Feature Article

A Study of L2 Spoken Corpora: The Gap between Productive Skills of Japanese EFL Learners and the Corpus-Based ELT Textbooks

Aika Miura
Tokyo Keizai University

Abstract

This study aims to observe the gap between productive skills of particular pragmatic features of Japanese EFL learners and the expectations and decisions the authors of corpus-based ELT textbook series make regarding the types of language to include. So-called "conversational strategies" in the Touchstone Series from Cambridge University Press with corpus-based syllabus design are examined in the NICT JLE Corpus, which contains more than one million words of 1,200 Japanese EFL learners with nine different proficiency levels. The research suggests that Japanese learners do not necessarily exhibit their productive skills as the way the textbook series presents target language and strategies. The learners at various proficiency levels performed relatively differently from each other. The intermediate learners tended to overuse particular words and phrases compared to those who were at upper intermediate and advanced proficiency levels. Research into the NICT JLE Corpus suggests a transitional nature of L2 development which is particular to Japanese EFL learners, and the results can be used to complement ELT textbooks and improve language instruction in Japan, specifically at the tertiary level.

Aika Miura currently teaches at Tokyo Keizai University, Toho University and Tokyo University of Science. She holds a B.A. (Linguistics) from Reading University and an M.A. (Language Studies) from Lancaster University in the UK. Her interests include second language acquisition and corpus linguistics. She can be contacted at dawn1110am@yahoo.co.jp.

Miura, A. (2009). A study of L2 spoken corpora: The gap between productive skills of Japanese EFL learners and the corpus-based ELT textbooks. *OnCUE Journal, 3*(2), 136-159. Copyright © 2009 Aika Miura. *OnCUE Journal*, ISSN 1882-0220, is available at http://jaltcue-sig.org one year after publication date.

本研究は、日本人英語学習者の話し言葉コーパスを用い、あるコーパス準拠のシラバスから成る英語教材で取り上げられる談話パターンの使用実態の習得段階の全容を調査し、英語教材で期待される学習者の言語運用と日本人英語学習者に見られる発話能力の差異を検証したものである。NICT JLE Corpusを用い、ケンブリッジ大学出版局のTouchstoneシリーズが提示する会話ストラテジーの使用実態を調査した。本コーパスは、1200人以上の学習者の100万語の話し言葉から構成され、9つの習得段階に分別されている。調査結果より、中級学習者による特定のストラテジーの過大使用など、習得段階ごとに異なる言語運用の傾向が示され、分析対象となった表現の使用頻度は必ずしも発達推移と一致しなかった。習得別の学習者コーパスを横断的に観察することで、日本人英語学習者に特有な発達指標を見出し、ネイティブコーパスに準拠した英語教材の補完や教育現場の示唆に有用な結果が得られたと言える。

SLA and Learner Corpora

In order to be an effective language teacher, it is advantageous to have considerable background regarding the study of second language acquisition as well as pedagogy related to L2 learners. In support of this Ellis (1994) states, “The study of SLA provides a body of knowledge which teachers can use to evaluate their own pedagogic practices” (p. 4). There have been a great number of different approaches to describe learner language. Ellis (1994) identified four major approaches; the study of learners’ errors, the study of developmental patterns, the study of variability, and the study of pragmatic features. The description of learner language enables language teachers to improve syllabus design, material design, task design and testing (Tono, 2000). Thus, when describing the study of developmental patterns, Ellis (1994) highlighted the importance of “unplanned language use” which is found in naturalistic settings (p. 82). This term corresponds to what Ellis (1994) called “natural language use,” which “occurs in the course of using the L2 in the kind of communication that learners engage in when they are not being studied” (p. 672).

Generally, the focus of the SLA researcher is to collect data that reflect learners’ attempts to use the L2 in either comprehension or production. However, Granger (2002) pointed out that much current SLA “tends to be dismissive of natural language use data” for it is

difficult to control the variables that affect learner output in a non-experimental context, and to subject a large number of informants to experimentation (p. 5-6). Therefore, questions have been raised whether the results of the studies on a small number of samples of learner data can be generalized.

Recently, researchers using computer based technologies are able to collect and store a large amount of learner data, and to analyze it automatically or semi-automatically using currently available linguistic software and tools (Granger, 2002).This advancement makes it possible to produce research results from a larger number of participants than previously possible, which leads to higher validity and reliability regarding statistical data. Therefore, situational specific data can be expanded and researchers can afford to make more generalized conclusions regarding the results. Granger (2002) refers to this large amount of learner data as "computer learner corpora" (CLC)(p. 4).

The research on CLC stems from recent advances in the field of corpus linguistics, specifically within the past decade CLC has had an extensive influence on language teaching regarding both second language acquisition as well as pedagogy (Granger, 2002; Ishikawa, 2008). Learner corpora of various kinds such as speakers of different mother tongues, proficiency levels, age, modes (spoken or written), and registers (academic or non-academic) have been developed and improved on. As for Japanese EFL learners and language researchers, learner corpora are available online, such as the Japanese English as a Foreign Language Learner (JEFLL) Corpus (Tono, 2007) and the Nagoya Interlanguage Corpus of English (NICE) (Sugiura, 2009). Web sites that support learner corpora data have unarguably supplied researchers and language teachers with an abundance of learner data that can be used for studies ranging from the smaller and more situational-specific classroom action research types to larger scale studies.

Access to a corpus enables researchers to investigate the learners' errors, developmental patterns, variability and pragmatic features.

