

Examining short conversation questions' use of English with the Exam Corpus

Hiroko USAMI*

1. Introduction

Corpora, which are electronically collected written or spoken texts that represent a particular language or genre for linguistic analysis (e.g. Baker et al., 2006; McEnery et al., 2006), have been applied to various linguistic fields. This includes the language teaching field when utilising computer-assisted language learning (CALL), data-driven learning (DDL) and compiling teaching materials and dictionaries (e.g. Hunston, 2002; Chambers, 2010; Cheng, 2010; Walsh, 2010).

Recently, various ways of employing different types of corpora for language testing and assessment have been recognised and discussed (Park, 2014; Cushing, 2017). For example, general corpora have been used to analyse vocabulary and grammar item frequency, authenticity, word senses and appropriate distractors presented in authentic texts of corpora for writing, editing or proofreading question items (Ball, 2001; Barker, 2006; Crossley et al., 2007; Jaén, 2007; Hughes, 2008). In addition to general corpora, learner corpora have been utilised in various ways: standardisation across and within different exam levels and suites using the Cambridge Learner Corpus, the Cambridge Corpus of Spoken and Learner English, characterising learners' English across proficiency levels, test development and validation in the English Profile Programme (EPP), and automated evaluation of learners' output (Boyle & Booth, 2000; Ball, 2001; Barker, 2004).

In addition to general and learner corpora applied to language testing and assessment, specialised corpora containing specific fields of English have been utilised in language testing in various ways. Presumably, the most basic application of specialised corpora in language testing would be constructed and analysed to examine specific and targeted English usages, such as academic, business and field-specific English, to develop wordlists and test items. Nevertheless, few specialised corpora containing actual past examination question items have been constructed to inform and validate the English used or to incorporate frequently targeted items tested in examinations.

Therefore, this study aims to demonstrate how the Exam Corpus, a specialised corpus containing past English proficiency exam question items constructed by the author, can be utilised in language testing and assessment. The Exam Corpus will help inform and examine English frequently and characteristically presented and tested in English proficiency tests. This paper investigates what multiword units are frequently and characteristically presented as n-grams and keyword lists and what specific conversation situations would be frequently tested in short conversation questions from English proficiency tests in the Exam Corpus.

2. Literature Review

2 – 1 Specialised Corpora for Language Testing and Assessment

The application of corpora in the field of language testing and assessment is relatively new (e.g. Hunston, 2002). However, recently, different types of corpora have been applied to language testing and assessment in various ways discussed in the journal of *Language Testing* (e.g. Park, 2014; Cushing, 2017). In addition to general and learner corpora, specialised corpora containing particular types of text have been applied to language testing and assessment. Barker (2006, 2010) suggested that field-specific (e.g. business, law, aviation and accountancy) and age-specific corpora should be constructed for future applications of corpus data. Recently, Park (2014) introduced specialised corpora used to develop tests, and Egbert (2017) introduced three types of corpora that can be used in language testing.

2 – 2 Various Ways of Applying Specialised Corpora

As general and learner corpora, specialised corpora have been compiled for each research purpose and applied to language testing and assessment in several ways.

Specialised corpora were employed to inform and validate existing examinations. The TOEFL 2000 Spoken and Written Academic Language Corpus (T2K-SWAL) was used to investigate university-level language skills and provide an empirically grounded alternative to the intuition of TOEFL test constructors and item writers (Biber et al., 2001; 2004; Biber, 2006). The University Entrance Examinations in Japan (UEEJ) Corpus (the previous version of the Exam Corpus), containing multiple-choice vocabulary and grammar questions from Japanese university entrance examinations, was constructed to examine and analyse the English presented in those examinations. The collocations used in the stems were analysed, and the frequently targeted grammar, answers and distractors presented in the questions were analysed with Japanese learners' productions in the Longman Learners' Corpus and authentic English in the British National Corpus (BNC) (Usami, 2005; 2012; 2015). Usami (2021) compiled the Exam Corpus containing various English proficiency tests and suggested possible applications through which the English presented in the stems, answers and distractors, and their CEFR levels and collocations could be examined across different English proficiency tests.

As well as informing and validating existing questions with specialised corpora, as a relatively challenging approach, specialised corpora have been utilised to develop wordlists. The University of Cambridge Local Examinations Syndicate (UCLES) constructed the Business English Texts Corpus, a web-based collection of business texts, and developed the Business English Certificate (BEC) Preliminary wordlist. Item writers can use the BEC to produce realistic examination tasks at specific levels by indicating the collocational patterns of certain words or phrases and suggesting different senses of words in real texts (Ball, 2001; 2002; Horner & Strutt, 2004).

Perhaps the most challenging way to apply specialised corpora to language testing and assessment is using it to create question items. The Michigan Corpus of Academic Spoken English (MICASE), a spoken corpus of language used in American academic settings, has been used to develop listening tests and pronunciation assessments (Read, 2002; Levis & Cortes, 2008). As another example, the British Academic Written English corpus, which contains the students' written productions assessed as proficient, was utilised to develop a grammar test that reflected the actual language used by students (Sharpling, 2010).

2 – 3 Construction and Analysis of the Spoken Corpora

There are more written corpora than spoken corpora as general corpora. However, corpora containing various types of spoken text have also been constructed, and many studies have been conducted using

spoken corpora.

