

Applying the Exam Corpus to Inform Targeted Vocabulary in Synonym Questions in Japanese University Entrance Examinations

Hiroko USAMI
Tokai University

Abstract

A corpus is an electronic collection of written or spoken texts selected to represent a particular language for the purpose of linguistic analysis (e.g., Baker et al., 2006; McEnery et al., 2006). Corpora have been applied to language teaching (Chambers, 2010; Cheng, 2010, Chapter 23; Walsh, 2010, Chapter 24) and recently to language testing and assessment (Park, 2014; Cushing, 2017). The application of corpora in the language testing and assessment field has been theoretically suggested since the mid-1990s. Since then, concrete research on the application of general, learner, and specialised corpora to language testing and assessment has been conducted (e.g., Ball, 2001; Biber et al., 2004). Different kinds of corpora have been utilised from the perspective of vocabulary testing. However, few studies have been conducted on the application of specialised corpora to inform and validate the English in vocabulary tests. The aim of this research was to inform and validate the English in vocabulary questions on English proficiency tests using a specialised corpus, the Exam Corpus, in terms of 1) the CEFR levels of the targeted vocabulary in the examinations, 2) the targeted vocabulary tested in the examinations, and 3) the stems, answers, and distractors presented in the most frequently targeted vocabulary item in the examinations. The most frequently targeted vocabulary was B2 level vocabulary and *put up with*. Furthermore, in the answer, four different verbs were presented as synonyms of *put up with*, and some patterns were found for the stems in the question concerning *put up with*.

Keywords: corpus linguistics, language testing and assessment, vocabulary test, CEFR, Japanese university entrance examinations

1. Introduction

A corpus is an electronically encoded collection of written or spoken texts selected to represent a particular language or genre collected for linguistic analysis (e.g., Baker et al., 2006; McEnery et al., 2006). Based on the idea that corpus linguistics is a

methodology, corpora have long been applied to language teaching areas, such as computer-assisted language learning (CALL), data-driven learning (DDL), and compiling teaching materials and dictionaries (Chambers, 2010; Cheng, 2010, Chapter 23; Walsh, 2010, Chapter 24).

The language testing and assessment area is an arena applying corpora to language teaching. Recently, various ways of applying corpora, including general, learner, and specialised corpora, to language testing and assessment have been increasingly recognised and discussed (Park, 2014; Cushing, 2017). General corpora can play a role in writing, editing, or proof-reading items by investigating frequency, authenticity, word senses, and appropriate distractors presented in authentic texts (Ball, 2001; Barker, 2006). In contrast, the main application of learner and specialised corpora has been to compile learner corpora (e.g., the Cambridge Learner Corpus, the Cambridge Corpus of Spoken and Learner English) or specialised corpora containing academic, business, and other field-specific English. Furthermore, learner corpora can be utilised for standardisation across and within different exam levels and suites, characterisation of learners' English across proficiency levels (English Profile Project), test development and validation, and automated evaluation of learners' output (Boyle & Booth, 2000; Ball, 2001; Barker, 2004). However, few studies have compiled specialised corpora containing English proficiency tests or analysed the English presented in these examinations to inform and validate the English used or vocabulary frequently targeted in examinations.

Therefore, this paper aims to discuss the application of a specialised corpus, the Exam Corpus, which was created by the author, to language testing and assessment, especially in order to inform and validate the English presented in vocabulary questions from English proficiency tests. The vocabulary questions in the Exam Corpus were analysed in terms of 1) the CEFR levels of the targeted vocabulary in the examinations, 2) the targeted vocabulary tested in the examinations, and 3) the stems, answers, and distractors presented in the most frequently targeted vocabulary item in the examinations.