Hunston (2002) further commented that "corpora lead to new descriptions of a language, so that the content of what the language teacher is teaching, is perceived to change in radical ways" (p. 137). Granger (2002) also suggested that "teachers and researchers often have useful intuitions about what does or does not constitute an area of difficulty for learners, but this intuition needs to be borne out by empirical data from learner corpora" (p. 23).

Corpora and its Application to ELT

Based on the literature, corpora can be defined as a principled collection of texts, written or spoken, stored on a computer (Biber, Conrad, & Reppen, 1998; Granger, 2002; McCarthy, 2004; O'Keeffe, McCarthy, & Carter, 2007). From this, researchers can study naturally occurring language use. According to Granger (2002), studies on corpora specifically highlight a quantitative approach as researchers are able to discern what is most likely to occur in language, or "frequency", which rules out the intuitive awareness factor in the collection of data and the preceding analysis of it. In this way, corpus linguistics has contributed to a number of areas in language pedagogy including syllabus design and materials development (Granger, 2002; Hunston, 2002; Reppen & Simpson, 2002; McEnery, Xiao, & Tono, 2006).

Kennedy (1998) summarizes the importance of syllabus and materials design by stating corpora supply "information on the distribution of the elements and processes of a language", and "it can influence the content of language teaching by affecting selection of what to teach, the sequencing of pedagogy, and the weight given to items or parts of the language being taught, thus contributing directly to the content of instructions" (p. 281). After comparing "internationally successful coursebooks" to the Cambridge and Nottingham Corpus of Discourse in English (CANCODE), Carter (1998) discovered that

dialogues taken from selected non-corpus-based textbooks lacked core spoken language features such as discourse markers, vague language, ellipsis and hedges.

Native vs. Non-Native Corpora

Several studies (Cook, 1998; Widdowson, 2000) have leveled criticisms against a corpus-based approach which tends to heavily rely on the aspect of "frequency" and "authenticity" when applied to language teaching (Hunston, 2002; Tono, 2008). Cook (1998) argued that "even within the native-speaker community it is often the infrequent word or expression which is most powerful and most communicatively effective" (p. 6). Therefore, "frequency information obtained from native-speaker corpora alone is not sufficient to inform curriculum and material design" (McEnery et al., 2006, p. 101).

In this respect, the research into CLC plays an important role in applying the data from corpora studies to language pedagogy (McEnery et al., 2006; O'Keeffe et al., 2007). Native corpora "give no indication whatsoever of the difficulty they present for learners in general or for a specific category of learners" (Granger, 2002, p. 21). The NS (native speakers) / NNS (non-native speakers) comparison enables us to notice gaps between learners' interlanguage and the language they are learning (McEnery et al., 2006) This is apparent in NNS learners' "overuse" and "underuse" of selected items, as well as "NL (native language) transfer" and "avoidance strategies", when compared to the NS corpora (Granger, 2002; Leech, 1998). As for objections to this type of comparison regarding the placing the importance on the native "norm" in language teaching see Widdowson (2000), Cook (1998), Granger (2002,), and Hunston (2002).

Corpora and Textbook Authoring

Apart from the aforementioned controversial issues related to the use of corpora in language teaching, material development and syllabus

design continue to be based on corpora. The importance of corpora is evident in the compilation of learners' dictionaries, which are now based on continually updated multi-million word databases of language, such as the Longman Corpus and the Cambridge International Corpus (O'Keeffe et al., 2007; Tono, Izumi, & Kaneko, 2005). Aside from the recent development of corpus based dictionaries, text book authors and publishers are also utilizing corpora in material development.

The *Touchstone Series* is one of the more recently published ELT textbooks whose dialogues and entire syllabi are informed by corpus data (McCarthy, 2004; McCarthy, McCarten, & Sandiford, 2005a, 2005b, 2005c; McCarthy et al., 2006a, 2006b; O'Keefe et al., 2007). The series is "based on the most common words and phrases in the North American segment of Cambridge International Corpus (CIC)" which contains 700 million words of text (O'Keeffe et al., 2007, p.22). According to McCarthy (2004), the authors interpret and mediate their corpus research by looking at word lists from the Corpus and searching for the most frequent and typical uses in everyday conversation. Then they decide which words are the most important ones to be included, considering whether they suit the interests of learners and are most relevant to their lives or not. For pedagogical considerations, it should be noted that "corpus data alone does not dictate an instructional syllabus" (McCarthy, 2004, p. 15), since high-frequency language is not always appropriate for beginners to start to learn (Ishikawa, 2008; McEnery et al., 2006; Tono, 2008).

Learner Corpus Data and its Proficiency Levels

The National Institute of Information and Communication Technology (NICT) developed the NICT JLE Corpus, a learner corpus of Japanese Learners of English (JLE), which contains more than one million words from approximately 1,200 Japanese subjects' 15-minute interviews (Izumi, Uchimoto, & Isahara, 2004). The data collected from

the interviews comes from the Standard Speaking Test (SST) created by ALC Press. The SST is based on the Oral Proficiency Interview of the American Council on the Teaching of Foreign Languages. Table 1 shows the distribution of subjects' proficiency levels, types (the total number of different word forms in each level), and tokens (the total occurrences of any given word forms in each level). Those who fall into levels 1 to 3 can be described as novice learners and those who fit levels 4 to 8 are considered intermediate ones, and finally, test takers with the highest scores are classified as the most advanced at Level 9 (see Izumi et al., 2004).

