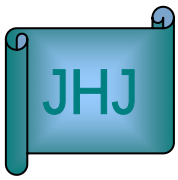
Proceedings



JALT Hokkaido Journal Vol.8 pp. 69-78 journal@jalthokkaido.org

A computational analysis of English vocabulary in textbook conversations

Junko Shirato
g03103@hokusei.ac.jp
Graduate School of Literature Hokusei Gakuen University

The practical importance of vocabulary acquisition has been generally recognized over the past few decades, along with a striking development of corpus linguistics (Hunston, 2002). It is becoming clear that a key element of successful native-like performance in English conversation for EFL learners is mastery of necessary vocabulary. This has resulted in a considerable number of corpus-based studies on the special nature of vocabulary use in face-to-face interaction and its lexical relations, including collocations and lexical phrases, as well as its pedagogical implications (McCarthy, 1998, 1999; De Cock, 2000; McCarthy and Carter, in press).

The aims of this paper are 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. Here, a computational analysis of the vocabulary in the concocted dialogues was conducted, followed by a comparison of the collected data with that of an authentic spoken corpus.

近年、英語教育において語彙の習得に関する興味は急速に拡大してきている。その理由の一つとして、コンピューターの革命的な進化に伴う大規模なコーパスの構築があげられる。話し言葉の語彙に関しても、英語母語話者の日常会話を集めた話し言葉コーパスが作られ、それを分析することにより、話しことばに特有な語彙の特徴が明らかにされている。本研究では、日本人英語学習者向けの英会話教材の中の会話部分に、既存研究によって明らかにされている話し言葉の語彙的特徴がどれほど反映されているかを分析していく。語彙の使用頻度に関しては、一語単位で比較するだけでなく、2 語、3 語、4 語の連語としての頻度表を比較検討することにより、両者の語彙選択における相違点および類似点をより明確にしていく。母語話者の語彙の分析結果から、連語の中にはかなり使用頻度が高いものが存在している事が明らかになっており、将来的に日本人英語学習者用の話し言葉の基本語彙表が構築される際には、一語単位の語彙のみならず、これらの連語も同様に基本語彙として考えていく事が効果的な語彙習得には重要であると考えられる。

Spoken core vocabulary

By examining the 3-million-word samples of CANCODE (Cambridge and Nottingham spoken corpus), McCarthy (1999) determined that there are nine specific broad categories that constitute a two thousand word basic corpus for spoken communication: (1) modal items referring to degree of certainty or necessity; (2) extremely high frequency delexical verbs, *do, make, take* and get in its collocation with nouns and 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) deictic items that relate the speaker to the world in relative terms of time and space; (7) basic adjectives representing especially positive and negative evaluations of people, situations, events and things; 8) basic adverbs referring to time, habituality and degree; (9) basic verbs for actions and events commonly used in everyday conversation. Single words included in these nine categories are regarded as the core corpus in oral communication.

McCarthy and Carter's corpus-based current research (in press) also identified that there are three categories of multi-word clusters, all of which encode interactive functions as follows: (1) discourse marking, (2) face, politeness and hedging, and (3) vagueness and approximation. They also discovered that many clusters are more frequently used than some high frequency single words. The findings of their 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)

Those clusters are therefore considered to be fundamental to successful interaction.

Materials.

In order to produce a sample corpus, NHK radio English program textbooks, *Let's speak* (2003.4 – 2004. 3) were selected and in the present research, because their dialogues provide many colloquial expressions with the aim of improving practical communication skills in a broad range of listeners, from senior high school students to the retired. A total of 172 dialogues between native speakers and Japanese are expanded upon based on a monthly topic covering everyday life, such as moving in, using a computer, family budget management, diet and physical check-ups, etc.

The conversational component of British National Corpus (BNC, hereafter) was also chosen as authentic data for the contrastive analysis. The corpus is comprised of 153 texts and about 4.2 million orthographic words which were collected from the everyday conversations of 124 adults, selected by taking into account age, gender and social group across the United Kingdom (Burnard, 2002). The BNC is presently the only publicly available large-scale spoken corpus that reveals how people actually talk in everyday conversation (Burnard, 2002).

Rank-order frequency lists of single word and two-, three-, and four-word sequences have been generated from both the *Let's Speak* Corpus (LSC, hereafter) and BNC.