The British National Corpus (BNC) accounts for approximately 10% of the spoken texts of the entire BNC, including spoken conversations and other spoken material collected in the early 1990s (Hoffmann et al., 2008). Using this corpus, the word frequencies of written and spoken corpora in the BNC were analysed (e.g. Leech et al., 1996). As the successor to the BNC, the British National Corpus 2014 (BNC2014), containing millions of contemporary spoken and written English words, was constructed. The Spoken BNC2014, which is a spoken component of the BNC2014, contains 11.5 million words from 1,251 informal spoken interactions by 672 speakers of British English collected between 2012 and 2016 (Love et al., 2017). The Corpus of Contemporary American English (COCA) contains more than one billion words from eight genres, including oral texts (<https://www.english-corpora.org/coca/>), and various studies have been conducted (e.g. Rafatbakhsh & Ahmadi, 2019).

Regarding spoken corpora containing learners' spoken productions, the Louvain International Database of Spoken English Interlanguage (LINDSEI) contains over one million words from 554 informal interviews (e.g. a warming-up activity, a free informal discussion and a picture description) with higher intermediate to advanced EFL learners of English from 11 mother tongue backgrounds, including Japanese (Gilquin et al., 2010). Nevertheless, for spoken learner corpora containing learners of English in Asian countries, the International Corpus Network of Asian Learners of English (ICNALE) contains spoken monologues (Ishikawa, 2014) and spoken dialogues (Ishikawa, 2019), as well as written essays, edited essays and a global rating archive. As for a learner corpus focusing on Japanese learners of English, the National Institute of Information and Communication Technology Japanese Learner English Corpus (NICT JLE Corpus) contains 1,281 Japanese learners' spoken productions taken from the Standard Speaking Test (SST) (Izumi et al., 2004).

For specialised corpora containing spoken texts, the Michigan Corpus of Academic Spoken English (MICASE) contains spoken texts used in American academic settings and has been utilised to develop listening tests and pronunciation assessment (Read, 2002; Levis & Cortes, 2008).

3. Research Questions

As seen in the above literature review, specialised corpora containing various specific texts have been employed for language testing in various ways. Moreover, some spoken corpora have already been constructed and analysed in terms of wordlists, keyword lists and collocations. However, few specialised corpora that contain short conversation questions presented in actual past English proficiency examinations have been compiled. Furthermore, little work utilising specialised corpora to inform, validate and analyse the English actually presented and tested in the examinations has been conducted, although the number of short conversation questions presented in past English proficiency examinations tends to be decreasing, as discussed later. Therefore, this study aims to inform and examine the English presented in short conversation questions from English proficiency examinations using the short conversation sections of the Exam Corpus, a specialised corpus created by the author. The following research questions (RQs) were posed:

RQ1: What multiword units are frequently and characteristically presented in short conversation questions?

RQ2: What conversation situations would be frequently tested in short conversation questions?

4. Method

4 – 1 The Exam Corpus

In this study, a specialised corpus, the Exam Corpus, was analysed to examine the n-grams and keywords in English proficiency tests, answering RQs 1 and 2, respectively. The Exam Corpus is a specialised corpus created by the author to be applied to language testing and assessment and examine existing English proficiency tests. Table 1 shows the number of questions across each skill and the English proficiency examination in the Exam Corpus (Ver. 1.3).

Table 1

Number of question items across each skill and examination in the Exam Corpus (Ver. 1.3)

Skill / Exam	UEEJ	TEAP	EIKEN	GTEC	TOEFL	TOEIC	CMB
Reading	✓	✓	✓	✓	✓	✓	✓
Listening	✓	✓	✓	✓	✓	✓	✓
Long conversation	✓	✓	✓	✓	✓	✓	✓
Short conversation	3,094	-	✓	-	-	-	-
Grammar	10,876	0	405	0	176	903	0
Vocabulary	9,442	60	1,995	92	0	463	380

Note. UEEJ = University Entrance Examinations in Japan. TEAP = Test of English for Academic Purposes. GTEC = Global Test of English Communication. TOEFL = Test of English as a Foreign Language. TOEIC = Test of English for International Communication. CMB = Cambridge English.

The Exam Corpus (Ver. 1.3) contains 934,682 words presented in 27,886 exclusively multiple-choice vocabulary, grammar and short conversation questions from four kinds of English proficiency tests, taken from only the official test books published after 2000 by authorised examination boards or publishers: 1) University Entrance Examinations in Japan (UEEJ) for 2001–2023, 2) English proficiency examinations administered in Japan (Test of English for Academic Purposes (TEAP), EIKEN—pre-1st, 2nd, pre-2nd and 3rd grades for 2011–2022, and Global Test of English Communication (GTEC)), 3) English proficiency examinations created by the Educational Testing Service (ETS) in the U.S. (Test of English as a Foreign Language (TOEFL) in 2012, and Test of English for International Communication (TOEIC) for 2005–2020); and 4) English proficiency examinations created by Cambridge Assessment English in the U.K. (Cambridge English—Key English Test (KET), Preliminary English Test (PET) and First Certificate in English (FCE) for 2006–2014). The Exam Corpus (Ver. 1.3) currently contains only vocabulary, grammar and short conversations questions but will add short conversation question items presented in EIKEN and question items presented in reading, listening and long conversations set in the near future, as indicated as the symbol ✓ in the Table 1. In this study, only short conversation questions, totalling 3,094 questions taken from university entrance examinations in Japan, were analysed.