Table 1

Distribution of Subjects, Types, and Token of the NICT JLE Corpus

Level	1	2	3	4	5	6	7	8	9
Subjects	3	35	222	482	236	130	77	56	40
Types	217	1516	6025	10120	8290	6867	5455	4981	4429
Token	1440	20788	211625	606951	365330	219646	139534	112185	85420

Conversational Strategies and the Touchstone Series

The primary aim of this study is to investigate the Japanese learners' use of pragmatic features based on the detailed analyses of the native speakers' corpus. This should be distinguished from research into "discourse markers" based on pragmatic theories and "formulaic expressions" (De Cock, Granger, Leech, & McEnery, 1998; Fraser, 1993).

O'Keeffe et al. (2007) and McCarthy (2008) defined conversational strategies as *chunks*, which are automatically extracted multi-word strings from the corpus that display pragmatic functions such as hedging, vagueness, and discourse marking, the preservation of face and the expression of politeness, rather than belonging to syntactic and semantic categories (McCarthy & Carter, 2006; O'Keeffe et al., 2007). The presence or absence of common chunks has been considered a

useful measure of comparison and evaluation of learner competence compared to native speaker competence (McCarthy & Carter, 2006; O'Keeffe et al., 2007). De Cock et al. (1998) offer their definition of chunks as "formulaic expressions, i.e. frequently used multi-word units that perform pragmatic or discourse structuring functions" (De Cock et al., 1998, pp. 68-69).

The *Touchstone Series* is composed of *Book 1* for beginners, *Book 2* for high beginners, *Book 3* for low-intermediate, and Book 4 for intermediate learners. Each book consists of 12 units, which includes a section devoted to conversational strategies presented in a dialogue followed by grammatical explanations, language practice exercises and listening comprehension quizzes. The following sample dialogue is taken from *Touchstone Book 2* (McCarthy et al., 2005b, p. 71) and focuses on the target language form I guess, which is used "when you're not 100% sure about something, or if you don't want to sound 100% sure":

> Chris: You know, we should take a few days off sometime.
> Adam: Yeah, we should. Definitely.
> Chris: We could go to Mexico or something.
> Adam: That's a great idea.
> Chris: We could even go for couple of weeks.
> Adam: Well, maybe. *I guess* we could, but…
> Chris: You know, we could just quit our jobs and maybe go backpacking for a few months…..
> Adam: Well, I don't know. I'd like to, but….. I guess I need to keep this job, you know, to pay for school and stuff.
> Chris: Yeah, me too, *I guess*.

In the data analysis, the list of conversational strategies was initially made from four books of the series (totaling 81 items). The functions of the strategies are defined in each book, including the inclusion of corpus-based information, i.e., "I guess is one of the top 20 expressions

in conversation" (McCarthy et al., 2005b, p. 71). Table 2 shows the items which are specifically discussed in the paper.

Table 2

Description of conversation strategies in the Touchstone Series Books 1, 2, 3, and 4 *(McCarthy et al., 2005a, 2005b; McCarthy et al., 2006a, 2006b)*

Language	Book	Unit (page)	Function of strategy	Corpus information
(1) I mean	1	5 (p. 48-9)	Repeating your ideas or saying more about something	"I mean" is one of the top 15 expressions.
	2	5 (p. 48-9)	Correcting yourself when you say the wrong word or name	"Mean" is one of the top100 words. About 90% of its uses are in the expression of "I mean."
(2) I guess	2	7 (p. 71)	When you're not 100 % sure about something, or if you don't want to sound 100 % sure	"I guess" is one of top 20 expressions
	3	7 (p. 70)	Softening comments	———
(3) of course, absolutely, definitely, really*, actually*, certainly, honestly, in fact, to be honest	4	3 (p. 26-7)	Sounding more direct	"Of course" is one of top 50 expressions

Table 2 (continued)

(4) I guess/I think, probably/ maybe*, sort of/kind of*, a little (bit), (just)	3	7 (p. 70-1)	Softening comments	———
(5) just*	4	6 (p. 58-9)	Making your meaning clearer	"Just" is one of the top 30 words. Over half of its uses are to make ideas stronger or softer.
(6) like	4	8 (p. 81)	Various functions (See the results and discussion section.)	"Like" is one of the top 15 words. It is about 6 times more frequent in conversation than in writing.
(7) so	4	4 (p. 39)	Various functions (See the results and discussion section.)	Checking your understanding

Note: Empty cells signify corpus information was not provided regarding target language in the table. The use of items marked by (*) is further discussed in the results section.

The principle data regarding the occurrence of each item was searched utilizing the NICT JLE Corpus. Figure 1 shows an extract of 10 concordance lines for I guess in a sub-corpus of Level 8 from the NICT JLE Corpus. It should be noted that most types of corpora generally have somewhat of an original annotation scheme, which is also true in the case of the NICT JLE Corpus as #F# means filler or filled pause, #SC# means self correction, "R" means repetition and last #JP# means Japanese (see Izumi et al., 2004).

