# A Corpus-based Analysis of Basic Spoken Vocabulary in EFL Textbook Conversations

Junko SHIRATO

## 1. Introduction

Vocabulary acquisition was once a neglected area of language study, but its practical importance has been generally recognized over the past few decades, along with a striking development of corpus linguistics and associated technology (e.g., Biber et al., 1998; Hunston, 2002). It is becoming clear that a key element of successful native-like performance in English conversation is mastery of necessary vocabulary and its lexical relations, including collocations and lexical phrases. This has resulted in a considerable number of corpus-based studies on the special nature of vocabulary use in face-to-face spoken interaction and its pedagogical implications (Carter and McCarthy, 1998; Schmitt and McCarthy 1997; Altenberg, 1998; De Cook, 1998; McCarthy, 1998; 1999; McCarthy and Carter, in press).

We must bear in mind, however, that research done so far, especially in Japan, has tended to take up vocabulary in terms of reading and writing competence, consequently imparting slight attention to speaking capability (Takefuta, 1982). Although faced with this situation, few attempts have been made to come up with a basic vocabulary list for spoken communication specifically designed to have a learner's burden reduced. Incidentally, it should be mentioned that we do have a certain number of vocabulary lists mainly for written communication (e.g., Sonoda, 1996; JACET, 2003). This discrepancy is rather startling because there are an increasing number of people who are trying hard to keep up with the times by honing their oral communicative skill in English.

The aims of the present paper are twofold. First, to determine the qualitative differences between the vocabulary in a specially prepared discourse and that in a natural authentic one, and to ascertain what constitutes the distinguishing characteristics in the former discourse. Second, to clearly show how adequately the cumulative findings made by foregone corpus-oriented research are incorporated into the lexical syllabus in Japanese EFL learners' materials. As a first step, a computational analysis of the vocabulary of the concocted dialogues will be conducted, followed by a comparison of the collected data with that of an authentic spoken English corpus and with a standard basic word list. A final application of this is to make useful suggestions, based on the results, as to which specific criteria should dictate vocabulary selection for Japanese EFL learners.

# 2. Previous studies

## 2.1. Specific features of spoken vocabulary

### 2.1.1 Vocabulary size

One of the aims of examining the vocabulary size of native English speakers is to provide an obtainable goal for EFL learners. Therefore there have been a considerable number of studies that investigate the lexical coverage of texts (e.g., Laufer 1992; Nation 1990, 2001, 2002; Schmitt 2000; Schmitt and McCarthy, 1997).

In order to measure vocabulary size, 'lemma' and 'word family' are often used as the basis for word counting (Nation, 2002). Computer software automatically counts every word form occurring more than once in the text as a 'token', and every different word form appearing in the text as a 'type'. The software, however, basically counts every 'type' including sets of the related words having the same stem forms, such as *arrive, arrives, arriving* and *arrived*, which are called inflections. Then all of the inflections, including plural, third person singular present tense, past tense, past participle, -ing form, comparative, superlative and possessive are converged under 'lemma' (Francis & Kucera, 1982). On the other hand, 'word family' consists of a base word and all its inflected forms mentioned above and its closely related derived forms, including affixes (Bauer & Nation, 1993).

As far as the vocabulary in casual conversation is concerned, McCarthy (1998) suggested that a pedagogical target of the first 2,000 words, or lemmas in order of frequency will safely cover the everyday core vocabulary. The Schonell et al. (1956) study of Australian oral English found that 2,000 'word families' cover nearly 99 per cent of lexicon in spoken discourse. On the contrary, a recent study(Adolphs & Schmitt, 2003) analyzing modern spoken corpora argues that 2,000 word families are not sufficient to communicate in casual conversation, because they cover only around 95 per cent of general spoken discourses. Their study concludes that 5,000 individual words, or lemmas are required to achieve about 96 per cent coverage of vocabulary.

In Japan, Takefuta (1982) attempted to investigate whether 2,000 basic words are sufficient for oral communication in a unique way. He analyzed the dialogue of *Love Story* in order to obtain the data on the total number of tokens, types and frequency of the words in spoken English using a computer. He demonstrated that only 1310 word types appeared in the script. He concluded that the number of the vocabulary indicated above might not be sufficient for EFL learners to speak as fluently as a native speaker, however it could enable EFL learners to conduct a certain level of communication in English.

Measuring the spoken vocabulary size of Japanese EFL learners is notably difficult. Although some researchers have attempted to assess the written vocabulary size of Japanese college students (e.g., Mochizuki & Aizawa, 2001), few attempts have been made to a learner's spoken ability.

It has been reported that the consensus of the spoken vocabulary size to suffice for everyday conversation in English is around 2,000 word families for native speakers, though recent corpus-mediated studies have provoked a controversy on the issue. However, it has not been fully discussed whether the figure indicated above is valid for EFL learners.

### 2.1.2 Type/Token Ratios (TTR)

There is every reason to believe that speakers make use of a narrower range of lexical choices than writers do (McCarty & Carter, 1997). As an empirical means for examining the difference use of vocabulary, Chafe and Danielewics (1987) put to use Type/Token Ratios (TTR). TTR, or lexical density, is defined as a measure of the ratio of different words, or types, to the total number of words, or tokens, in a text (Richard and Scmidt, 2002). And its practical use is a measurement of the difficulty of a passage or a text.