2. Literature Review

2.1 Research on Applying Corpora to Language Testing and Assessment

How corpora can be applied to the language testing and assessment field has been discussed since the mid-1990s. First, a language tester, Alderson (1996, Chapter 15) proposed 1) test construction, compilation, and selection; 2) test presentation; 3) response capture; 4) test scoring; and 5) calculation and delivery of results. A few years later, a corpus linguist, Hunston (2002, Chapter 8) dedicated one section in her book, *Corpora in Applied Linguistics*, to the use of corpora in language testing, suggesting 1) a

measurement of typicality of the materials used, 2) the marking of tests, and 3) the development of tests. Subsequently, an exam board, the University of Cambridge Local Examinations Syndicate (UCLES) (currently Cambridge Assessment English) has been publishing corpus-informed research since the early 2000s. In future applications, Barker (2006) suggested 1) automated scoring of spoken performance, 2) new technologies to detect cheating and malpractice, and 3) the creation of new corpora such as field-specific reference corpora (e.g., business, law, aviation, and accountancy), and age-specific corpora.

Following their suggestions, the role of corpora in language testing and assessment has been deeply and specifically discussed from a different point of view. First, general corpora have been utilised in various ways (Ball, 2001; Barker, 2006; 2010), such as in writing items by accessing authentic texts and in editing or proofreading items by checking the frequency and authenticity of their collocations, word senses in contexts, and appropriate distractors (Crossley et al., 2007; Hughes, 2008).

Regarding the application of learner corpora, the basic and most direct way is to compile corpora containing test-takers' written and spoken output collected from examinations (e.g., the Cambridge Learner Corpus, the Cambridge Corpus of Spoken and Learner English) (Boyle & Booth, 2000; Ball, 2001; Barker, 2010). Learner corpora have been utilised in various ways (Boyle & Booth, 2000; Ball, 2001; Barker, 2004; 2010): standardisation across and within different exam levels and suites; characterising learners' language, including collocational patterns and errors at different proficiency levels in the framework of the English Profile Programme (EPP) (Alexopoulou, 2008); test development, including compiling word lists; validation purposes; and automatically scoring learners' written and spoken productions (Burstein et al., 2004).

Although these have not been as popular as general and learner corpora, specialised corpora containing particular types of texts can play a role in language testing and assessment. Like learner corpora, the basic way to apply corpus approaches is by actually compiling a specialised corpus, although they have been applied in various ways. For example, the Michigan Corpus of Academic Spoken English (MICASE) and the British Academic Written English corpus were used to develop a listening test, pronunciation assessment, and a grammar test, respectively (Read, 2002; Levis & Cortes, 2008; Sharpling, 2010). Usami (2021) compiled the Exam Corpus, a specialised corpus containing various English proficiency tests, and suggested possible applications, which this paper pursues: the English presented in the stems, answers, and distractors, and their CEFR levels and collocations can be examined across different English proficiency tests.

2.2 Applying Corpora to Vocabulary Tests

General, learner, and specialised corpora have been applied to vocabulary tests in various ways. Business English Texts Corpus, a web-based collection of business texts constructed by UCLES, was utilised to develop the Business English Certificate (BEC) Preliminary wordlist. It can be used by item writers to produce realistic examination tasks at specific levels, as it indicates the collocational patterns of certain words or phrases, and to suggest the different senses of words in real texts (Ball, 2001; 2002; Horner & Strutt, 2004).

Specialised corpora have also been applied to inform and validate existing examinations. The TOEFL 2000 Spoken and Written Academic Language Corpus (T2K-SWAL) was developed to investigate university-level language skills and provide an empirically grounded alternative to the intuitions of TOEFL test constructors and item writers (Biber et al., 2001; 2004; Biber, 2006). Usami (2005) examined the English presented in multiple-choice vocabulary and grammar questions in Japanese university entrance examinations, focusing on the collocations used in the stems, frequently targeted grammar, answers, and distractors using the University Entrance Examinations in Japan (UEEJ) Corpus (containing multiple-choice vocabulary and grammar questions from Japanese university entrance examinations), the Longman Learners' Corpus, and the British National Corpus (BNC).