4 – 2 Short Conversation Questions in The Exam Corpus

In this research, out of the question items included in the Exam Corpus, short conversation questions presented in Japanese university entrance examinations are examined and analysed, as Japanese university entrance examinations contain many short conversation questions. In this study, short conversation questions were categorised as conversation questions, including up to four turns, whereas long conversation questions included more than five turns. Presenting conversations up to four turns without gap-filling questions or those with gap-filling questions did not matter.

(1) Short conversation question with gap-filling question

Jim: Thanks to your help, my brother got a good grade.

Tom: () I'm relieved to hear that. I know that she studied so hard.

A) Good for him! B) That's too bad. C) Are you kidding? D) We don't know.

(Obunsha, 2023: 37)

(2) Short conversation question without gap-filling question

Hiro: Jane! It's me, Hiro. Do you remember me?

Jane: Hi, Hiro! Long time no see. How have you been?

Hiro: I've moved to Kyusyu. I got a job there as a system engineer.

Jane: Good for you. You were always good with computers at school.

(Obunsha, 2021: 239)

Table 2 shows the number of question items across each year in the Exam Corpus (Ver. 1.3).

Table 2

Number of short conversation question items across each year in the Exam Corpus (Ver. 1.3)

Year	# of question items	Year	# of question items	Year	# of question items
2001	144	2009	194	2017	80
2002	199	2010	156	2018	72
2003	217	2011	144	2019	63
2004	217	2012	99	2020	61
2005	242	2013	110	2021	72
2006	221	2014	87	2022	72
2007	201	2015	89	2023	65
2008	189	2016	100	Total	3,094

The Exam Corpus (Ver. 1.3) contains 3,094 short conversation questions from only Japanese university entrance examinations. Although the number of questions slightly differs depending on the year, the average number of questions presented is 134.52 questions per year; nonetheless, the number of short conversation questions decreased slightly after 2012.

4 – 3 #LancsBox

To analyse the conversation questions in the Exam Corpus, the software package #LancsBox (see <http://corpora.lancs.ac.uk/lancsbox/>), which was developed at Lancaster University, was used. Among the functions available in #LancsBox, the functions called *Ngrams* (for n-grams), *KWIC* (for concordances) and *GraphColl* (for collocations) were used.

5. Analysis and Discussion

5 – 1 Statistics in the Short Conversation Questions

Before analysing n-grams and keywords, first, the following statistics were obtained to examine the lexical density in the short conversation questions and other two spoken corpora accessible in the #LancsBox.

Table 3 shows the statistics, such as token, type, type-token ratio (TTR) and R in the short conversation questions and other spoken corpora, Spoken BNC2014 containing informal speech of British English

collected in the 2010s, and BNC64 containing informal dialogic speech of British English collected in 1994. The number of tokens in these three corpora was considerably different. Therefore, it is advisable to compare lexical density based on the value of R rather than that of TTR. The value of R in the Spoken BNC2014 was the highest, whereas that of R in the short conversation questions was the lowest, indicating that the short conversation questions contained the least number of words compared to the other two spoken corpora. Thus, it is assumed that specific words or multiword units may be used or repeated in specific situations.

Table 3

Statistics in the short conversation questions, Spoken BNC2014 and BNC64

Corpora / Statistics	Token	Type	TTR	R
Short conversation questions	134,156	6,053	4.51%	16.52
Spoken BNC2014	10,395,068	68,093	0.65%	21.11
BNC64	1,686,930	24,056	1.42%	18.52

5 – 2 N-grams in the Short Conversation Questions

Second, n-grams in the short conversation questions were obtained to analyse frequently presented multiword units. In short conversation questions, relatively long multiword units, 4-, 5- and 6-grams, which are fixed expressions with specific meanings, rather than 1-, 2- and 3-grams, seem to tend to be used as follows:

Table 4

4-, 5- and 6-grams in the short conversation questions in the Exam Corpus

Rank	4-gram	Freq.	5-gram	Freq.	6-gram	Freq.
1	would you like to	115	do you mind if I	25	A what do you think of	11
2	are you going to	59	would you like to go	24	A do you mind if I	11
3	do you want to	58	question A B C D	23	what are you going to do	10
4	what do you think	46	what do you think of	22	would you like to join us	9
5	may I help you	39	do you want to go	21	could you tell me how to	8
6	do you have any	39	A what do you think	20	what can I do for you	8
7	A would you like	37	may I help you B	20	you tell me the way to	8
8	A what do you	36	A would you like to	17	do you want to go to	8
9	B zzz A I	34	what would you like to	15	what do you think of the	8
10	you mind if I	33	what are you going to	14	A what are you going to	8
11	do you have a	32	how long have you been	14	you tell me how to get	8
12	A zzz B I	32	you tell me how to	14	tell me the way to the	8
13	A do you have	31	A excuse me do you	12	A have you ever been to	7
14	could you tell me	30	are you going to do	12	A hello may I speak to	7
15	can I help you	30	would you like me to	12	where would you like to go	7
16	I help you b	29	do you have the time	12	may I help you B zzz	7
17	A I'd like to	28	would you like to join	12	A excuse me do you know	7
18	do you mind if	27	where would you like to	12	A would you like to go	7
19	A zzz B yes	27	have you ever been to	12	A would you mind if I	6
20	how do you like	26	A do you have a	11	A do you have any plans	6