TTR, automatically computed, can be a reliable indicator of specific features of vocabulary in spoken language, because it is known to be lower in spoken texts than in the total of the written registers (Biber et. al., 1999). Yates (1996) analyzed three media including LOB written corpus and London-Lund spoken one in regard to TTR. He established that the mean TTR of writing samples is 0.624 while that of speech samples is 0.395 (ibid: 34). It must be conceded, however, that TTR fluctuates depending on how long the text is. In other words, a longer text with more repeated words is accorded with much lower TTR (Biber et al., 1999). Accordingly such useful software as WordSmith tools (Scott, 1999) is available as a mean of standardizing it.

### 2.1.3 Spoken Core Vocabulary

Some specific features to spoken vocabulary were identified based on recent corpus-based analysis (e.g., McCarthy and Carter, 1997, McCarthy, 1998, 1999; Schmitt, 2000; Cowie, 2001; Altenberg, 2003). McCarthy (1998) demonstrates noticeable differences in frequency in the use of single words between spoken and written language examining 100,000 words of Cambridge International Corpus written data (CIC) and 100,000 words of Cambridge and Nottingham Corpus of Discourse (CANCODE). He investigated that the top 100 frequency word list of written texts consists of mostly grammatical words, while that of spoken texts include several lexical words such as *know, well, get/got, think, right,* which are proved to be core elements of oral interaction.

Furthermore, by examining the 3-million samples of CANCODE, McCarthy (1999) also determined nine specific broad categories that constitute a two thousand-word basic vocabulary for spoken communication:

(1) modal items referring to degree of certainty or necessity including modal verbs and other high frequency items that carry related meanings, such as adjectives and adverbs
(2) extremely high-frequency delexical verbs such as *do, make, take* and *get* in their collocations with nouns, prepositional phrases and particles
(3) interactive words representing speakers' attitudes and stances towards the content communicated
(4) discourse markers whose function is to organize the talk and monitor its progress
(5) basic nouns having very general, non-concrete and concrete meanings
(6) general deictic items including demonstratives that relate the speaker to the world in relative terms of time and space
(7) basic adjectives for communicating everyday positive and negative evaluations of people,

situations, events and things

⑻ basic adverbs referring to time, frequency and habituality, and manner and degree

⑼ basic verbs for actions and events denoting everyday activity.

Single words included in the nine categories mentioned above are regarded as the core vocabulary in oral communication.

McCarthy and Carter's current research (in press) analyzing CANCODE containing 5 million words of spontaneous speech collected between 1995 and 2000, also identified that some multi-word units encode interactive functions in face-to-face communication. Those are as follows:

⑴ frequent clusters encoding discourse marking functions, such as *you know, I mean, do you know what I mean*. Those clusters are used to signal a transition in a conversation, and an interactive relationship between a speaker and a listener.

⑵ clusters encoding face and politeness function, such as *do you think, do you want me to* and *I don't know if/ whether*. Those clusters save face of a receiver and to show politeness.

⑶ clusters encoding hedging function such as *I think, sort of, a bit* and *I don't know*. They convey imprecision and make less assertive and less open to challenge or refutation.

⑷ high frequency clusters having vagueness and approximation function such as, *a couple of, and something like that* and *that sort of thing*. They refer to semantic categories in an open-ended way and make the conversation go smoothly.

McCarthy and Carter also discovered that many high-frequency clusters occur with greater frequency than some common single words. The findings of McCarthy and Carter's research are reflected in the following statement (in press):

> Word lists, which focus only on single words risk losing sight of the fact that many high frequency clusters are more frequent and central to communication than even very frequent words. (p. 18 , from the proofs)

As we have seen above, computer-generated frequency word lists have brought us specific features of spoken vocabulary. They revealed that both single words and multi-word clusters encoding interactive functions are considered as core vocabulary in spoken discourses.

### 2.2. Criteria of vocabulary selection in spoken communication

Criteria for choosing vocabulary necessary for functional oral communication have been discussed among some researchers (e.g., West, 1953; Richards, 1969; Nation, 1990). Nation (1990) demonstrated that frequency counts do not always give us enough information and there are several problems associated with them. The most serious problem is that certain useful and important words for casual conversation do not occur in the 2,000 word level, therefore these words should be reinforced into the learner's basic vocabulary list.

Richards (1969) also argued about the limitations of frequency lists. He illustrated that high frequency items have multiple meanings and are context-free, while low frequency items have fewer meanings and are context-bound, therefore for measures of the usefulness of such words we must look elsewhere. The following are possible criteria proposed by Richards: (1)

frequency, (2) range, (3) language needs, (4) availability and familiarity (5) coverage, (6) regularity and (7) ease of learning or learning burden (ibid: 87-102).

Based on the findings of the previous studies mentioned above, Nishizawa (2003) also suggested that the following six practical criteria should be taken into consideration for vocabulary selection for Japanese EFL learners: (1) frequency, (2) range, (3) common words for everyday life, (4) classroom language for students, (5) words useful and specific to Japanese culture, and (6) Japanized English words which are not English words but created in Japan. (ibid:.10)

## 3. The present study

### 3.1. Materials

#### (1) Conversation textbooks, Let's Speak

In order to produce a corpus of concocted dialogues, NHK radio English program textbooks, *Let's speak* (April 2003 - March 2004) were selected, because of the following two reasons. First, these are textbooks from long-term popular programs, with a broad range of listeners from senior high school students to the retired. Secondly, the present program titled *Let's speak* provides many colloquial expressions commonly used in everyday life, with an aim to improve learners' practical communication skills. NHK provides six radio English programs, with three of them aiming at improving learners' basic English skills, and the other three aiming at putting their basic skills into practical use. *Let's Speak* is categorized in the latter group, targeting both elementary and intermediate level learners. The dialogues between native speakers and Japanese are expanded based on a monthly topic covering everyday life, such as moving, using a computer, figuring out a family budget management, going on a diet, getting a physical check-up, and traveling abroad.