Various corpora have been applied to create and develop vocabulary tests, and this is perhaps the most challenging way of applying corpora to language testing and assessment. Rees (1998) developed cloze tests in which the omitted items were selected based on frequency in a large corpus, collocation in the text, repetition of the target word in the text, and word class. Coniam (1997) described how a multiple-choice vocabulary nth-deletion cloze test could be produced automatically based on word frequency data from a corpus. Shillaw (1994) constructed a blank-filling vocabulary test in which the first letter of each missing word was given. The definition of each word was extracted from the COBUILD dictionary, and the targeted words were selected based on their frequency in the High School Corpus (composed of reading textbooks used in senior high schools in Japan) and the Birmingham Corpus (created in the 1980s). Merino (2000) developed receptive and productive vocabulary tests using the Academic Word List (AWL) and the BNC. Usami (2012; 2015) created and improved the stems and answers by checking authentic English from the BNC, and effective distractors using the learners' errors from the Longman Learners' Corpus.

Thus, different types of corpora have been employed in various ways in language testing and assessment, especially in terms of vocabulary tests. However, few specialised

corpora containing English proficiency tests have been compiled, meaning that there is little work applying them to inform, validate, and analyse the English presented and tested in examinations. The aim of this paper is to discuss the application of a specialised corpus, the Exam Corpus created by the author, on language testing and assessment for the purpose of informing and validating the English presented in questions related to vocabulary in English proficiency tests. In this paper, the following specific research questions are addressed: RQ1. What are the most frequent CEFR levels of targeted vocabulary in Japanese university entrance examinations? RQ2. What is the most frequently targeted vocabulary in Japanese university entrance examinations? RQ3. What are the most frequent stems, answers, and distractors presented in the most frequently targeted vocabulary item in Japanese university entrance examinations?

3. Method

3.1 The Exam Corpus

The Exam Corpus is a specialised corpus created by the author for application on language testing and assessment. The current version, 1.1, contains 1,191,850 words presented in 23,837 multiple-choice vocabulary and grammar questions from the four kinds of English proficiency tests, as per official test books published after 2000 by the examination boards or the publishers: 1) university entrance examinations in Japan for 2001–2021, 2) English proficiency examinations held in Japan (e.g., Test of English for Academic Purposes (TEAP) and EIKEN—Grade 1, Grade Pre-1, Grade 2, Grade Pre-2, and Grade 3 for 2011–2019), 3) English proficiency examinations created by the Educational Testing Service (ETS) in the USA (e.g., 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 UK (e.g., Cambridge English–Key English Test (KET), Preliminary English Test (PET), and First Certificate in English (FCE) for 2006–2014).

In the Exam Corpus, each vocabulary or grammar question is stored in one plain text file, and the following meta-information is added: 1) RF=a nine-digit reference number; 2) TN=test name (UEEJ, TEAP, EIKEN, TOEFL, TOEIC, or Cambridge English); 3) TD=test details (names of reference books, grades for EIKEN and Cambridge Examination, year when the question was administered); 4) QS=question skill (e.g., VC, vocabulary; GR, grammar), 5) QC=big-small question category (e.g., modal-can); 6) CR=CEFR level of the targeted vocabulary or grammar (<https://www.englishprofile.org>) (A1, A2, B1, B2, C1, or C2); 7) IF=item facility, indicating difficulty of the item; 8) DI=discrimination index, indicating how well discriminated the item is. As seen in Figure

1, the stem and options for each vocabulary or grammar question are presented below <DI>. The missing part of each stem is labelled *zzz*; instead of brackets, each option is prefaced with *A*), *B*), *C*), or *D*), and the correct option, that is, the answer, is prefaced with * to distinguish it from incorrect answers, or distractors.

Figure 1

A sample plain text file in the Exam Corpus

```
<RF>S3113V013
<TN>EIKEN
<TD>Grade 3 2011-3
<QS>GR
<QC>gerund-verb
<CR>A1
<IF>.76
<DI>.34
A: Do you like zzz in Japan, Mr. Kent? B: Yes, I do.
A)live B)lived C)lives *D)living
```

Furthermore, each question is stored in one row in a separate sub-sheet across different types of examinations in one Excel file, as well as in each plain text file. Hence, someone unfamiliar with analysing the Exam Corpus in plain text files using #LancsBox can search test items using the filter function in Excel across particular meta-information such as examination type, targeted vocabulary and grammar, CEFR level, or the annotated values of item facility or discrimination index.