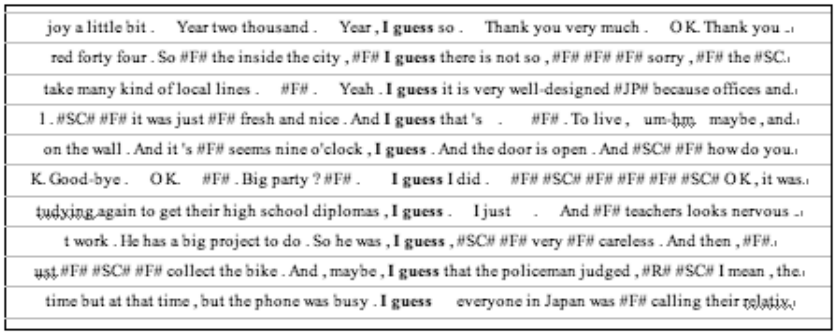
joy a little bit . Year two thousand . Year , **I guess** so . Thank you very much . O K. Thank you .
red forty four . So #F# the inside the city , #F# **I guess** there is not so , #F# #F# #F# sorry , #F# the #SC.
take many kind of local lines . #F# . Yeah . **I guess** it is very well-designed #JP# because offices and.
1 . #SC# #F# it was just #F# fresh and nice . And **I guess** that 's . #F# . To live , um-hm, maybe , and.
on the wall . And it 's #F# seems nine o'clock , **I guess** . And the door is open . And #SC# #F# how do you.
K. Good-bye . O K. #F# . Big party ? #F# . **I guess** I did . #F# #SC# #F# #F# #F# #SC# O K , it was.
tudying again to get their high school diplomas , **I guess** . I just . And #F# teachers looks nervous .
t work . He has a big project to do . So he was , **I guess** , #SC# #F# very #F# careless . And then , #F#.
ast #F# #SC# #F# collect the bike . And , maybe , **I guess** that the policeman judged , #R# #SC# I mean , the.
time but at that time , but the phone was busy . **I guess** everyone in Japan was #F# calling their relativ.

Figure 1. Sample concordance lines for *I guess* in Level 8 of the NICT JLE Corpus.

Results

The distribution of *I mean* and *I guess* in the NICT JLE Corpus is shown in Figure 2. Since tokens in each proficiency level were different as indicated in Table 1, the frequency of analyzed words or phrases was normalized at 100,000 word tokens. A sample taken from the NICT JLE Corpus shows that a student of Level 7 produced the following utterance: "In my hometown, *I mean* I don't have a hometown right now since my parents moved to another town." The student correctly uses the term *I mean* as a conversational strategy to clarify meaning.

According to Figure 2, *I mean* and *I guess* were especially underused by novice and lower intermediate learners of Level 1 to 5, compared to upper intermediate and advanced learners. The frequency of *I mean* showed a sudden increase at Level 6. On the other hand, the use of *I guess* suddenly increased at Level 8 while the frequency of *I mean* in Level 9 dropped from 90 to 60, which may suggest that the degree of self-correcting by advanced learners decreases as they improved their proficiency with increased vocabulary.

Figure 3 shows the frequency of strategies whose function is to sound more direct, in contrast to the strategies that soften comments. In the analysis just was excluded from the "softening" groups as it also has the function of making the utterance stronger. According to

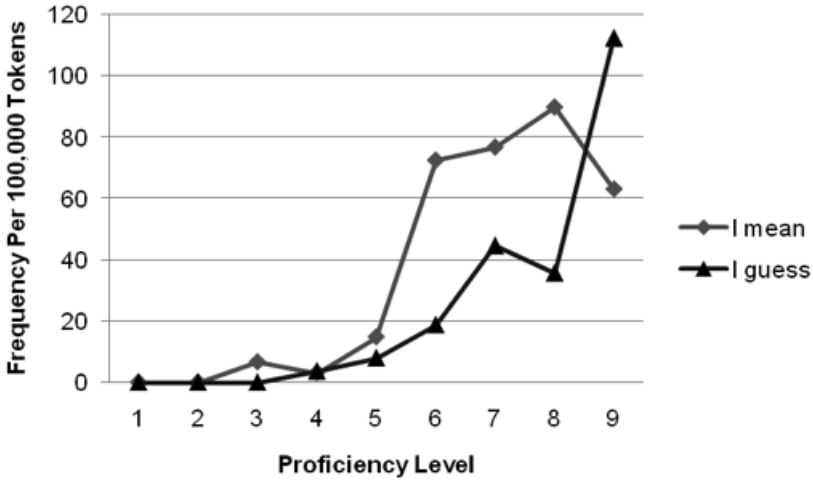

Figure 2. Distribution of the use of *I guess* and *I mean* in the NICT JLE Corpus.

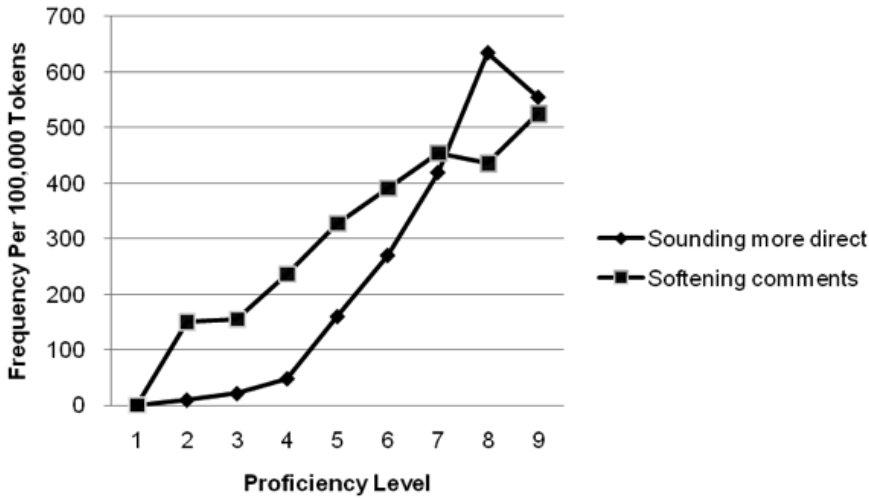

Figure 3. Distribution of direct and softening strategies in the NICT JLE Corpus.