Its monthly textbooks are composed of several sections, including 16 English dialogues which are basically completed every 4 days, from Monday to Thursday on a year-round basis. The textbooks for August and December are, however, exceptions because the former is a review material and does not carry any new dialogues and the latter contains only 12 dialogues from the first week to the third week. Consequently the samples amounted to 172 dialogues.

#### (2) British National Corpus (BNC)

The conversational components of British National Corpus (BNC, hereafter) were chosen as authentic data for the contrastive analysis. BNC is made up of 4,124 different text files, comprising over 100 million orthographic words of both spoken and written data. Its conversational corpus is comprised of 153 texts and about 4 million orthographic words, which were collected from everyday face-to-face conversation of 124 adults aged 15 and over, selected by taking into account age, gender and social class across the United Kingdom. (Leech, et al., 2001) There are two reasons why I chose it in this present research. First, the BNC is presently the only publicly available spoken corpus that reveals how people actually talk in everyday conversation. Secondly it has been utilized to produce such common dictionaries as Longman Dictionary of Contemporary English and Oxford Advanced Learner's Dictionary

(Saito et al., 1998), as well as it has had major effects across the whole range of ELT resources since the 1990s (Rundell, 1995)

### (3) *A General Service List of English Words*

West's (1953) *A General Service List of English Words* (GSL, hereafter), covering all 2,000 high frequency headwords with their derivative forms, is still considered the most suitable list as the basis of vocabulary for learning English as a foreign language (e.g., Richards, 1974; Carter and McCarthy, 1988; Schmitt and McCarthy, 1997; McCarthy, 1998), though it is in need of minor adjustments because of its age. West found that frequency alone were not sufficient criteria for deciding what goes into GSL designed for teaching purposes. Other considerations are made on the following five criteria: easiness or difficulty of learning, necessity, cover, stylistic level and intensive and emotional words. In a variety of studies it has been proven that GSL has provided coverage of 78 to 92 per cent of various kinds of texts, averaging around 82 per cent coverage (Engels, 1968; Richards, 1974; Harlech-Jones, 1983).

## 3.2. Procedures

Before conducting a computer-mediated analysis of the vocabulary in *Let's Speak*, a few general features have come to my attention while the written texts have been being compiled.

(1) Many back channels, such as *uh-huh*, and *uh-oh* and discourse markers, such as *you know* and *okay* appear in the dialogues.
(2) There are many ellipses and substitutions in the dialogues.
(3) Some words are described as actually heard, such as *wanna* and *gonna*
(4) The maximum length of each sentence appears to be limited: the maximum number of words per turn is twenty-three.
(5) Repetition and rephrasing are few.

Scanning through the dialogues, I have assumed that the first three characteristics show the typical features of authentic spoken discourses (McCarthy, 1998). On the other hand, some restrictions for concocted discourses in a published textbook, including limitation of space and listeners' concentration might be reflected in the last two features.

Then a computational analysis has been conducted. Let's Speak Corpus (LSC, hereafter) was produced after compiling all of the 172 conversations appearing in the textbooks. The LSC word lists are produced in alphabetical and frequency order using specifically designed software tools, WordSmith Tools (Scott 1999). Then, the data of LSC are compared to that of BNC in order to examine how close or far the both vocabularies are and whether a certain number of words and multi-word clusters representing specific features to spoken communication are included, referring to the findings of the research by McCarthy (1999) and McCarthy and Carter (in press). We do the same frequency count on both single words and multi-word clusters of BNC as McCarthy and Carter did on those of CANCODE.

Then, alphabetical word lists of LSC have been compared against GSL (West, 1953). All of the words on the list are input with their word class, frequency and some other encoded information in Microsoft Excel. Then computational comparative analysis has been conducted to

identify specific features of the vocabulary in the textbook conversation as well as to find a clue useful in vocabulary selection.

Finally, I will suggest specific criteria of lexical syllabus for Japanese EFL learners based on the findings of the present research.

## 4. Results and Discussion

### 4.1. Vocabulary size

Table 1 summarizes the statistics of the LSC single word list. A total of 15,605 tokens and 2,469 types are indicated. All of the types indicated are then manually converted into 'word family' based on Bauer and Nation's definition (1993). In the present research, each tense form of irregular verbs are counted as a separate head word, because learning each form of irregular verbs is essential for beginner level EFL learners (Nation, 2001). In addition, in order to adequately assess the vocabulary size of LSC, all of the proper names and Japanese words were extracted from the sum total of the words in question. The final result shows that the total of 1,789 word families has been enumerated in LSC. As a consequence, it has been identified that the vocabulary size of the textbook dialogues is apparently smaller than the consensus of native speakers' vocabulary size, 2000 word families.

Table 1 Statistics of the LSC wordlist

| | |
|---|---|
| Bytes | 98, 132 |
| Tokens | 15, 605 |
| Types | 2, 469 |
| Word families | 1, 789 |
| Type/Token Ratio | 0. 1582 |
| Standardized Type/Token | 0. 3995 |
| Sentences | 849 |

### 4.2. Type/ Token Ratios (TTR)

The Type/Token Ratio (TTR) of LSC automatically counted by WordSmith Tools is 0.1582. Since TTR varies with the length of the text, the standardized TTR, also automatically calculated, indicates a ratio of 0.3995, while that of BNC indicates 0.3241. It seems that according to the findings of Yates's research (1996), in which he demonstrated that the mean TTR of writing samples is 0.624 while that of speech samples is 0.395, LSC's standardized TTR represents a specific feature of spoken text. It is, however, concluded that the result may suggest that the concocted dialogues in textbooks are more explicit and have fewer repetitions and rephrasing compared to authentic ones. I assume that this is partly due to the higher information load per line in the textbooks.