3.2 Synonym Types of Vocabulary Questions

The current version of the corpus, 1.1, contains both multiple-choice grammar and vocabulary questions. Vocabulary questions are categorised into synonym questions, whereby test-takers are required to choose the answer most similar in meaning to the underlined word(s), and gap-filling questions, such that test-takers are required to select the correct answer to fill in the blank, as follows:

Synonym question:

The primary purpose of his visit is to improve trading relations.

A) further B) solitary C) political D) main

Gap-filling question:

Please () your papers to me by the end of the month.

A) hand out B) hand in C) hand down D) hand of

(Usami, 2018: 37)

In this study, only synonym-related vocabulary questions were examined and analysed; grammar questions and gap-filling vocabulary questions were excluded. In addition, only 2,507 synonym-related vocabulary questions presented in Japanese university entrance examinations were examined because other English proficiency examinations do not provide synonym-related vocabulary questions.

3.3 #LancsBox

To analyse synonym-related vocabulary questions in the Exam Corpus, a software package, #LancsBox (see <http://corpora.lancs.ac.uk/lancsbox/>) developed at Lancaster University was used. The advantage of #LancsBox for this research is being able to employ both plain text files and an Excel file in the analysis. Among the functions available in #LancsBox, the functions *Words* (for wordlists) and *KWIC* (for concordances) were used in this research.

4. Analysis and Discussion

4.1 The Frequently Targeted CEFR Level

RQ1 examined the most frequent CEFR levels of targeted vocabulary in Japanese university entrance examinations. The Excel file of the Exam Corpus was analysed using #LancsBox to obtain the CEFR level frequency breakdown of the targeted vocabulary.

Table 1

Frequency breakdown of the CEFR levels of the targeted vocabulary

| Rank | CEFR level | Frequency |
|------|------------|-----------|
| 1 | B2 | 816 |
| 2 | NA | 550 |
| 3 | B1 | 407 |
| 4 | C2 | 351 |
| 5 | C1 | 275 |
| 6 | A2 | 88 |
| 7 | A1 | 20 |

Note. NA indicates any cases where CEFR levels were not assigned.

As Table 1 shows, vocabulary on the CEFR B2 level was by far the most frequently targeted and tested of the synonym-related vocabulary questions in Japanese university entrance examinations—almost double the frequency of the B1 level. Vocabulary in the CEFR B1, C2, and C1 levels was more frequently tested, as compared to the A2 and A1 CEFR levels. Therefore, almost half of the synonym-related vocabulary questions tested in Japanese university entrance examinations target the CEFR B level (B1 and B2). Unfortunately, 550 targeted vocabulary items in the examinations could not be assigned any CEFR level.

Table 2

The frequently targeted CEFR B2 level vocabulary

| Rank | Vocabulary | Part of speech | Frequency |
|------|-----------------|----------------|-----------|
| 1 | put up with | phrasal verb | 19 |
| 2 | get over | phrasal verb | 16 |
| 3 | figure out | phrasal verb | 14 |
| 3 | turn down | phrasal verb | 14 |
| 5 | look into | phrasal verb | 12 |
| 5 | look up to | phrasal verb | 12 |
| 7 | call off | phrasal verb | 11 |
| 7 | fed up with | phrasal verb | 11 |
| 9 | make out | phrasal verb | 9 |
| 9 | make up for | phrasal verb | 9 |
| 9 | take after | phrasal verb | 9 |
| 12 | count on | phrasal verb | 8 |
| 12 | give in | phrasal verb | 8 |
| 12 | let down | phrasal verb | 8 |
| 12 | run into | phrasal verb | 8 |
| 12 | stand for | phrasal verb | 8 |
| 17 | come up with | phrasal verb | 7 |
| 17 | once in a while | idiom | 7 |
| 19 | on account of | idiom | 6 |
| 19 | make up | phrasal verb | 6 |

Table 2 shows the top 20 most frequently targeted CEFR B2 level vocabulary items. Most of them were phrasal verbs, apart from the adverbial idiom *once in a while* and the prepositional idiom *on account of*. The most frequently targeted CEFR B2 level

vocabulary item was *put up with*, and various phrasal verbs with various verbs (e.g., *put*, *get*, *figure*, *turn*, *look*, *call*, *make*, etc.) were tested.

