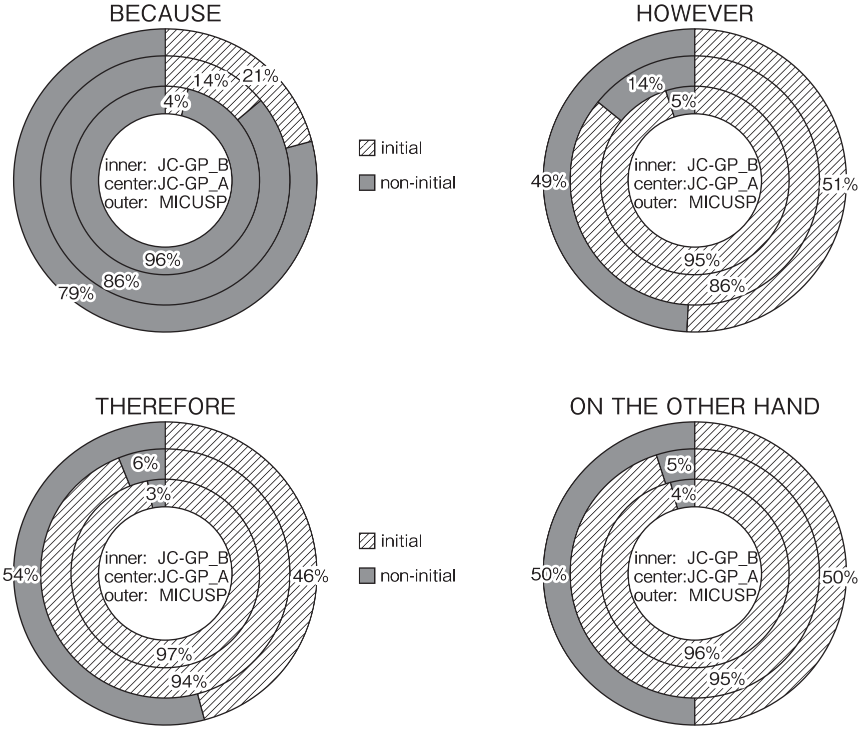

Figure 3. Comparative analysis results of transitions' colligation[1] patterns.

position more frequently than the native speaker students; in contrast, the native speakers in MICUSP use modal adjuncts/adverbs mostly in the medial position. (See Figure 4.)

The analysis results of modal adjuncts/adverbs, in answer to the fourth research question as to whether or not any modal adjunct/adverb collocates with an auxiliary in the same sentence (e.g. '... will undoubtedly...', '...would surely...', and '...can possibly...'), also showed differences in the average ratio between the students corpora and the native corpus: JC-GP_B & JC-GP_A versus MICUSP. To sum up, the students tend to use modal adjuncts/adverbs insufficiently by collocating them with modal auxiliaries as often as the native speaker students.

Furthermore, throughout all of the above-mentioned results, it was prominent that the students' peculiar tendencies in the overuse/underuse of transitions, colligation patterns, and collocation occurrences are more obvious in the lower grade papers corpus, JC-GP_B, than in the higher grade corpus, JC-GP_A, compared to the native speaker students corpus, MICUSP. This may reflect the students' process of development in attaining discourse competence.

Therefore, in addition to the fact that sentence length and complexity of vocabulary have been treated as the indicator of writing ability, the analysis results show that a diversity of colligation patterns overall indicates a higher writing ability as well.