### 4.3. Computer-based frequency count

#### 4.3.1 Single words

Table 2 shows the 50 most frequent words of LSC and BNC. The shaded cells in the table

indicate forms which occur significantly more frequently in the spoken than in the written based on the findings of McCarthy's research (1998). The top 50 most frequent word list of LSC includes 10 high frequency words commonly used in spoken communication, while BNC includes 16 words. Words appearing on the both lists are as follows: *I, you, so, get, just,* and *know*. Some extremely high-frequency conversational markers, such as yeah, oh, well and yes are not on the LSC list but are on the BNC list, however right is vice versa.

**Table 2 Comparative frequencies of top 50 high frequency single words**

| | LSC | BNC | | LSC | BNC |
|---|---|---|---|---|---|
| 1 | THE | I | 26 | ME | BUT |
| 2 | I | YOU | 27 | ALL | DON'T |
| 3 | YOU | THE | 28 | HOW | ONE |
| 4 | TO | AND | 29 | ARE | SHE |
| 5 | IT | IT | 30 | AT | SO |
| 6 | AND | A | 31 | THEY | WE |
| 7 | THAT | TO | 32 | GET | THERE |
| 8 | OF | THAT | 33 | CAN | THAT'S |
| 9 | IN | YEAH | 34 | NOT | FOR |
| 10 | WE | IN | 35 | UP | LIKE |
| 11 | FOR | OF | 36 | WITH | MM |
| 12 | IS | OH | 37 | YEAR | NOT |
| 13 | ON | NO | 38 | HE | ER |
| 14 | DO | WELL | 39 | LIKE | THIS |
| 15 | YOUR | HE | 40 | OUR | GET |
| 16 | HAVE | IT'S | 41 | TIME | JUST |
| 17 | SO | ON | 42 | I'M | ALL |
| 18 | MY | WHAT | 43 | JUST | SAID |
| 19 | WAS | WAS | 44 | HERE | BE |
| 20 | WHAT | KNOW | 45 | NO | GO |
| 21 | BE | THEY | 46 | RIGHT | UP |
| 22 | THIS | IS | 47 | WORK | THIS |
| 23 | BUT | HAVE | 48 | LOOK | YES |
| 24 | IT'S | GOT | 49 | KNOW | THEN |
| 25 | ABOUT | DO | 50 | GOOD | IF |

*Words highlighted in gray represent specific features to spoken discourses.

Based on McCarthy's nine specific broad categories (1999), the characteristics of the words in LSC are discussed below.

Some modal verbs, such as *can, will, would and could* frequently occur, however, the frequency of other common modal items, such as *seem, sound, certain definitely* and *probably* is relatively low in LSC. The frequency of these items is extremely high in the authentic corpus, and they serve a key role in everyday talk. There may be, however, duplication of close synonyms between the modal verbs and other related modal items, therefore some justification are

necessary for lexical syllabus depends on learners' level. Since the learnability of model verbs is generally higher than that of the other modal items for elementary level learners, the latter may be excluded from their syllabus

As for interactive words, *really* and *pretty* frequently show up, in contrast, *actually* and *basically* hardly ever appear in LSC. The speaker who cannot use these words is regarded as an impoverished speaker, because these words play such an important role in oral communication as they may soften or make indirect potentially face-threatening utterances, or intensify and emphasize affective stance towards the content of utterances (ibid, 1999).

As far as adjectives are concerned, *good, great, bad* occur most frequently, however, *lovely, horrible* and *terrible*, representing more specific evaluations, much less frequently occur in LSC. On the other hand, all of the adjectives mentioned above show an extremely high frequency in BNC. Since these adjectives offer the speaker a range of responding functions, and can be used very simply, even for elementary level learners, they should be presented earlier in the syllabus (O'Dell, 1997). It is, however, important to ascertain how the different adjectives commonly form patterns with other items. *Horrible* and *terrible*, for example, are close in meaning, but the corpus data show that *terrible* is commonly combined with *situation* and *state*, but *horrible* is much less frequently combined with those nouns.

In addition, there are many contracted forms, which are common in spoken and informal English (Leech & Svartvik, 1994), such as *I'll, you've, he's* in LSC. In addition, as particular linguistic forms specific to spoken English, such as "*wanna*" and "*gonna*" are shown in the whole word list of the material examined. Furthermore, back channel responses including *uh-huh, mm-hmm, aha, umm, boo-boo, uh-uh, hush* are shown on the list. McCarthy (1998) argued that these word forms are considered more worthy candidates for the title of word items on the grounds that they express meanings such as acknowledgement, topic pausing, agreement, hesitation. They are not necessarily put on word lists, however, they may indeed be useful vocalizations to learn (ibid: 237).

### 4.3.2 Multi-word clusters

Tables 3 and 4 show the top 20 items in 2-word and 3-word clusters relatively, and table 5 shows the top 10 items in 4-word clusters in LSC and BNC. The clusters showing distinctive features of spoken discourses are highlighted in gray on the basis of McCarthy and Carter (in press) in the tables. It is obvious that there are significant differences between both word lists. Most of the clusters in the text conversations are combinations of function words and do not have specific meanings themselves.