Results and Discussion

Single words - Computer-based frequency count

Single word frequency lists of both corpora have been analyzed and compared based on the findings of the research by McCarthy (1999) by using WordSmith tools (Scott, 1999).

Table 1 shows the 20 most frequent words of LSC and BNC. The shaded cells in the

table represent high frequency words showing specific features of a spoken discourse. .Four high frequency words, including *I, you*, *have* and *so* appear on LSC, while seven high frequency words, including extremely high-frequency conversational markers such as *yeah*, *oh*, *well*, appear on *the* BNC list.

Table 1. Top 20 high frequency single word lists

	LSC		BNC
1		**THE**	**I**
2		I	**YOU**
3	**YOU**		THE
4	TO		AND
5	IT		IT
6	AND		A
7	THAT		TO
8	OF		THAT
9	IN		**YEAH**
10	WE		IN
11	FOR		OF
12	IS		**OH**
13	ON		**NO**
14	DO		**WELL**
15	YOUR		HE
16	**HAVE**		IT'S
17	**SO**		ON
18	MY		WHAT
19	WAS		WAS
20	WHAT		**KNOW**

*Shaded cells represent specific features of spoken discourses.

Based on McCarthy's nine specific broad categories (1999), the characteristics of 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* are 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, duplications of close synonyms between the modal verbs and other related modal items; therefore, some justification is necessary for a lexical syllabus depending on learners' levels. 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 an affective stance towards the content of utterances (McCarthy, 1999).

As far as adjectives are concerned, *good, great, bad* occur most frequently, however, *lovely, horrible* and *terrible*, representing more specific evaluations, hardly ever occur in LSC. On the other hand, all of the adjectives mentioned above show an extremely high frequency in the BNC. Since these adjectives offer the speaker a range of response functions, and can be used very simply, even for elementary level learners, they should be presented earlier in the syllabus (O'Dell, 1997). However, it is 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.

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)

Multi-word clusters - Computer-based frequency count

The two-, three-, four-word clusters appearing in the LSC were compared to those in the BNC (Table 2, 3 and 4). It is obvious that there are significant differences between both word lists. On closer examination of the lists of the top 10 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* (Table 2). On the contrary, the following specific strings to a spoken discourse, such as *I know, I mean, I think* are listed in the top 3 on the BNC frequency. *I think* is the only one cluster that appears in the top 10 two-word cluster list of the LSC. There is the same tendency among the three-word and four-word cluster lists (Table 3 and 4).

Table 2. Top 10 high frequency two-word cluster lists

	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

*Shaded cells represent specific features of spoken discourses.

Table 3 Top 10 high frequency three-word cluster lists

	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**

* Shaded cells represent specific features of spoken discourses.

Table 4 Top 10 high frequency four-word cluster lists

	LST	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

* Shaded cells represent specific features of spoken discourses.

The results clearly show that the concocted dialogues include many fewer multi-word strings encoding such interactive functions as discourse marking, vagueness and approximation, and hedging than authentic ones do. 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 example, *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 a lexical syllabus for EFL learners.

In addition, some collocations with delexical verbs which are considered as important combinations for vocabulary teaching (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 do not frequently 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 in the textbook conversations are not commonly used in authentic discourse. The string, *as a matter of fact*, occurs 256 times per 1,000,000 tokens in LSC, while only 0.1 times in BNC, as well as *It's good to meet* occurs 192 times in LSC, meanwhile, it never occurs in the enormous amount of the BNC conversational samples. The usage of these strings should be closely examined.

Conclusion

Computer-generated frequency word lists have revealed how much core spoken vocabulary is in place in the materials examined. Overall, the materials examined represent in 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.

The lack of a reliable spoken vocabulary list for both EFL material writers and learners is a clear concern. 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. In addition to this, further research into producing a basic spoken wordlist would significantly boost learners' speaking ability.

References

Burnard, L. (2002). Where did we Go Wrong? A Retrospective Look at the British National Corpus. In B. Kettemann, & G. Marko (Eds.), *Teaching and Learning by Doing Corpus Analysis* (pp.51-70). New York: Rodopi.

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

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

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

NHK (April, 2003- March, 2004). *Let's Speak*. NHK Publisher.

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.

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

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