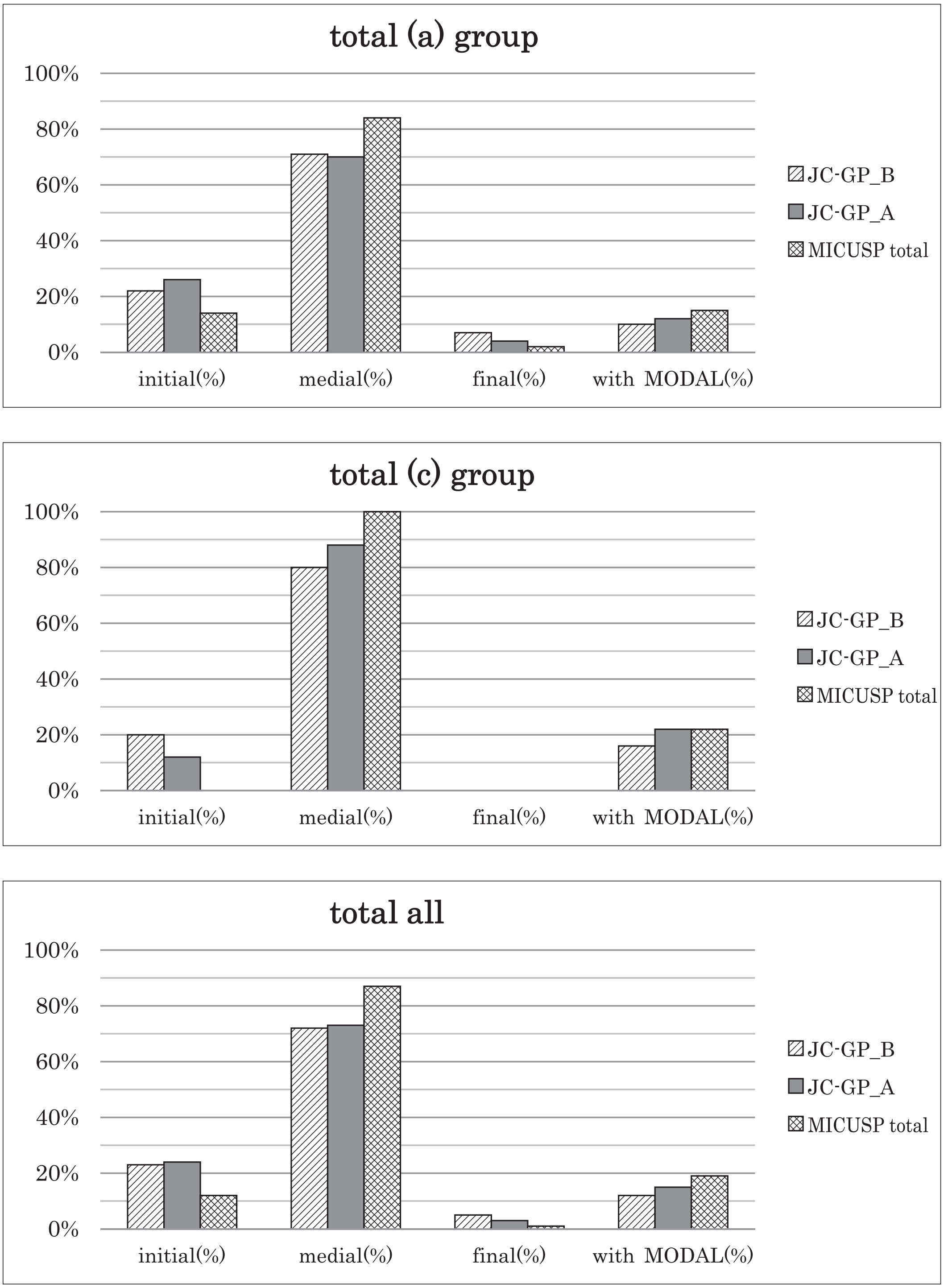

Figure 4. Comparative analysis results of modal adjuncts/adverbs’ colligation patterns and collocation with modal auxiliaries.

Pedagogical Implications and Developing the Online Tutorial System

As for the ultimate goal of the authors' project, that is to complete a web-based study-aid system, the research results as described above should be successfully applied to the tutorial part of the system. In the first place, proper example sentences should be carefully composed based on the data of the three corpora for the students to learn how the transitional words or phrases and the modal adjuncts/adverbs researched can be used. Next, the materials must be arranged in an effective way taking into consideration the students' developmental process of discourse competence as suggested by the analyses' results. Finally, the contents will be loaded into the system by using the editorial mode, which has already been installed, to supplement the already-installed tutorial contents for teaching grammatical forms. Then, these materials in the tutorial page plus the sentence search concordance system, which has already been developed, (Fukushima et al., 2010) and other linked out-resources on the system, which include the online dictionary Weblio[2], and a native speakers' English corpus reference site NativeChecker[3], as well as the AWL Exercise Page by Averil Coxhead, are certainly expected to enhance the students' discourse and rhetorical competence regarding cohesion. Further research based on the study results must involve an empirical study, which is aimed at assessing the effect of the independent e-learning system, which will be completed with these materials and equipment.

References

Chen, C. (2006). The use of conjunctive adverbials in the academic papers of advanced Taiwanese EFL learners. *International Journal of Corpus Linguistics, 11*(1), 113-130.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Coxhead, A. (2011). The Academic Word List ten years on: Research and teaching implications. *TESOL Quarterly, 45*(2): 355-362.