On closer examination of the top 20 high frequency two-word clusters, most of them in LSC are regarded as "fragmentary strings" (De Cock 2000) having neither syntactic nor semantic integrity, such as *in the, of the* and *for a*. On the contrary, the following specific strings to a spoken discourse, such as *I know, I mean* and *I think* are listed in the top 3 in the BNC frequency list. *I think* is the only one cluster that appears in the top 20 two-ward cluster list of LSC.

In addition, the same tendency was observed among three-word and four-word clusters, as

Table 4 and 5 shows. There are very few strings which represent specific features to spoken language in LSC. *I don't know* is only one listed in the top 20 three-word list as well as *what do you mean* is the only one in the top 10 four-word cluster list of LSC.

Table 3 Comparative frequencies of top 20 high frequency two-word clusters

| | LSC | BNC |
|---|---|---|
| 1 | DO YOU | YOU KNOW |
| 2 | IN THE | I DON'T |
| 3 | OF THE | IN THE |
| 4 | TO THE | I MEAN |
| 5 | ON THE | I THINK |
| 6 | A LOT | DO YOU |
| 7 | I THINK | IT WAS |
| 8 | FOR A | ON THE |
| 9 | HAVE TO | AND I |
| 10 | TO DO | I KNOW |
| 11 | ARE YOU | I SAID |
| 12 | GOING TO | DON'T KNOW |
| 13 | I WAS | OF THE |
| 14 | THIS IS | AND THEN |
| 15 | A FEW | HAVE TO |
| 16 | WANT TO | I WAS |
| 17 | AT THE | YOU CAN |
| 18 | HAVE A | IF YOU |
| 19 | TO BE | IS IT |
| 20 | TO GET | GOT A |

*Clusters highlighted in gray represent specific features to spoken discourses.

Table 4 Comparative frequencies of top 20 high frequency three-word clusters

| | LSC | BNC |
|---|---|---|
| 1 | WHAT DO YOU | I DON'T KNOW |
| 2 | A LOT OF | I DON'T THINK |
| 3 | YOU HAVE TO | DO YOU WANT |
| 4 | BY THE WAY | A LOT OF |
| 5 | I DON'T KNOW | WHAT DO YOU |
| 6 | TO MEET YOU | A BIT OF |
| 7 | I HAVE A | HAVE YOU GOT |
| 8 | I WANT TO | DO YOU KNOW |
| 9 | THAT WOULD BE | YOU HAVE TO |
| 10 | TO BE A | YOU WANT TO |
| 11 | YOU WANT TO | YOU KNOW WAHT |
| 12 | A FEW YEARS | I MEAN I |
| 13 | GOING TO BE | AND I SAID |
| 14 | IS GOING TO | HAVE A LOOK |
| 15 | OUT OF THE | I DON'T WANT |
| 16 | WHY DON'T YOU | YOU'RE GOT TO |
| 17 | WOULD YOU LIKE | DON'T KNOW WAHT |
| 18 | A FEW DAYS | MM MM MM |
| 19 | A MATTER OF | BE ABLE TO |
| 20 | AS A MATTER | DO YOU THINK |

*Clusters highlighted in gray represent specific features to spoken discourses.

Table 5 Comparative frequencies of top 10 high frequency four-word clusters

| | LSC | BNC |
|---|---|---|
| 1 | A MATTER OF FACT | MM MM MM MM |
| 2 | AS A MATTER OF | I DON'T KNOW WHAT |
| 3 | GOOD TO MEET YOU | WHAT DO YOU WANT |
| 4 | IS GOING TO BE | I THOUGHT IT WAS |
| 5 | WHAT DO YOU DO | DO YOU WANT TO |
| 6 | WHAT DO YOU MEAN | I DON'T KNOW WHETHER |
| 7 | AS YOU CAN SEE | DO YOU KNOW WHAT |
| 8 | GOING TO BE A | WELL I DON'T KNOW |
| 9 | I WAS GOING TO | DO YOU WANT A |
| 10 | IT'S GOOD TO MEET | YOU KNOW WHAT I |

*Clusters highlighted in gray represent specific features to spoken discourses.

The results show that the concocted dialogues include much fewer multi-word strings encoding discourse making, vagueness and approximation and hedging functions than authentic ones. Among them, we may find that there are distinct differences in the use of the strings of words encoding vagueness and approximation functions, which are inherently at work. For ex*amples, and things like that* occurs only once, and the other strings in this category, such as *that sort of thing, this that and the other, all the rest of it* and *all this sort of thing* never appear in LSC, while those strings mentioned above occur very frequently in the BNC. Since vagueness, approximation, and hedging are central to informal conversation and its absence can make utterances blunt and pedantic, it is reasonable to suppose that the strings mentioned above would be included in the lexical syllabus for EFL learners.

In addition, some collocations with delexical verbs which are considered as important combinations for vocabulary teaching (Aisenstadt, 1981; Sinclair & Renouf, 1998) do not frequently occur in the text conversations. The *get*-passives, such as *get locked in, get done* are very frequently used by native speakers to reflect the speaker's opinion on an event (Carter & McCarthy, 1999), however they rarely occur in textbook dialogues. It is suggested that these collocations should be incorporated in the lexical syllabus because they are considered as spoken core vocabulary (O'Dell, 1997).