21	I wish I could	26	you can say that again	11	A do you want to go	6
22	you like to go	24	would you like to have	11	excuse me is this seat taken	6
23	A B C D	23	A do you mind if	11	tell me how to get to	6
24	question A B C	23	A do you have any	11	what would you like to do	6
25	you want to go	23	A are you going to	10	what seems to be the problem	6
26	to go to the	23	A how often do you	10	A excuse me is this seat	6
27	do you think of	22	I'm glad to hear that	10	A would you like me to	5
28	can you tell me	22	do you have any plans	10	what did you think of the	5
29	I have no idea	22	what do you think about	10	A may I help you B	5
30	do you have the	21	how can I help you	10	A excuse me can you tell	5
31	what would you like	21	me the way to the	9	do you think of the new	5
32	that's a good idea	20	B that's a good idea	9	excuse me can you tell me	5
33	what are you doing	20	you like to join us	9	A what do you think about	5
34	A do you know	19	A do you want to	9	A excuse me could you tell	5
35	A what are you	19	tell me the way to	9	you want to go to the	5
36	how long have you	19	you do me a favor	9	can you tell me where the	5
37	nice to meet you	18	can I help you B	9	me is this seat taken B	5
38	how often do you	18	would you mind if I	8	A how long have you been	5
39	may I speak to	18	I don't think I can	8	do you have any plans for	5
40	A zzz B I'm	18	tell me how to get	8	could you tell me the way	5

Note. The symbol *zzz* is used to replace a bracket. The upper cases, *A*, *B*, *C* and *D* are the symbols placed before each option.

First, relatively long multiword units asking likeness are frequently presented in various sentence types and following verbs; probably a part of interrogative sentences (e.g. *would you like to*, *do you want to*, *what would you like*, *what would you like to*, *what would you like to do*), such as (3); ones presented in a part of option (e.g. *A would you like*, *A I'd like to*, *A would you like to*, *A do you want to*); ones following specific verbs such as *go*, *join* and *have* (e.g. *you like to go*, *you want to go*, *would you like to go*, *do you want to go*, *do you want to go to*, *you want to go to the*, *where would you like to go*, *would you like to join*, *would you like to join us*, *you like to join us*, *would you like to have*, *where would you like to*), such as (4); ones presented in a part of option (e.g. *A would you like to go*, *A do you want to go*); and ones containing an object (e.g. *would you like me to*, *A would you like me to*), such as (5).

(3) A: Well, what would you like to do?

B: Nothing. I'm just going to stay home.

(Underline is mine.)

(Obunsha, 2001: 211)

(4) A: Where would you like to go?

B: I'd like to go to the museum. Is it very far from here?

(Underline is mine.)

(Obunsha, 2003: 566)

(5) A: I was so busy I completely forgot to call her.

B: Would you like me to take care of it for you?

(Underline is mine.)

(Obunsha, 2012: 450)

Second, relatively long multiword units asking about and answering thoughts were also frequently presented, probably as a part of interrogative sentences (e.g. *what do you think, do you think of, what do you think of, what do you think about, what do you think of the, what did you think of the, do you think of the new*), such as (6); ones presented as an option (e.g. *A what do you, A what do you think, A what do you think of, A what do you think about*), and the answer (e.g. *I have no idea, I don't think I can*), such as (7).

(6) A: What do you think of the new theater?

B: I think it's ugly.

(Underline is mine.)

(Obunsha, 2003: 530)

(7) A: What do you think she'll say?

B: I have no idea.

(Underline is mine.)

(Obunsha, 2004: 356)

Third, relatively long multiword units asking for plans are also presented, probably as a part of interrogative sentences using *be going to* (e.g. *Are you going to, what are you going to, are you going to do, what are you going to do*), such as (8); using *plan(s)* (e.g. *do you have any, do you have any plans, do you have any plans for*), such as (9) and those presented in a part of the option (e.g. *A are you going to, A what are you going to, A do you have any, A do you have any plans*).

(8) A: What are you going to do next year?

B: I'll probably go to law school.

(Underline is mine.)

(Obunsha, 2003: 187)

(9) A: Do you have any plans for the winter holidays?

B: I'm going to visit my cousin in Boston.

(Underline is mine.)

(Obunsha, 2006: 426)

Multiword units using *how* are also presented; most likely as a part of interrogative sentences (e.g. *how do you like, how long have you, how long have you been, how often do you*), such as (10) and ones presented in a part of the options (e.g. *A how long have you been, A how often do you*).

(10) A: How long have you been playing volleyball?

B: For about 10 years.

(Underline is mine.)

(Obunsha, 2005: 477)

More specifically, multiword units using the specific verb *mind* are also presented; probably as a part of interrogative sentences (e.g. *you mind if I, do you mind if, do you mind if I, would you mind if I*), such as (11) and ones presented in a part of the option (e.g. *A do you mind if, A do you mind if I, A would you mind if I*).