my analysis, there was an increase in the use of strategies regarding softening comments and when sounding more direct as student proficiency levels improved. Novice and lower intermediate learners tended to use softening comment strategies more frequently than direct

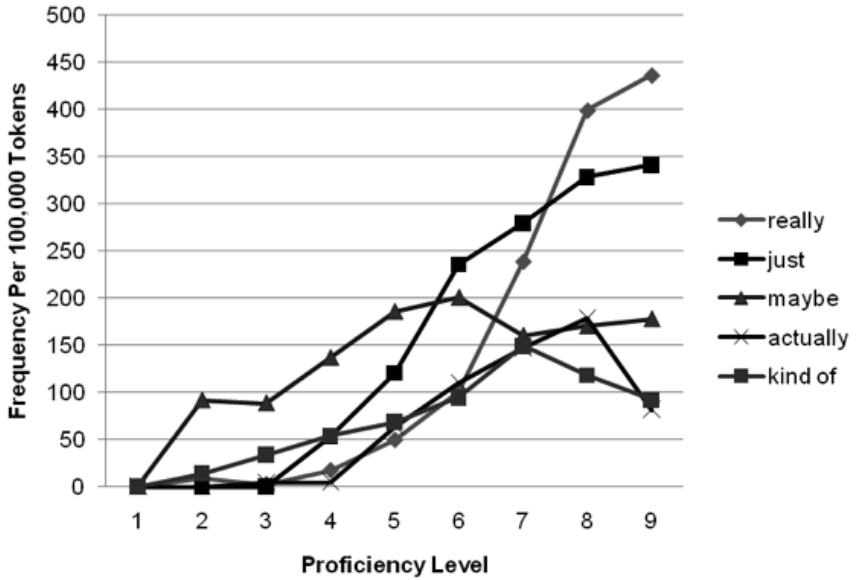

Figure 4. Distribution of the use of *really, just, maybe, actually* and *kind of* in the NICT JLE Corpus.

ones.

As shown in Table 2, the *Touchstone Series* introduces nine expressions for "sounding more direct" (*of course, absolutely, definitely, really, actually, certainly, honestly, in fact, to be honest*), and eight for "softening comments" (*I guess, I think, probably, maybe, sort of, kind of, a little (bit), just*). Except for the following five expressions *really, actually, maybe, kind of* and *just*, the other nine expressions for "sounding direct" and the remaining eight expressions used for "softening comments" were not analyzed for this study due to their infrequent use in the NICT JLE Corpus. Figure 4 presents the use of these five frequently used phrases, and clearly shows an increase specifically from Level 5 onwards. The highest increase occurs with the two phrases *really* and *just*, respectively. There is also a sharp increase for both phrases between Levels 6 and 8, which may either indicate overuse of the terms or an increased level of fluency. On the other hand, there was a decrease in the frequency of the phrase actually,

kind of and maybe as the proficiency levels increased.

Finally, regarding *like* and *so*, Japanese learners showed completely different tendencies from how they are actually introduced in the *Touchstone Series*. In Table 3, conversational analysis of dialogues extracted from Book 4 reveals six different functions of *like* and the percent of how each function is used by native speakers in the far right column.

Table 3

Six Different Usages of "Like" introduced in Touchstone Series *(McCarthy et al., 2006b)*

Type	Function	Example	Percent
(1)	To say something is similar	He acted *like* we were in his way.	34%
(2)	To highlight something	They were *like* totally blocking the doors.	18%
(3)	To mean other things, including the verb "like"	I *like* to ski.	17%
(4)	To give an example	*Like*, I get upset……	16%
(5)	To report what someone said	They were *like*, "What's your problems?"	10%
(6)	To say "approximately"	Isn't he *like* 80 years old?	5%

Figure 5 presents the distribution of the use of *like* among Japanese learners (see Appendix A for sample data taken from the NICT JLE Corpus). According to the results, *like* used as a verb was most frequent among Japanese learners, peaking at 500 tokens at Level 9, compared to native speakers who used *like* as a verb at only 17% (see table 3). There is quite a discrepancy between the two groups which becomes evident due to the data revealed from corpora studies.

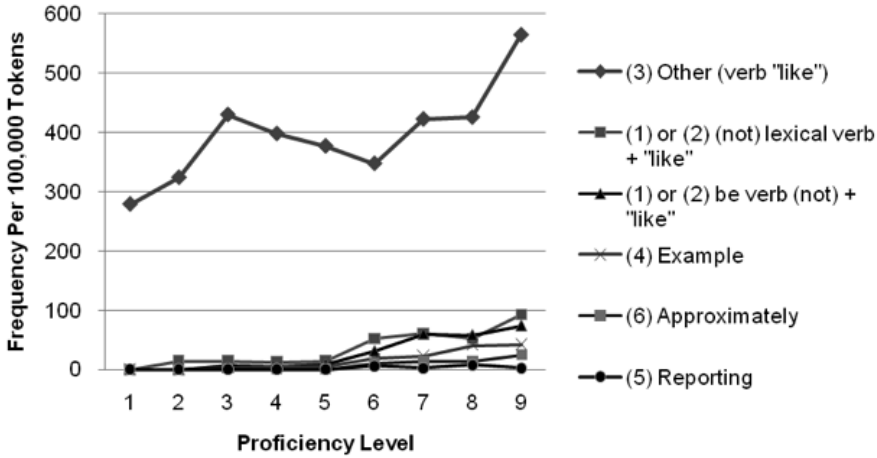

Figure 5. Distribution of the use of *like* in the NICT JLE Corpus.