Field, Y., & Yip, L. (1992). A comparison of internal conjunctive cohesion in the English essay writing of Cantonese speakers and native speakers of English. *RELC Journal, 23*(1), 15-28.

Fukushima, S., Kinjo, Y., Suzuki, C., Watanabe, Y., & Yoshihara, S. (2010). Development of a web-based concordance system based on a corpus of English papers written by Japanese university students, *EUROCALL 2010 Abstracts,* 182-183.

Granger, S., & Tyson, S. (1996). Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes, 15*(1), 17-27.

Huddleston, R., & Pullum, G. K. (2002). *The Cambridge grammar of the English language.* Cambridge: Cambridge University Press.

Kinjo, Y., Suzuki, C., Fukushima, S., Watanabe, Y., & Yoshihara, S. (2010). Eigo ronbun sakusei shien wo mokutekitoshita Nihonjin daigakusei no gakushusha koopasu kouchiku (Building up a corpus with a view to aiding the students with writing English graduation papers). *NLP 2010 Conference Proceedings* by The Association for Natural Language Processing, 876-879.

Liu, D. (2008). Linking adverbials: An across-register corpus study and its implications. *International Journal of Corpus Linguistics, 13*(4), 491-518.

Michigan Corpus of Upper-level Student Papers. (2009). Ann Arbor, MI: The Regents of the University of Michigan.

Milton, J., & Tsang, E. (1993). A corpus-based study of logical connectors in EFL students' writing: Directions for future research. In R. Pemberton & E. S. C. Tsang (Eds.), *Studies in Lexis* (pp.215-246). Hong Kong: The Hong Kong University of Science and Technology.

Paquot, M. (2010). *Academic vocabulary in learner writing: From extraction to analysis.* London: Continuum.

Suzuki, C., Fukushima, S., Watanabe, Y., Kinjo, Y., & Yoshihara, S. (2009). Compiling a corpus of English graduation papers written by Japanese university students to develop an independent learning system, in *The 7th Asia TEFL International Conference Abstracts,* 254-255.

Suzuki, C., Fukushima, S., Watanabe, Y., Kinjo, Y., & Yoshihara, S. (2011). A study of textual colligation of transitional words in corpora of academic papers written by NS/NNS of English, in *12th International Pragmatics Conference Abstracts,* 534.

Suzuki, D. (2011). Eigo hou-fukushi no tayosei ni tsuite: Goyouronnteki hensuu to toukeiteki shuhou (On the diversity of modal adjuncts/adverbs: Their pragmatics variables and statistical methods) handout at JAECS (Japan Association for English Corpus Studies) East Chapter Symposium, 09/07/2011, Keio University, Tokyo.

Römer, U., & O'Donnell, M. B. (2009). Exploring the variation and distribution of academic phrase -frames in MICUSP in M. Mahlberg, V. González-Díaz, & C. Smith (Eds.) *Proceedings of the Corpus Linguistics Conference CL 2009,* URL: http://ucrel.lancs.ac.uk/publications/cl 2009/

West, M. (1953). *A general service list of English words.* London: Longman, Green and Co.

Footnotes

1) Textual colligation is defined as "words and phrases (which) may carry with them particular associations for occurrence at a specific location in text, i.e. beginning or end of a text, paragraph, or sentence" (Römer & O'Donnell, 2009).

2) Weblio, which is a free site for public use, provides a thesaurus, a variety of Japanese/English & English/Japanese dictionaries for different specific purposes, a collection of example sentences,

and the KWIC display of any target word, along with related information of vocabulary and language.

3) NativeChecker was developed by software engineer Kaisei Hamamoto (http://kaiseh.com/); the reference site had offered frequency counts of any word or phrase from collected corpus data by the developer until March 2013, when Yahoo! Japan changed terms of service for using its Web Search API.

4) This paper was originally presented as an oral presentation at the 10th International AELFE (European Association of Languages for Specific Purposes) Conference on September 7th, 2011 at Valencia, Spain.

Acknowledgements

JSPS (Japan Society for Promotion of Science) KAKENHI supported this study (Grant-in-Aid for Scientific Research: Grant No.21520621). Dr. Averil Coxhead, with her expertise as a research cooperator, also constantly consulted with the authors about corpus studies and TEFL. In addition, the authors sincerely appreciate Dr. Ute Römer for agreeing to allow them to use the corpus data of MICUSP.