(11) A: Would you mind if I sit here?

B: ()

A) I don't know where. B) Mind your own business. C) No, not at all. D) Yes, I'd love to. E) Yes, please sit.

(Underline is mine.)

(Obunsha, 2005: 437)

In more specific situations, multiword units providing an offer and making a request are also presented, most likely as a part of interrogative sentences (e.g. *may I help you, can I help you, how can I help you, you do me a favor, what can I do for you*), such as (12) and (13) and ones presented in some options (e.g. *I help you b, may I help you b, may I help you b zzz, A may I help you B, can I help you b*).

(12) A: How can I help you?

B: The TV turns on, but I can't change channels.

(Underline is mine.)

(Obunsha, 2006: 438)

(13) A: Tony, will you do me a favor?

B: Sure. What is it?

(Underline is mine.)

(Obunsha, 2004: 164)

In more specific situations, asking a direction, the following multiword units are also presented; most likely as a part of interrogative sentences (e.g. *could you tell me, can you tell me, excuse me can you tell me, you tell me how to, tell me how to get, could you tell me how to, you tell me how to get, tell me how to get to, you tell me the way to, me the way to the, tell me the way to, tell me the way to the, could you tell me the way, can you tell me where the*), such as (14) and (15) and ones presented in a part of options (e.g. *A excuse me could you tell, A excuse me can you tell*).

(14) A: Excuse me. Can you tell me where the spaghetti sauce is?

B: Yes, it's in the Italian food section.

(Underline is mine.)

(Obunsha, 2019: 341)

(15) Excuse me, could you tell me the way to the station, please?

(Underline is mine.)

(Obunsha, 2006: 520)

In more specific situations asking for a seat, the following multiword units are also presented: most likely as a part of interrogative sentences (e.g. *excuse me is this seat taken*), such as (16) and ones presented in a part of the options (e.g. *A excuse me is this seat, me is this seat taken b*).

(16) A: Excuse me, is this seat taken?

B: No, it isn't.

(Underline is mine.)

(Obunsha, 2008: 412)

In the situation used in the phone conversation, the following multiword units are also presented: most likely as a part of interrogative sentences (e.g. *may I speak to*), such as (17) and ones presented in a part of the options (e.g. *A hello may I speak to*).

(17) A: May I speak to Mr. Green?

B: I'm sorry. He is not here right now.

(Underline is mine.)

(Obunsha, 2007: 225)

Finally, a fixed expression area is presented (e.g. *you can say that again, I'm glad to hear that, that's a good idea, do you have the time, nice to meet you*), as follows:

(18) A: I really enjoyed the show tonight.

B: I'm glad to hear that. It was fun, wasn't it?

(Underline is mine.)

(Obunsha, 2012: 450)

(19) A: Maybe we should contact security.

B: That's a good idea. There's a guard over there by the information counter.

(Underline is mine.)

(Obunsha, 2009: 43)

5 – 3 Keyword List in the Short Conversation Questions

Finally, in this section, the keyword list in the short conversation questions was obtained to examine what specific situations would be tested using the specific keywords in the short conversation questions compared to normal spoken conversations taken from the general spoken corpus, BNC64.

Table 5

Keyword list of the short conversation questions in the Exam Corpus

Rank	4-gram	Stats.	5-gram	Stats.	6-gram	Stats.
1	would you like to	7.12	do you mind if I	2.81	A what do you think of	1.81
2	may I help you	3.90	question A B C D	2.71	A do you mind if I	1.81
3	do you have any	3.81	would you like to go	2.63	would you like to join us	1.67
4	A would you like	3.75	what do you think of	2.49	could you tell me how to	1.59
5	A what do you	3.66	may I help you b	2.49	what can I do for you	1.59
6	B zzz A I	3.53	A what do you think	2.47	you tell me the way to	1.59
7	you mind if I	3.39	A would you like to	2.26	A what are you going to	1.59
8	A zzz B I	3.38	do you want to go	2.17	you tell me how to get	1.59
9	A do you have	3.31	you tell me how to	2.04	tell me the way to the	1.59
10	do you have a	3.26	how long have you been	1.99	what do you think of the	1.58
11	could you tell me	3.21	what would you like to	1.98	what are you going to do	1.58
12	can I help you	3.21	A excuse me do you	1.89	do you want to go to	1.55
13	are you going to	3.16	do you have the time	1.89	A have you ever been to	1.52
14	I help you B	3.16	have you ever been to	1.89	A hello may I speak to	1.52