Table 4 shows four different uses of so in *Touchstone Series* (McCarthy et al., 2006b, p.38).

Table 4

Four Different Usages of "So" introduced in Touchstone Series (McCarthy et al., 2006b)

Type	Function	Example
(1)	To start a topic, often with a question	*So* it's your birthday Friday, right?
(2)	To check your understanding	*So* they all came, huh?
(3)	To pause or let the other person draw a conclusion	They all came, *so*……
(4)	To close a topic	*So* that's what happened. They all came.

Figure 6 presents the distribution of the use of *so* among Japanese learners (see Appendix B for sample data taken from the NICT JLE Corpus). There is a sharp increase in frequency between Level 1 and

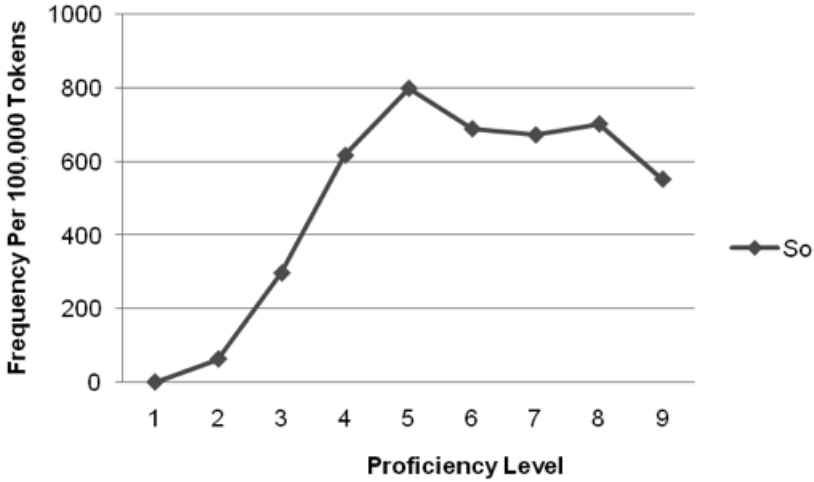

Figure 6. Distribution of the use of *so* in the NICT JLE Corpus.

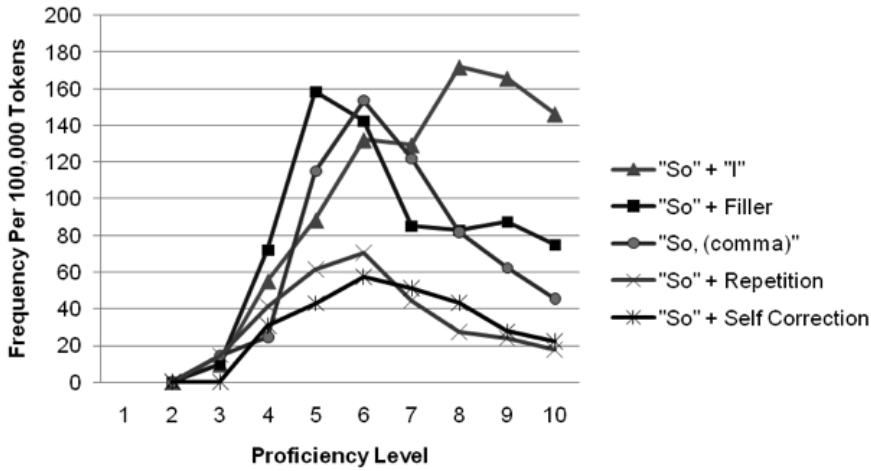

Figure 7. Distribution of the use of the items following after so in the NICT JLE Corpus.

Level 5 followed by a steady decline. The highest frequency at Level 5 suggests the overuse by lower intermediate learners, as compared to upper intermediate and advanced learners.

The unexpected finding with the high frequency counts in the level of lower intermediate learners led to further analysis on the

different contexts in which so is used. Particularly, the linguistic item following so was measured. Figure 7 shows that "I", "Filler", "Self-Correction", "Comma" (which may indicate a certain length of pause), and "Repetition" are the top five items following so.

Discussion

First, an interesting result generated from the data was that frequency counts of the strategies such as *I mean, actually, kind of* and *maybe* were comparatively higher in the level of intermediate learners than that of upper intermediate and advanced learners. Intermediate learners possibly overused these words as fillers because of their lack of vocabulary and productive skills. Based on observations of my learners, the overuse of these strategies interferes with their conversational fluency. From a pedagogical perspective, students, specifically intermediate learners, should be made aware that the overuse of the term may hinder communication.

Secondly, the analysis of direct and softening strategies suggests that novice and lower intermediate learners show higher frequency of using "softening comments" than "sounding more direct". This may be due to their L1 background, where the Japanese language tends to be less direct. However, Kaneko's (2004) research into the development of speech acts in the NICT JLE Corpus revealed that intermediate learners were the most direct when making requests, which is contrary to the data collected for this study.