Furthermore, the present research has identified that some high-frequency strings appearing the textbook conversations are not commonly used in authentic discourses. The string, *as a matter of fact,* occurs 260 times per 1,000,000 tokens in LSC, while only 1 times in BNC, as well as *It's good to meet* occurs 190 times in LSC, meanwhile, it never occurs in the enormous amount of the BNC spoken data. The usage of these strings should be closely examined.

So far we have seen the vocabulary of the concocted dialogues in the textbooks, the results lead us to the conclusion that it is significantly different from that of authentic ones. In terms of single words, there appear to be certain degree of similarities between the both dialogues, however, regarding the multi-word clusters, the results show that textbook dialogues include much fewer clusters as units of interaction than authentic ones do. It is identified that a very limited number of the clusters encoding the functions of vagueness, approximation and

discourse marking occur in text conversations, though they are essential for taking turns, giving responses, and for making the conversation go smoothly (McCarthy, 1998). It is obvious that without acquiring such clusters, it is difficult for EFL learners to achieve mastery of communication skills in English.

### 4.4. Basic spoken vocabulary

It has been demonstrated that frequency alone is not sufficient criteria for selecting vocabulary for teaching purposes (West, 1953). Therefore, all of the words on LSC are checked against GSL. As a result, LSC contains 1569 high frequency word families listed on GSL, which accounts for 78.5 per cent of the total number. In other words, 431 high frequency word families are not contained in the textbooks. There seems to be two reasons for this.

First as the size of the examined corpus is rather small; the total number of tokens is less than 20,000. Therefore it appears difficult to encompass 80 per cent of the basic high frequency words.

Secondly, some of the high frequency words may not necessarily need to be listed, because they are already known to learners. The material writer may exclude words which are assumed to be very basic words introduced at junior high school level, as well as loanwords written in *katakana* such as *address, bank, body, camp, ceremony, clock, cool, fish, fork, kitchen, garden, mistake, pen, post, river, road, safe, smile, speed, staff, stamp, tea, wool, zero.* I assume that those words are considered to be absolutely familiar to the targeted learners. However, keen attention should be paid to a certain number of loan words written in *katakana* (McCreary, 1990; Tamaoka & Miyaoka, 2003). *Katakana* words, which are created by simply transforming the original sounds of foreign words into those of the Japanese phonetic system (Tamaoka & Miyaoka, 2003), are categorized in the following three groups. First, those having phonetic similarity with the original English words but having a quite different meaning from the original ones, such as *challenge, smart, mansion,* and secondly, those having Japanized pronunciation, which are phonetically different from original ones, including *model, metal, thrill.* And thirdly, especially close attention should be paid to Japanese-English words, which are created in Japan and commonly misconceived as loan words from English words. Following words are categorized in this group: *handoru* (steering wheel), *mishin* (sewing machine) and *kuulaa* (air conditioner) (Reischauer, 1997; cited in Takefuta, 1982).

The remaining words listed in GSL, which are not included in the textbooks, are considered to be incorporated into the material, however, there seems to be several exceptions. Since GSL listed useful words for both written and spoken communication, there are many words that display significant differences in distribution between them. Let us see some examples. Table 6 shows that *despair, elastic, inquire*, and *tremble* occur much less frequently in spoken communication. This result shows that low frequency words need careful evaluation and further observation in word selection.

Table 6 Comparative frequencies in spoken and written samples of BNC

| | spoken samples | written samples |
|---|---|---|
| despair | 17 | 1816 |
| elastic | 14 | 712 |
| inquire | 0.5 | 256 |
| tremble | 1.7 | 329 |

Then let me give some examples of the common words which are not listed in GSL but appear in LSC in order to identify the characteristics of the words in LSC.

*aerobic, apartment, area, airport, beach, blanket, cancel, cash, communicate, computer, concert, contract, couple, credit, culture, dentist, deposit, design, diet, documentary, double, exit, festival, final, goal, golf, guy, handicap, hello, hi, incredible, internet, jazz, laundry, license, magazine, mall, menu, okay, passport, phone, relax, schedule, student, TV, video, vacation*

All of the words indicated above seem to be very useful in casual conversation. Some of them are newly created words related to advanced technology such as *computer, internet*, and some of them are used in overseas travel such as *airport, passport,* however it is surprising to notice some words, such as *hi, hello, okay*. Although those words are not listed in GSL, they are indispensable for everyday conversation.

A multi-word unit is, however, not completely listed not only in GSL but most of the other basic word lists as a headword. This discrepancy appears to be a fundamental problem for the basic word lists. If the frequency of such items is high enough to get them into basic word lists in direct competition with single words, then perhaps they should be included. (Nation & Meara, 2002). These findings have turned up a clue useful in word selection for Japanese EFL learners.

### 4.5. Implications for lexical syllabus

Finally let us consider the appropriate criteria of the vocabulary selection for the Japanese EFL learners' lexical syllabus in spoken communication. Referring to the findings of the current analysis, I suggest that specific consideration be taken into the following six criteria based on Nishizawa's criteria (2003).

First, 2000 most frequent single words and a certain number of multi-word clusters demonstrated in the present paper are indispensable for spoken communication (McCarthy, 1998; McCarthy and Carter, in press). Learners are especially encouraged to learn adequate usage of modal items, delexical verbs and their collocations, interactive words and clusters encoding functions such discourse marking, face and politeness, and vagueness and approximations.