As analysed in this section, the most frequently tested CEFR level among all 2,507 synonym-related vocabulary questions was B2. In the following section, frequently targeted vocabulary, regardless of CEFR level, is examined.

4.2 The Frequently Targeted Vocabulary Questions

To answer RQ2 (What is the most frequently targeted vocabulary in Japanese university entrance examinations?), that is, to obtain frequently targeted vocabulary, the Excel file of the Exam Corpus was again analysed using #LancsBox.

Table 3

Frequency of the targeted vocabulary

| Rank | Vocabulary | Part of speech | CEFR level | Frequency |
|------|--------------------|----------------|------------|---------------|
| 1 | put up with | phrasal verb | B2 | 19 |
| 2 | account for | phrasal verb | C2 | 17 (1 for N) |
| 3 | figure out | phrasal verb | B2 | 16 (2 for N) |
| 3 | get over | phrasal verb | B2 | 16 |
| 5 | turn down | phrasal verb | B2 | 14 |
| 6 | call for | phrasal verb | C2 | 12 |
| 6 | do away with | phrasal verb | C1 | 12 |
| 6 | get rid of | phrasal verb | B1 | 12 |
| 6 | look into | phrasal verb | B2 | 12 |
| 6 | look up to | phrasal verb | B2 | 12 |
| 6 | stand for | phrasal verb | B2 | 12 (4 for C2) |
| 12 | call off | phrasal verb | B2 | 11 |
| 12 | fed up with | phrasal verb | B2 | 11 |
| 12 | make up one's mind | idiom | B1 | 11 |
| 15 | on purpose | idiom | B1 | 10 |
| 16 | give in | phrasal verb | B2 | 9 (1 for B1) |
| 16 | make up | phrasal verb | B2 | 9 (3 for N) |
| 16 | make up for | phrasal verb | B2 | 9 |
| 16 | make out | phrasal verb | B2 | 9 |
| 16 | take after | phrasal verb | B2 | 9 |

The most frequently targeted CEFR level item in synonym-related vocabulary questions

was B2, as noted above and indicated in Table 3 (14 for B2, three for B1, one for C1, and two for C2). In addition, most of the top 20 frequently targeted B2 vocabulary items were verb-related; indeed, the most frequently targeted vocabulary in synonym-related vocabulary questions in Japanese university entrance examinations was the phrasal verb *put up with*, tested 19 times in 12 different universities. In addition, the C2 level of vocabulary (e.g., *account for* and *call for*), the C1 level (e.g., *do away with*), and the B1 level (e.g., *get rid of*, *make up one's mind*, and *on purpose*) were also frequently tested.

As indicated in Table 3, the part of speech among the top 20 frequently targeted vocabulary items was unanimously either a phrasal verb or an idiom (18 for phrasal verbs and two for idioms). In addition, most of the top 20 frequently targeted vocabulary were verb-related, the exception being adverbial idiom *on purpose*. Surprisingly, the word class/part of speech of any single word was not tested frequently in synonym-related vocabulary questions in Japanese university entrance examinations.

On examining frequent idioms or phrasal verbs in Table 3 in more detail, we find that some were assigned two different CEFR levels according to their meaning. For example, out of 12 questions on *stand for*, eight questions where it meant 'represent' were assigned B2, as in (1), whereas four questions where it meant 'accept' were assigned C2, as in (2).

(1) Do you know what "WHO" stands for?

*A) represents B) announces C) observes D) recognizes

(Obunsha, 2003: 511)

(2) I won't stand for any nonsense!

A) fail B) save *C) endure D) select

(Obunsha, 2016: 33)

As analysed in this section, the most frequently targeted vocabulary was *put up with*, and most of the frequently targeted vocabulary were phrasal verbs or idioms. In the following section, the most frequently targeted vocabulary item is examined in terms of their stems, answers, and distractors.