15	A I'd like to	3.08	would you like to join	1.88	where would you like to go	1.52
16	A zzz B yes	3.01	where would you like to	1.88	may I help you B zzz	1.52
17	do you mind if	2.94	would you like me to	1.85	A excuse me do you know	1.52
18	do you want to	2.93	A do you have a	1.81	A would you like to go	1.52
19	how do you like	2.90	A do you mind if	1.81	A would you mind if I	1.44
20	I wish I could	2.83	A do you have any	1.81	A do you have any plans	1.44
21	what do you think	2.81	you can say that again	1.80	A do you want to go	1.44
22	question A B C D	2.71	would you like to have	1.80	excuse me is this seat taken	1.44
23	A B C D	2.69	what are you going to	1.75	tell me how to get to	1.44
24	you like to go	2.61	A are you going to	1.74	what seems to be the problem	1.44
25	can you tell me	2.56	A how often do you	1.74	A excuse me is this seat	1.44
26	do you have the	2.55	I'm glad to hear that	1.74	what would you like to do	1.42
27	I have no idea	2.54	do you have any plans	1.74	A would you like me to	1.37
28	do you think of	2.47	how can I help you	1.74	A may I help you B	1.37
29	A do you know	2.41	what do you think about	1.71	A excuse me can you tell	1.37
30	A what are you	2.41	me the way to the	1.67	do you think of the new	1.37
31	may I speak to	2.34	B that's a good idea	1.67	excuse me can you tell me	1.37
32	A zzz B I'm	2.34	you like to join us	1.67	A what do you think about	1.37
33	B zzz yes I	2.34	A do you want to	1.67	A excuse me could you tell	1.37
34	nice to meet you	2.32	tell me the way to	1.67	me is this seat taken B	1.37
35	how often do you	2.30	you do me a favor	1.67	A how long have you been	1.37
36	how long have you	2.29	can I help you B	1.67	do you have any plans for	1.37
37	what would you like	2.29	are you going to do	1.61	could you tell me the way	1.37
38	A do you mind	2.26	would you mind if I	1.59	you like to join us B	1.37
39	B zzz A I'm	2.26	tell me how to get	1.59	what did you think of the	1.36
40	A zzz B no	2.19	you tell me the way	1.59	can you tell me where the	1.36

Note. The symbol *zzz* is used to replace a bracket. The upper cases, *A*, *B*, *C* and *D* are the symbols placed before each option.

Compared to the n-grams in Table 4, first, the multiword units *may I help you* and *can I help you* are more characteristically presented in Table 5 (2nd and 12th ranks, respectively). Additionally, the multiword unit *you mind if I* would be more characteristically presented in Table 5 (7th rank). Finally, the multiword units *could you tell me*, *you tell me how to*, *you tell me how to get* and *tell me the way to the* are more characteristically presented in Table 6 (11th, 9th, 8th and 9th, respectively). Therefore, the collocates of keywords such as *I help you*, *you mind if* and *you tell me* were obtained to examine in what situations these multiword units collocate with what multiword units.

Table 6

Collocates (left 5 to right 5 span) in the short conversation questions in the Exam Corpus

I help you			you mind if I		you tell me	
Rank	position	collocate	position	collocate	position	collocate
1	left	A	right	I	right	the
2	left	may	left	A	left	A
3	right	B	left	do	right	to
4	left	can	right	B	left	could
5	right	you	right	the	left	can
6	left	how	right	you	left	you
7	right	yes	right	here	right	is
8	right	zzz	left	would	right	how
9	right	like	right	zzz	left	me
10	right	is	left	to	right	where
11	right	I'm	right	window	left	excuse
12	left	hello	right	ask	left	I
13	right	I	right	smoke	right	get
14	middle	the	right	your	right	B
15	right	do			right	way

Note. The symbol *zzz* is used to replace a bracket. The upper cases, *A*, *B*, *C* and *D* are the symbols placed before each option.

First, as shown in Table 6, the multiword unit *I help you* tends to occur with *may*, *can*, *how* and *hello* in the left position, whereas it collocates with *you*, *yes*, *like*, *is*, *I'm*, *I* and *do* in the right position, as in (20), (21) and (22), respectively.

(20) A: Hello. May I help you?

B: Yes. I'd like a bag. It's a birthday gift for a friend of mine.

(Underline is mine.)

(Obunsha, 2004: 570)

(21) A: Hello. Welcome to Thailand. How may I help you?

B: Hi. I'd like some information on hotels along the beach.

(Underline is mine.)

(Obunsha, 2008: 376)

(22) A: May I help you?

B: No, thank you. I'm just looking.

(Underline is mine.)

(Obunsha, 2001: 22)

As assumed in the above short conversation questions, the multiword unit *I help you* tends to be used in telephone conversations or shops where receptionists or shop clerks tend to start with *hello* and customers say *yes* or *no* and tell them what they are doing or what they want to do.

Second, as shown in Table 6, the multiword unit *you mind if I* tends to collocate with *do*, *would* and *to* in the left position, whereas it cooccurs with *I*, *the*, *you*, *here*, *window*, *ask*, *smoke* and *your* in the right

position, as shown in (23), (24) and (25). Therefore, the multiword unit *you mind if* tends to start with *do* or *would* in the situation asking whether smoking, asking something or opening the window is permitted.

(23) A: Excuse me. Do you mind if I smoke here?

B: No, go ahead.

(Underline is mine.)

(Obunsha, 2004: 547)

(24) A: Would you mind if I ask your price range?

B: No, not at all. We are thinking of something between 50 and 60 dollars.

(Underline is mine.)

(Obunsha, 2006: 371)

(25) A: Do you mind if I open the window?

B: Not a bad idea. Let's get some fresh air in here.

(Underline is mine.)