Some studies utilize discourse completion tasks as a method of data collection. Cole and Anderson's (2001) longitudinal study showed their students used language that indicated an increase in more indirect conversational strategies after a ten-month homestay. Takahashi's (1996) study into the transferability of Japanese indirect request strategies to corresponding English request contexts suggested that the learners' pragmatic competence was not native speaker like,

regardless of their L2 proficiency. However, the results of the previously mentioned studies have varied and are inclusive regarding direct and indirect conversational strategies. One reason for this might be due to the small number of participants in those studies.

Following that, the distribution of various usages of *like* indicates that Japanese students rarely used *like* other than as a verb. This may be because it is common that Japanese learners have been introduced only to the usage of *like* as a verb in their secondary and tertiary educations, compared to more abstract uses such as when *like* is used to approximate. Although the frequency slightly increases as the proficiency level goes up, there is still a wide gap between the more concrete or easier uses of *like* than what might be perceived as the more difficult uses of *like*. According to McCarthy et al (2006b), the native speakers use of *like* as types 2, 5, and 6 (see table 3) are used "in very informal conversations only" (p. 81)

However, one overwhelming appeal of the *Touchstone Series* is that the dialogues and instructional points in the series are based on empirical data taken from native speakers' corpus. The results in the corpus of Japanese learners suggest that they should produce *like* with more varieties, but language teachers may have to be careful regarding the use of *like* in appropriate contexts. Yet, the more informal uses of *like* based on native speakers' corpora (see table 3 sections 2, 5, and 6) may not be readily taught by language instructors, as some teachers tend not to encourage their students to use overly colloquial, informal and redundant language in their communication.

Finally, an interesting finding with the frequency counts of *so* and the analyses of its context should be "I" is the only lexical form, and others are hesitations or mumbles in the interviews. Level 4 and 5 show the highest frequency of the non-word items, and the occurrence decreases as the proficiency level increases. Lower proficiency level learners tend to seize speaking after uttering *so*, which can be related to the results of overusing *I mean, actually, maybe* and *a kind of*. From

the data, instructors should focus on the variety of different functions of like and so presented in *Touchstone Series* regarding pedagogical practices. In summary, the concern regarding colloquial usages in pedagogy may be realized in terms of the stance against heavy reliance on native speakers' corpora as claimed by Granger (2002). This might provide us with a sense of caution when placing the norm on native speakers as discussed earlier in the present study.

Conclusion

The results of this study show how learners of each acquisition stage in the NICT Corpus performed differently in terms of the language usage related to conversational strategies. One important discovery was that Japanese learners did not necessarily exhibit their productive skills as the way textbook authors interpreted the native speakers' corpus, or the North American segment of Cambridge International Corpus (CIC), which the *Touchstone Series* was based on. Another finding was that intermediate learners tended to overuse particular words and phrases, compared to those who were at upper intermediate and advanced levels. Thus, throughout the analysis, the features overused by intermediate learners seemed to function as fillers when they stopped speaking in order to search for what to say next.

Finally, the shortcomings with this study should be mentioned. There are items which show relatively low frequency or no occurrences, for example, *absolutely, definitely, honestly, in fact*, and *to be honest* as well as *Did you?* The last example, *Did you?* functions as an echo question to keep the conversation going (McCarthy et al. 2006a). Their absence or extremely low frequency from the NICT JLE Corpus may be due to the settings of the interviews, where students complete the assigned tasks so that they are assessed in terms of their oral proficiency. The SST interview, which the NICT JLE Corpus is based on, actually has five stages: (1) answering warm-up questions, (2) describing a single

picture, (3) having a role-play with the interlocutor, (4) narrating picture sequences, and (5) answering "wind-down questions." Therefore, the whole corpus can be divided into subcorpora depending on the stages. However, the present study does not go into a deeper analyses focusing on each stage and task of the SST. The warm-up or wind-down, and role-play stages may be the best situations to explore the language of conversational strategies where a student and the interlocutor interact equally, while in the other stages the interlocutor produces utterances only to encourage a student to speak out.

In conclusion, although the *Touchstone Series*, with its corpus-informed syllabi, is carefully designed to meet the needs of learners at different developmental levels, the research into the NICT JLE Corpus suggests that language teachers still need to accommodate the syllabi, target language, and activities based on their own students' needs. The study supports conclusions made by McCarthy (2004), McEnery et al. (2006), and O'Keeffe et al. (2007) regarding corpora and syllabus design, specifically when McCarthy (2004) himself noted "corpus data alone does not dictate an instructional syllabus" (p.15).

The results of the analyses also reinforces the notion that corpus-informed language pedagogy should not solely be based on native speakers' corpora, but should apply the extensive analyses of learner corpora as well. The application of learner corpora may even bridge the gap of claims made between corpus-based educationalists and their critics such as Cook (1998) and Widdowson (2000).

Acknowledgement

I would like to thank Dr. Yukio Tono at Tokyo University of Foreign Studies for access to the NICT JLE Corpus and his thoughtful comments on this research.

References

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Carter, R. (1998). Orders of reality: CANCODE, communication and culture. *ELT Journal, 52*, 43-56.

Cook, G. (1998). The uses of reality: A reply to Ronald Carter. *ELT Journal, 52*(1), 57-63.

Cole, S., & Anderson, A. (2001). Requests by young Japanese: A longitudinal study. *The Language Teacher, 25*(8), 7-11.

De Cock, S., Granger, S., Leech, G., & McEnery, T. (1998). An automated approach to phrasicon of EFL learners. In S. Granger (Ed.), *Learner English on computer* (pp. 67-79). London: Longman.

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University Press.