4.3 Stems, Answers, and Distractors in the Most Frequently Targeted Vocabulary Item

RQ3 (What are the most frequent stems, answers, and distractors presented in the most frequently targeted vocabulary item in Japanese university entrance examinations?) was examined in this section. To obtain the questions and options used in the frequently

targeted vocabulary item *put up with*, the Excel file of the Exam Corpus was again analysed using #LancsBox.

Table 4

Frequency of the answers for put up with

| Rank | Answer | CEFR level | Frequency |
|------|-----------|------------|-----------|
| 1 | *endure | B2 | 7 |
| 2 | *stand | C1 | 5 |
| 3 | *bear | B2 | 4 |
| 4 | *tolerate | B2 | 3 |

Note. “*” indicates that the word is the correct answer.

Table 5

Wordlist used in the stems of the question put up with

| Rank | Word | Frequency | Rank | Word | Frequency |
|------|----------|-----------|------|----------|-----------|
| 1 | put | 19 | 14 | bad | 2 |
| 1 | up | 19 | 14 | anymore | 2 |
| 1 | with | 19 | 14 | weather | 2 |
| 4 | the | 13 | 14 | you | 2 |
| 5 | I | 9 | 14 | made | 2 |
| 6 | longer | 6 | 14 | rude | 2 |
| 6 | noise | 6 | 14 | must | 2 |
| 6 | any | 6 | 14 | no | 2 |
| 6 | cannot | 6 | 14 | of | 2 |
| 10 | his | 5 | 14 | behavior | 2 |
| 11 | she | 4 | 14 | can't | 2 |
| 12 | couldn't | 3 | 14 | to | 2 |
| 12 | this | 3 | 14 | he | 2 |

First, the answers presented as synonyms of *put up with* were examined. As shown in Table 4, four different verbs with different CEFR levels were used to answer the question. Out of 19 questions, the verb *endure* was the most frequent answer, followed by *stand* and *bear*. The word *tolerate* was the least chosen synonym of *put up with*. Furthermore, the CEFR levels of *endure*, *bear*, and *tolerate* were B2, whereas only *stand* was C1. The target *put up with* was assigned B2. However, the same CEFR level B2 or a higher CEFR level C1, not lower CEFR levels such as B1, A2, or A1, was used for the answer.

Second, as for other options used as distractors in this question, various words with various CEFR levels ranging from A1 to C1 were presented. Out of 19 questions, the verbs *enjoy*, *improve*, and *sleep* were presented twice, in two different questions.

Table 6

2-grams and 3-grams used in the stems of the question put up with

| Rank | Word | Frequency | Rank | Word | Frequency |
|------|---------------|-----------|------|-------------------|-----------|
| 1 | up with | 19 | 1 | put up with | 19 |
| 1 | put up | 19 | 2 | up with the | 10 |
| 3 | with the | 10 | 3 | cannot put up | 6 |
| 4 | cannot put | 6 | 4 | I cannot put | 5 |
| 4 | any longer | 6 | 5 | with the noise | 4 |
| 6 | I cannot | 5 | 5 | noise any longer | 4 |
| 7 | the noise | 4 | 7 | up with his | 3 |
| 7 | noise any | 4 | 7 | up with this | 3 |
| 9 | with this | 3 | 7 | couldn't put up | 3 |
| 9 | couldn't put | 3 | 10 | must put up | 2 |
| 9 | with his | 3 | 10 | longer I cannot | 2 |
| 12 | behavior I | 2 | 10 | this noise any | 2 |
| 12 | longer I | 2 | 10 | with this noise | 2 |
| 12 | can't put | 2 | 10 | to put up | 2 |
| 12 | anymore I | 2 | 10 | behavior I cannot | 2 |
| 12 | rude behavior | 2 | 10 | she couldn't put | 2 |
| 12 | this noise | 2 | 10 | rude behavior I | 2 |
| 12 | must put | 2 | 10 | the noise any | 2 |
| 12 | she couldn't | 2 | 10 | can't put up | 2 |
| 12 | to put | 2 | 10 | any longer I | 2 |

Third, words or phrases presented in the stems of the questions on *put up with* were examined. As shown in Tables 5 and 6, some patterns can be observed. First, pronouns such as *I*, *she*, *you*, and *he* shown in Table 5 may have been used as the subject of the stems. Actually, in most cases, it was the subject *I* that *put up with* something, such as in (3).