(Obunsha, 2002: 200)

Third, as shown in Table 6, the multiword unit *you tell me* would tend to collocate with *could*, *can*, *you*, *me*, *excuse* and *I* in the left position, whereas it collocates with *the*, *to*, *is*, *how*, *where*, *get* and *way* in the right position, as in (26), (27) and (28). Therefore, the multiword unit *you mind if* tends to be used in situation where the directions are sometimes preceded by *excuse me* in various expressions asking the directions.

(26) A: Excuse me, but could you tell me where the library is?

B: The library? Go up this street and turn eight on State Street.

(Underline is mine.)

(Obunsha, 2004: 12)

(27) A: Can you tell me the way to the National History Museum?

B: Certainly. Just walk down two blocks and it'll be on your left.

(Underline is mine.)

(Obunsha, 2006: 540)

(28) A: Can you tell me how to get to the nearest subway station?

B: I am sorry, but I can't help you. I am a stranger around here.

(Underline is mine.)

(Obunsha, 2006: 216)

6. Conclusion

This study analysed and discussed how the Exam Corpus can be applied to language testing and assessment by informing and validating the English frequently presented in the examinations. Specifically, this study examined which multiword units would be frequently and characteristically presented and which conversation situations would be frequently tested in short conversation questions from English proficiency tests in the Exam Corpus.

First, as imagined, lexical density in short conversation questions is the lowest compared to that of other general spoken corpora. The result indicates that the least number of multiword units might be used repeatedly and presented in more specific situations in short conversation questions.

Second, multiword units frequently used and presented in short conversation questions are examined. N-grams expressing likeness, asking about and answering thoughts, asking plans, the words *how* and *mind*, giving an offer and making a request, asking a direction, asking for a seat, phone conversations and various fixed expressions were frequently presented.

Third, according to the keywords obtained in the short conversation questions, the keywords *may/can I help you*, *you mind if I* and *could/can you tell me how to get/the way to the* are frequently presented. The multiword unit *may/can I help you* tend to be used in telephone conversations or shops, and receptionists or shop clerks tend to start with *hello* and customers reply *yes* or *no*, telling them what they are doing or wanting to do. The multiword unit *you mind if I* would tend to be used with *do* or *would* in the situation asking whether smoking, asking something, or opening the window is permitted. Finally, the multiword unit *you tell me* tends to be used when directions are sometimes asked preceded by *excuse me* in various expressions.

In this study, short conversation questions taken from university entrance examinations in Japan in the Exam Corpus were examined and analysed. Compared to the authentic English presented in general corpora, n-grams and keywords obtained in short and long conversation questions in English proficiency examinations would be worth examining and analysing to see whether they are actually used in authentic English. However, in the future, the short conversation questions presented in EIKEN are worth examining and analysing. Additionally, the long conversations in which more than five turns occurred are going to be examined and analysed.

Acknowledgements

This research was funded by a Japan Society for the Promotion of Science (JSPS) KAKENHI grant (20K00814). I would like to thank Editage (www.editage.com) for English language editing.

References

- Baker, P., Hardie, A., & McEnery, T. (2006). *A glossary of corpus linguistics*. Edinburgh University Press.
- Ball, F. (2001). Using corpora in language testing. *Research Notes*, 6, 6-8.
- Ball, F. (2002). Developing wordlists for BEC. *Research Notes*, 8, 10-13.
- Barker, F. (2004). Using corpora in language testing. *Modern English Teacher*, 13(2): 65-67.
- Barker, F. (2006). Corpora and language assessment: Trends and prospects. *Research Notes*, 26, 2-4.
- Barker, F. (2010). How can corpora be used in language testing? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 633-645). Routledge.
- Biber, D., Reppen, R., Clark, V., & Walter, J. (2001). Representing spoken language in university settings: The design and construction of the spoken component of the T2K-SWAL Corpus. In R. C. Simpson & J. M. Swales (Eds.), *Corpus linguistics in North America. Selections from the 1999 symposium* (pp. 48-57). The University of Michigan Press.
- Biber, D., Conrad, S.M., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004). *Representing language use in the university: analysis of the TOEFL 2000 spoken and written academic language corpus*. TOEFL Monograph Series No. 25. ETS.
- Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. John Benjamins.
- Boyle, A. & Booth, D. (2000). The UCLES/CUP Learner Corpus. *Research Notes*, 1, 10.
- Chambers, S. (2010). What is data-driven learning? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 345-358). Routledge.
- Cheng, W. (2010). What can a corpus tell us about language teaching? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 319-332). Routledge.