Secondly, words related to everyday life are essential to casual conversations even though their frequencies are low. For example, the following words are categorized in this group: *message, envelop, medicine, customer, retired*. These words are not listed in the 2000 high frequency word list of spoken demographic samples of BNC. In addition, some multi-word units, including collocations with delexical verbs, such as *get a job, make coffee* as well as some

adjacency pairs, such as *thank you* and *you're welcome* are very commonly used in casual conversation and should be incorporated into the syllabus.

Thirdly, *chopstick, shrine, temple*, are absolutely essential to explain Japanese culture to foreigners. Furthermore, it is also important for beginner leveled learners to recognize that such Japanese words as *sushi, tofu, sake, kimono, haiku,* and *futon* have already become English words, which appear in the latest Longman dictionary of contemporary English (LDOCE, hereafter) (Gadsby, 1995).

Fourthly, particular attention should be paid to some loan words from English, and written in *katakana* when they are incorporated into lexical syllabus. Especially learners should be given an awareness of Japanized-English words, which are created in Japan and commonly misconceived as loan words (Takefuta, 1982)

Fifthly, some newly created words along with development of technology and lifestyle, such as *computer, electronic, Internet, desktop, software, mobile phone, homestay,* are often used in everyday conversation and essential for communication nowadays. These words are not listed in GSL, but appear in the latest LDOCE (Gadsby, 1995). *Electronic* and *software* have been listed in the Longman Defining Vocabulary of around 2000 common words, which are constantly being researched and checked to make sure that they are commonly used by speakers in everyday conversations. (Gadsby, 1995).

Finally, lexical sets which are likely to interest learners in any particular group should be included. Since many Japanese EFL learners aim to speak English overseas, words necessary for traveling abroad, such as *check-in, flight, departure, boarding* are categorized in this group. In addition, traveling abroad is one of the most popular topics among ESL and EFL learners, therefore, it often comes up in computer-mediated communication forums. (e.g., Dave's ESL Cafe).

## 5. Limitations

Although the results of the current paper highlighted the significant lexical differences and similarities between textbook conversations and authentic discourse, there are two notable limitations. The first limitation is the size of the textbook corpus produced in the present research. It is undeniable that a total of less than 20,000 tokens is rather small, though it assumed to be acceptable because it is said that even relatively small samples can yield original insights or can raise awareness for future observation and verification in the field (McCarthy, 1997). However, more data is necessary for further investigation. The second one is found in the selection of the criterial authentic data. It is apparent that more emphasis has been given to American English than British English in linguistic pedagogy in Japan. There is, however, no equivalent spoken corpus of American English to that of BNC available at the present moment. The American National Corpus (ANC, hereafter) has just announced their first release, though only a limited amount of spoken data which do not include face-to-face conversations could be obtained. Therefore, further separate investigation based on the final version of the ANC spoken components would provide a more useful perspective for teaching vocabulary to Japanese learners.

## 6. Conclusion

Computer-generated frequency word lists have revealed how much core spoken vocabulary is in place in the form of the cited teaching materials. Overall, the material examined represent the nature of vocabulary use in face-to-face interaction in its single words, however there are clearly limitations in the selection of its multi-word clusters. Specifically, those units encoding interactive functions, such as discourse marking and vagueness and approximation do not frequently appear, although they are explicitly essential for successful communication. Therefore more emphasis should be put on them in vocabulary selection based on the findings of corpus-based research in future teaching materials.

It is regarded as a concern for both EFL material writers and learners that no reliable spoken vocabulary list exists in Japan. The present research has led me to believe that a comprehensive single word and multi-word cluster frequency list would be extremely important for acquiring natural usage ability. However, raw lists of items need careful evaluation and further observations before a vocabulary syllabus can be established for EFL learners. Therefore, further research into producing a list representative of the criteria outlined here would significantly boost a learners' speaking ability.

References

Adolphs, S. and Schmitt, N. (2003). Lexical coverage of Spoken Discourse. *Applied Linguistics* 24(4), 425-438.

Aisenstadt, E. (1981). Restricted collocations in English lexicology and lexicography. *IRAL*. 53, 53-61.

Altenberg, B. (1998). On the Phraseology of Spoken English: The Evidence of Recurrent Word-Combinations. In A.P. Cowie (Ed.), *Phraseology: Theory, Analysis, and applications*. New York; Oxford University Press.

American National Corpus (2003). American National Corpus First Release. Retrieved September 30, 2004. from http://www.americannationalcorpus. org/FirstRelease/

Bauer, L and Nation, I.S.P. (1993). Word families. *International Journal of Lexicography*. 6, 253-279.

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

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999). *Longman grammar of spoken and written English*. Essex; Longman.

Carter, R. and McCarthy, M. (1998). *Vocabulary and Language Teaching*. New York: Longman.

Carter, R. and McCarthy, M. (1999). The English get-passive in spoken discourse: description and implications for an interpersonal grammar. *English Language and Linguistics*. 3(1), 41-58.

Chef, W. and Danielewicz, J. (1987). Properties of Spoken and Written Language. In R. Horowitz, & S.S. Samuels (Eds.), *Comprehending Oral and Written Language*. San Diego, CA: Academic Press.

Cowie, A.P. (2001). *Phraseology: Theory, Analysis, and Applications*. New York: Oxford University Press.

Dave's ESL Café. Retrieved August 10, 2004, from.http://www.eslcafe.co

De Cock, S. (2000). Repetitive phrasal chunkiness and advanced EFL speech and writing. In C. Mair, & M. Hundt (Eds.), *Corpus Linguistics and Linguistic Theory. Paper from ICAME 20 1999*. (pp. 51-68). Amsterdam: Rodopi.