(3) I can't put up with this condition.

(Obunsha, 2001: 339)

Second, modal verbs—*cannot*, *couldn't*, *must*, and *can't* (in Table 5)—were used in the stems of the questions. Also, the phrasal verb *put up with* was accompanied by *cannot*, *couldn't*, *must*, and *can't*, as in (3). Also, other modal expressions such as *have to* and *had no choice but to* were also used, as follows:

(4) The local residents have to put up with the growing number of tourists.
(Obunsha, 2004: 206)

(5) I had no choice but to put up with my client's rude behavior.
(Obunsha, 2021: 256)

Third, the adverbial expressions *longer*, *any*, and *anymore* in Table 5, and the collocations including *longer*, *any*, or *anymore*, that is, *any longer*, *noise any*, *longer I*, *anymore I*, *noise any longer*, *longer I cannot*, *this noise any*, *the noise any*, and *any longer I* in Table 6 were used in the stems of the questions as follows:

(6) I couldn't put up with his anger anymore.
(Obunsha, 2001: 50)

(7) I cannot put up with this noise any longer.
(Obunsha, 2002: 348)

Fourth, the words or phrases probably used as the subject were also presented: *noise*, *weather*, *rude*, and *behavior* in Table 5, and *the noise*, *noise any*, *behavior I*, *rude behavior*, *this noise*, *with the noise*, *noise any longer*, *this noise any*, *with this noise*, *behavior I cannot*, *rude behavior I*, *the noise any* in Table 6 as follows:

(8) How can you put up with the hot weather without an air conditioner?
(Obunsha, 2004: 54)

(9) She couldn't put up with the noise he made.
(Obunsha, 2004: 261)

(10) We must put up with George's rude behavior.
(Obunsha, 2007: 262)

5. Conclusion

This paper has analysed and discussed how a novel specialised corpus, the Exam Corpus, can be applied on language testing and assessment to inform and validate the English presented in vocabulary questions of English proficiency tests. Multiple-choice

synonym-related vocabulary questions, all from Japanese university entrance examinations, were analysed in terms of 1) the CEFR levels of the targeted vocabulary in the examinations, 2) the targeted vocabulary items tested in the examinations, and 3) the stems, answers, and distractors presented in the most frequently targeted vocabulary item in the examinations.

As for the frequency of the CEFR levels with targeted vocabulary, B2 level vocabulary was most frequently targeted and tested, followed by CEFR B1, C2, C1, and finally A1 and A2 levels. In addition, most of the vocabulary was not assigned a CEFR test level. Among the CEFR B2 vocabulary, the top 20 most frequent words apart from two idioms were phrasal verbs.

Regarding frequently targeted vocabulary, the CEFR level of the most frequently targeted vocabulary was B2, although other CEFR level questions were also tested (three questions for B1, one for C1, and two for C2). The most frequently targeted vocabulary was a phrasal verb *put up with*, and most of the parts of speech in the top 20 frequently targeted vocabulary were phrasal verbs, besides the adverbial idiom, *on purpose*.

As for the answers presented in relation to the frequently targeted vocabulary *put up with*, four different verbs (*endure*, *stand*, *bear*, and *tolerate*) with different CEFR levels were used, three of which were the same CEFR level as the target B2, and one of which was the higher CEFR level C1. Regarding other options, various words with a range of CEFR levels from A1 to C1 were used as distractors, and some verbs were used twice in two different questions. Furthermore, some patterns were observed for the stems used in the question. Specific pronouns such as *I*, *she*, *you*, and *he* were used as the subject in the stems. Modal verbs *cannot*, *couldn't*, *must*, and *can't* were followed by the target *put up with*. Moreover, the adverbial expressions *longer*, *any*, and *anymore* were collocated with the target. Some specific patterns were also seen in the subject someone *put up with* such as *noise*, *weather*, and *rude behavior*.