- Crossley, S. A., Louwerse, M., McCarthy, P. M., & McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *Modern Language Journal*, 91, 15-30.
- Cushing, S. T. (2017). Corpus linguistics in language testing research. *Language Testing*, 34(4), 441-449.
- Egbert, J. (2017). Corpus linguistics and language testing: Navigating uncharted matters. *Language Testing*, 34(4), 555-564.
- Gilquin, G., De Cock, S. & Granger, S. (Eds.) (2010). *The Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM*. Presses universitaires de Louvain.
- Hoffmann, S., Evert, S., Smith, N., Lee, D., & Ylva, B.P. (2008). *Corpus Linguistics with BNCweb – a Practical Guide*. Peter Lang.
- Horner, D., & Strutt, P. (2004). Analysing domain-specific lexical categories: evidence from the BEC written corpus. *Research Notes*, 15, 6-8.
- Hughes, G. (2008). Text organisation features in an FCE reading gapped sentence task. *Research Notes*, 31, 26-31.
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press.
- Ishikawa, S. (2014). Design of the ICNALE Spoken: A new database of multi-modal contrastive interlanguage analysis. *Learner Corpus Studies in Asia and the World*, 2, 63-76.
- Ishikawa, S. (2019). The ICNALE Spoken Dialogue: A new dataset for the study of Asian learners' performance in L2 English interviews. *English Teaching*, 74(4), 153-177.
- Izumi, E., Uchimoto, K., & Isahara, H. (2004). *Nihonjin 1,200 nin no eigo speaking corpus*. [English speaking corpus by 1,200 Japanese learners of English]. Alc.
- Jaén, M. (2007). A corpus-driven design of a test for assessing the ESL collocational competence of university students. *IJES*, 7(2), 127-147.
- Leech, G., Rayson, P., & Wilson, A. (1996). *Word Frequencies in Written and Spoken English*. Routledge.
- Levis, J., & Cortes, V. (2008). Minimal pairs in spoken corpora: Implications for pronunciation assessment and teaching. In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive call: Natural language processing for diagnostic language assessment* (pp.197-208). Iowa State University.
- Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014. *International Journal of Corpus Linguistics* 22(3), 319-344.
- McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: an advanced resource book*. Routledge.
- Obunsha. (2001-2009, 2012, 2019, 2021, 2023). *Zenkoku daigaku nyushi mondai seikai eigo—shiritsu dai hen*. [Compilation of English past exam papers – private university exam version]. Obunsha.
- Park, K. (2014). Corpora and language assessment: The state of the art. *Language Assessment Quarterly*, 11(1), 27-44.
- Rafatbakhsh, E., & Ahmadi, A. (2019). A thematic corpus-based study of idioms in the Corpus of Contemporary American English. *Asian-Pacific Journal of Second and Foreign Language Education*, 4:11, 1-21.
- Read, J. (2002). The use of interactive input in EAP listening assessment. *Journal of English for Academic Purposes*, 1, 105-119.
- Sharpling, G. P. (2010). When bawe meets welt: The use of a corpus of student writing to develop items for a proficiency test in grammar and English usage. *Journal of Writing Research*, 2, 175-189.
- Usami, H. (2005). *Using corpora for language testing—examining the English used in university entrance examinations in Japan with the British National Corpus and the Longman Learner Corpus*. Unpublished MA Dissertation, Lancaster University.
- Usami, H. (2012). *The application of corpora to language testing—identifying and improving problematic multiple choice grammar questions in Japanese university entrance exams*. Unpublished PhD Thesis, Lancaster University.
- Usami, H. (2015). *The application of corpus linguistics to language testing—improving multiple choice questions*. Saarbrücken, LAP Lambert Academic Publishing.

- Usami, H. (2021). Applying specialised corpora to language testing: Development of the Exam Corpus (vocabulary and grammar sections). *Tokai University: The Bulletin of the International Education Center*, 3(1), 1-13.
- Walsh, S. (2010). What features of spoken and written corpora can be exploited in creating language teaching materials and syllabuses? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 333-344). Routledge.

要旨

Exam Corpus を用いた短い会話問題における英語使用に関する調査

宇佐美裕子

コーパス(言語分析のために特定の言語の代表となるよう厳選して収集された書き言葉と話し言葉の電子データ)は長い間言語教育に応用されており、最近特に言語テスト・評価の分野において、様々な方法での応用が議論されている。例えば、一般コーパスは語彙・文法項目の頻度、真性、意味、適切な錯乱肢が分析され、語彙・文法問題の作成、編集、校正に用いられている。また、学習者コーパスは、試験レベル間の標準化、能力別の学習者の英語の特徴、テスト開発・検証、学習者の産出の自動評価に使用されている。さらに、特定の分野の英語を収集した特殊コーパスも、様々な方法で言語テスト・評価の分野に使用されている。最も基本的な応用は、特殊コーパスを構築して、特定の英語の使用を調査し、語彙リストやテスト問題を作成することであろう。しかし、実際に出題された過去の試験問題を収集した特殊コーパスはほとんど構築されていなく、そのような特殊コーパスは頻繁にテストされている語彙を調査するのに応用されてもいなければ、試験で使用される英語を調査するのにも十分に応用されていない。従って、本論文は、著者が構築した特殊コーパスである **Exam Corpus** がどのように言語テスト・評価に応用できるかを議論することを目的とする。特に、**Exam Corpus** に収集された英語能力試験の短い会話問題において、どのような **multiword unit** が頻繁に、特徴的に **n-gram** とキーワードリストに現れているか、またどのような会話の状況がテストされているかを検証することを目的とする。結果、好み、思考、計画を問うもの、*how* と *mind* を使用した表現、申し出、依頼、道案内、座席の空き状況、電話での会話などが多く出題されていたことが判明した。また、キーワードリストから、特に電話やお店での会話、*mind* を使用した許可を求める会話、道順を尋ねる会話が多く出題されていることも判明した。