Engels, L.K. (1968). The fallacy of word-counts. *International Review of Applied Linguistics in Language Teaching*. 6 (3), 213-231.

Francis, W.N. and Kucera, H. (1982). *Frequency Analysis of English Usage*. Boston: Houghton Mifflin Company.

Gadsby, A. (1995). *Longman dictionary of contemporary English.* (3rd Ed.) Harlow, Essex: Longman.

Harlech-Jones, B. (1983). ESL Proficiency and a Word Frequency Count. *ELT Journal* 37, 62-70

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

Laufer, B. (1992). What Percentage of Text-Lexis is Essential for Comprehension? In C. Lauren, & M. Nordmann (Eds.), *Special Language: From humans to thinking machines* (pp.316-323). Clevedon: Multilingual Matters.

Leech, G. and J. Svartvik (1994). *A Communicative Grammar of English.* New York; Longman.

Leech, G., Rayson, P. and Wilson, A. (2001). *Word Frequencies in Written and Spoken English:* based on the British National Corpus. Harlow: Pearson Educational Limited.

McCarthy, M (1998). *Spoken Language and Applied Linguistics.* Cambridge: Cambridge University Press

McCarthy, M (1999). "What constitutes a basic vocabulary for spoken communication?" *SELL* 1, 233-249.

McCarthy, M and Carter, R. (1997). Written and spoken vocabulary. In N. Schmitt, & M. McCarthy (Eds.), *Vocabulary: Description, acquisition, and pedagogy.* (pp. 20-39). Cambridge: Cambridge University Press.

McCarthy, M. and Carter, R. ( in press ) This that and the other: Multi-word clusters in spoken English as visible pattern of interaction. *Teanga.* 21 (Yearbook of the Irish Association for Applied Linguistics).

McCreary, D.R. (1990). Loan words in Japanese. *Journal of Asia Pacific Communication.* 1, 61-69.

Mochizuki, M. and Aizawa, K. (2001). A validity study of the vocabulary size test of controlled productive ability. *Reitaku University Journal.* 73, 85-102.

Nation, I.S.P. (1990). *Teaching and Learning Vocabulary.* Massachusetts: Newbury House.

Nation, I.S.P. (2001). *Learning Vocabulary in Another language.* Cambridge: Cambridge University Press.

Nation, I.S.P. and Meara, P. (2002). Vocabulary. In N.Schmitt (Ed.), *An introduction to Applied Linguistics.* (pp. 35-54). London: Arnold.

O'Dell, F.(1997). Incorporating vocabulary into the syllabus. In N. Schmitt, & M. McCarthy.(Eds.), *Vocabulary Description, Acquisition and Pedagogy.* (pp. 258-276). Cambridge: Cambridge University Press.

Reischauer, E. (1978). *The Japanese.* Tokyo: Charles E. Tuttle Co.

Richard, J.C. (1969). A Psycholinguistic Measure of Vocabulary Selection. *International Review of Applied Linguistics in Language Teaching.* 8(2), 87-102.

Richard, J.C. (1974.) Word Lists: Problems and Prospects. *RELC Journal* 5(2), 69-84.

Richard, J.C. and Schmidt, R. (2002) *Longman dictionary of language teaching and applied linguistics.* (*Third Ed.*). London: Longman.

Rundell, M. (1995). The BNC: A Spoken Corpus. *Mordern English Teacher* 4(2), 13-15.

Schmitt, N. (2000). *Vocabulary in Language Teaching.* Cambridge: Cambridge University Press.

Schmitt, N. and McCarthy, M. (1997). *Vocabulary Description, Acquisition and Pedagogy.* Cambridge: Cambridge University Press.

Schonell, F., Meddleton, I., Shaw, B.(1956). *A study of the oral vocabulary of adults.* Brisbane: University of Queensland Press.

Scott, M. (1999). *Wordsmith Tools.* Software. Oxford: Oxford University Press.

Sinclair, J & Renouf, A. (1988). A lexical syllabus for language learning. In R. Carter & M. McCarthy (Eds.), *Vocabulary and Language Teaching.* (pp.140-158). London: Longman.

Tamaoka, K. and Miyaoka, Y. (2003). The cognitive Processing of Japanese Loanwords in Katakana. *Japanese Psychological Research.* 45(2), 69-80.

Wray, A. (2002). *Formulaic Language and the Lexicon.* Cambridge: Cambridge University Press.

West, M. (1953). *A General Service List of English Words.* London: Longman, Green & Co.

Yates, J. S. (1996). Oral and Written Linguistic Aspects of Computer Conferencing: A Corpus Based Study. In S.C. Herring (Ed.), *Computer-Mediated Communication.* (pp.29-46). Amsterdam/Philadelphia: John Benjamins Publishing Company.

斉藤俊雄・中村純作・赤野一郎（編）(1998)『英語コーパス言語学』基礎と実践. 研究社　pp.24-26.

園田勝英 (1996)『大学生用英語語彙表のための基礎的研究』言語文化部研究報告書叢書７. 北海道大学言語文化部.

大学英語教育学会基本語改定委員会（編）.（2003）『大学英語教育学会基本語リスト（JACET8000）』. 大学英語教育学会.
竹蓋幸夫（1982）『日本人の英語の科学』研究社. pp. 95-114.
西澤正幸（2003）『語彙数はどれだけ必要か』語彙習得のメカニズム. 英語教育52巻７号. 大修館. pp.8-10.
日本放送協会・日本放送出版協会編（2003-2004）『NHK ラジオ英会話レッツスピーク』（2003 April－2004March）日本放送出版協会