Fraser, B. (1993). Discourse markers across language. *Pragmatics and Language Learning, 4*, 1-16.

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

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3-18). London: Longman.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Ishikawa, S. (2008). *Eigo corpus to gengo kyouiku: data toshite no text.* [English corpora and language teaching: Texts as data]. Tokyo: Taishukan.

Izumi, E., Uchimoto, K., & Isahara, H. (Eds.). (2004). *Nihonjin 1200 nin no eigo speaking corpus* [L2 spoken corpus of 1200 Japanese learners of English]. Tokyo: ALC.

Kaneko, A. (2004). Nihonjin eigogakushuusha no yokyu no hatsuwa no hattatsu. [The developmental processes of making requests as speech act by Japanese EFL learners]. In E. Izumi, K. Uchimoto, & H. Isahara (Eds.), *Nihonjin 1200 nin no eigo speaking corpus* [L2 spoken corpus of 1200 Japanese learners of English] (pp. 113-

129). Tokyo: ALC.
Kennedy, G. (1998). *An introduction to corpus linguistics*. London: Longman.
Leech, G. (1998). Learner corpora: What they are and what can be done with them. In S. Granger (Ed.), *Learner English on computer* (pp. xiv-xx). London: Longman.
McCarthy, M. J. (2004). *Touchstone: From corpus to course book*. Cambridge: Cambridge University Press.
McCarthy, M. J. (2008). Profiling spoken fluency. *The Language Teacher, 32*(7), 32-33.
McCarthy, M. J., & Carter, R. A. (2006). This that and the other: Multi-word clusters in spoken English as visible patterns of interaction. In M. J. McCarthy (Ed.), *Explorations in corpus linguistics* (pp. 7-26). Cambridge: Cambridge University Press.
McCarthy, M. J., McCarten, J., & Sandiford, H. (2005a). *Touchstone. Student's book 1*. Cambridge: Cambridge University Press.
McCarthy, M. J., McCarten, J., & Sandiford, H. (2005b). *Touchstone. Student's book 2*. Cambridge: Cambridge University Press.
McCarthy, M. J., McCarten, J., & Sandiford, H. (2005c). *Touchstone. Teacher's edition 2*. Cambridge: Cambridge University Press.
McCarthy, M. J., McCarten, J., & Sandiford, H. (2006a). *Touchstone. Student's book 3*. Cambridge: Cambridge University Press.
McCarthy, M. J., McCarten, J., & Sandiford, H. (2006b). *Touchstone. Student's book 4*. Cambridge: Cambridge University Press.
McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London: Routledge.
O'Keeffe, A., McCarthy, M. J., & Carter, R. A. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.
Reppen, R., & Simpson, R. (2002). Corpus linguistics. In N. Schmitt (Ed.), *An introduction to applied linguistics* (pp. 92-111). London: Oxford University Press.
Sugiura, M. (2009). NICE: Nagoya Interlanguage Corpus of English. Retrieved June 6, 2009, from http://sugiura5.gsid.nagoya-u.ac.jp/~sakaue/nice/index.html
Takahashi, S. (1996). Pragmatic transferability. *Studies in Second Language Acquisition, 18*(2), 189-223.

Tono, Y. (2000). A computer learner corpus based analysis of the acquisition order of English grammatical morphemes. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective: Papers from the third international conference teaching and language corpora* (pp. 123-132). Frankfurt am Main: Peter Lang.

Tono, Y. (2007). The JEFLL Corpus Project. Retrieved June 6, 2009, from http://jefll.corpuscobo.net/

Tono, Y. (2008). Kyozai to corpus [The materials and corpora]. In J. Nakamura & S. Hotta (Eds.), *Corpus to eigo kyouiku no setten* [The interface between corpora and ELT]. Tokyo: Shohakusha.

Tono, Y., Izumi, E., & Kaneko, E. (2005). The NICT JLE Corpus: the final report. In K. Bradford-Watts, C. Ikeguchi, & M. Swanson (Eds.), *JALT 2004 Conference Proceedings* (pp. 345-356). Tokyo: JALT Publications. Retrieved August 10, 2007, from http://jalt-publications.org/proceedings/2004/contents.php

Widdowson, H. G. (2000). On the limitations of linguistics applied. *Applied Linguistics, 21*(1), 3-25.

Appendix A

Lists of Sample data of "like" in the NICT JLE Corpus

Type	Level	Example
(1) or (2)	Level 3	My room looks *like* this.
(1) or (2)	Level 7	And one guy is *like* falling down on a ski.
(3)	Level 2	I *like* Giants.
(4)	Level 9	*Like* there are hundreds of language schools nationwide.
(5)	Level 9	All my neighbors were pretty old *like* fifty or sixty.
(6)	Level 7	And his girlfriend was *like* "Oh what are you doing?".

Appendix B

Lists of Sample data of "so" in the NICT JLE Corpus

Type	Level	Example
"So" + "I"	Level 2	How much is this? Yes. *So*, I will take this one.
"So" + Filler(#F#)	Level 3	The Charlie's Angel is starting. *So* #F# #F# do you want to see it with me?
"So" + Self Correction(#SC#)	Level 6	*So* #SC# if I watch #F# #F# if I watch it #SC# in movie theater, …..
"So, (comma)"	Level 9	I don't think ATM is working. *So*, please come and help me.
"So"+ Repetition(#RC#)	Level 5	My mother is house wife. *So* #R# #R# she make breakfast and lunch and dinner for us everyday.