In this paper, synonym-related vocabulary questions presented in Japanese university entrance examinations, taken from the Exam Corpus, were analysed. For future research, skills and question types can be expanded, and multiple-choice vocabulary and grammar questions presented in other English proficiency examinations stored in the Exam Corpus may be examined. Furthermore, only the most frequently targeted phrasal verb *put up with* was investigated here, in terms of stems, answers, and distractors. It would be worth examining other vocabulary items, their synonyms, and multiple-choice vocabulary and grammar questions as well as phrasal verbs and idioms from a different point of view. Furthermore, the English used in stems in English proficiency examinations may be compared or contrasted in terms of authenticity and frequency with that presented

in general corpora such as the BNC.

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

- Alderson, J. C. (1996). Do corpora have a role in language assessment? In J. Thomas & M. Short (Eds.), *Using corpora for language research: Studies in the honour of Geoffrey Leech* (pp. 248-259). Longman.
- Alexopoulou, T. (2008). Building new corpora for English Profile. *Research Notes*, 33, 15-19. <https://www.cambridgeenglish.org/Images/23152-research-notes-33.pdf>
- 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. <https://www.cambridgeenglish.org/Images/23117-research-notes-06.pdf>
- Ball, F. (2002). Developing wordlists for BEC. *Research Notes*, 8, 10-13. <https://www.cambridgeenglish.org/Images/23120-research-notes-08.pdf>
- 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. <https://www.cambridgeenglish.org/Images/23145-research-notes-26.pdf>
- 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. <https://www.ets.org/Media/Research/pdf/RM-04-03.pdf>

- 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.
<https://www.cambridgeenglish.org/Images/23112-research-notes-01.pdf>
- Burstein, H., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The criterion online writing evaluation service. *AI Magazine*, 25(3), 27-36.
https://www.researchgate.net/publication/220605112_Automated_Essay_Evaluation_The_Criterion_Online_Writing_Service
- 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.
- Coniam, D. (1997). A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests. *CALICO Journal*, 16(2-4), 15-33.
- Crossley, S. A., Louwse, 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. <https://journals.sagepub.com/doi/pdf/10.1177/0265532217713044>
- Horner, D., & Strutt, P. (2004). Analysing domain-specific lexical categories: evidence from the BEC written corpus. *Research Notes*, 15, 6-8.
<https://www.cambridgeenglish.org/Images/23131-research-notes-15.pdf>
- Hughes, G. (2008). Text organisation features in an FCE reading gapped sentence task. *Research Notes*, 31, 26-31.
<https://www.cambridgeenglish.org/Images/23150-research-notes-31.pdf>
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press.
- 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.
- McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: an advanced resource book*. Routledge.
- Merino, M. (2000). *The use of corpora-based input in language testing: developing an academic vocabulary test from ‘The Academic Word List’ and ‘The British National Corpus’*. Unpublished MA Dissertation, Lancaster University.

- Obunsha. (2001-2004, 2007, 2016, 2021). *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.
- Read, J. (2002). The use of interactive input in EAP listening assessment. *Journal of English for Academic Purposes*, 1, 105-119.
- Rees, J. (1998). A new approach to rational deletion in cloze tests. *Paper read at the University of Birmingham* in 1998.
- 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.
https://www.jowr.org/articles/vol2_2/JoWR_2010_vol2_nr2_Sharpling.pdf
- Shillaw, J. (1994). Using a corpus to develop vocabulary tests. In L. Flowerdew & A.K.K. Tong (Eds.), *Entering Text* (pp.166-177). The Hong Kong University of Science and Technology.
- 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*. LAP Lambert Academic Publishing.
- Usami, H. (2018). A validation study of the CEFR vocabulary levels of Japanese English learners in the English Vocabulary Profile. *Tokai University: The Bulletin of the International Education Center*, 38, 35-49.
- 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.
<https://www.u-tokai.ac.jp/uploads/2021/04/8c5ab3983897fe462af464eb6e31fd59-1.pdf>
- 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.