The Bulletin of International Education Center Tokai University
Vol. 3, No.1, 2021, pp. 1-13
DOI: 10.18995/24347337.3-1.1

Applying Specialised Corpora to Language Testing: Development of the Exam Corpus (Vocabulary and Grammar Sections)

Hiroko USAMI
Tokai University

Abstract

A corpus is an electronically collected written or spoken text that represents a particular language for the purpose of linguistic analysis (e.g., Baker, Hardie, & McEnery, 2006; McEnery, Xiao, & Tono, 2006). As a methodology, corpora have been applied to language teaching topics such as computer-assisted language learning (CALL), data-driven learning (DDL), and compiling teaching materials (Chambers, 2010; Cheng, 2010, Chapter 23; Walsh, 2010, Chapter 24), among others. It has been suggested that corpora can play an important role in language testing (Cushing, 2017; Park, 2014). Perhaps the most basic application for corpora in language testing is the use of specialised corpora containing academic and field-specific English terms that have been compiled for the main purpose of developing wordlists and test items. In contrast, there have been few corpora that contain past test items to inform, validate, and assess the English used in examinations, and to analyse aspects of grammar or vocabulary frequently tested in examinations. Therefore, the aim of this paper was to introduce an original, specialised corpus created by the author for language testing, the Exam Corpus. Currently, this corpus contains 1,191,850 words used in multiple-choice vocabulary and grammar questions in different worldwide English proficiency examinations. This paper describes the design, data, metadata collection, annotation, and application of the Exam Corpus after reviewing previous studies on applying corpora to language testing, especially about utilising specialised corpora in this area.

Keywords: corpus linguistics, language testing, English proficiency examinations, vocabulary and grammar, specialised corpus

1. Introduction

A corpus is an electronically collected written or spoken text that is representative of a particular language for the purpose of linguistic analysis (Baker et al., 2006; McEnery et al., 2006), and has been applied to various linguistic fields as a methodology. Both general corpora and learner corpora have been applied to and utilised in language teaching, such as in computer-assisted language learning (CALL), data-driven learning (DDL), and to compile teaching materials (Chambers, 2010; Cheng, 2010, Chapter 23; Walsh, 2010, Chapter 24).

However, the role of general and learner corpora in language testing and assessment has been only recently recognised and discussed (Cushing, 2017; Park, 2014). Besides the general corpora and reference corpora that have been applied to language testing, most large examination boards, publishers, and academic institutions have compiled learner corpora, which contain test takers' written and spoken output (e.g., Cambridge Learner Corpus).

Furthermore, as the most basic application of corpora to language testing, specialised corpora have been compiled to examine specific and targeted English usage, such as academic, business, and field-specific English, for the development of wordlists and test items. However, few specialised corpora contain actual past examination items to inform and validate the English used or incorporate frequently targeted grammar or vocabulary tested in examinations.

Therefore, this paper introduces the Exam Corpus, a specialised corpus which has been created by the author for language testing. This corpus currently contains 1,191,850 words presented in 23,837 multiple-choice vocabulary and grammar questions in worldwide English proficiency tests. This study reviews two streams of literature: one on theoretical suggestions to apply corpora to language testing, and the second, research that has used specialised corpora in language testing. The paper then describes the design, data, metadata collection, annotation, and application of the Exam Corpus.

2. Literature Review

2.1 Theoretical Suggestions of Applying Corpora to Language Testing

It has been suggested that corpora, including general and learner corpora, should be applied to language testing in various ways, as is prevalent in other language teaching fields. Alderson (1996, Chapter 15) was the first paper to discuss the possibilities of utilising computers and corpora in language assessment. From a language tester's point of view, Alderson (1996, Chapter 15) suggested that corpora should be applied to 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 after his suggestion, researchers from the University of Cambridge Local Examinations Syndicate (UCLES) (currently Cambridge Assessment English) began to construct corpora and carry out corpus-informed research. Firstly, Ball (2001) described corpus-building (the Cambridge Learner Corpus, the Cambridge Corpus of

Spoken Learner English, and Business English Texts Corpus) and corpus-informed research, such as the 1) development of examination materials, 2) standardisation across examinations, and 3) development of comparative activities undertaken at UCLES. Furthermore, Barker (2006) introduced the applications of two learner corpora such as the Cambridge Learner Corpus and the Cambridge Corpus of Spoken Learner English, developing wordlists for various examinations and for English Profile, which is an interdisciplinary research programme to enrich the learning, teaching, and assessment of English worldwide. Barker (2006) also suggested future applications for corpus data: 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 and age-specific corpora. More recently, Barker (2010, Chapter 45) introduced corpora which have been created and utilised for language testing and discussed how both general corpora and learner corpora have been and can be informed by language testing. Since the special issue of *Language Testing* focused on improving the connection between corpus linguistics and language testing (Cushing, 2017), there have been more papers discussing various aspects of the application of corpora to language testing (e.g., Egbert, 2017; Xi, 2017).

2.2 Specialised Corpora for Language Testing

Besides utilising general corpora and creating learner corpora as a way of applying corpora to language testing, specialised corpora containing particular types of texts have been compiled and utilised in language testing. Barker (2006, 2010, Chapter 45) from UCLES suggested that field-specific (e.g., business, law, aviation, accountancy) and age-specific corpora should be compiled as future applications for corpus data. Furthermore, Park (2014) introduced specialised corpora that have been used to develop tests. Moreover, Egbert (2017) introduced three kinds of corpora that can be used in language testing.

As for specialised corpora of academic English, the Michigan Corpus of Academic Spoken English (MICASE), which is a spoken corpus of language used in American academic settings, has been utilised to develop listening tests for academic purposes (Read, 2002). Furthermore, the British Academic Written English corpus, which contains the writing of students assessed as proficient, was used to develop a grammar test that reflected the actual language used by students (Sharpling, 2010).

Regarding specialised corpora containing business English, UCLES constructed the Business English Texts Corpus as a project-specific corpus, which is a web-based collection of business texts. They utilised this corpus to develop the Business English

Certificate (BEC) Preliminary wordlist so that item writers can use it to produce realistic examination tasks at specific levels by indicating the collocational patterns of certain words or phrases and to suggest the different senses of words in real texts (Ball, 2001, 2002).

Specialised corpora have been applied not only to create wordlists and test items, but also 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 provides an empirically grounded alternative to the intuitions of TOEFL test constructors and item writers (Biber, Reppen, Clark, & Walter, 2001; Biber et al., 2004).

As seen in the discussion above, there have been various specialised corpora applied to language testing. However, few specialised corpora have been compiled from actual past test items and utilised to inform, validate, and analyse the English actually presented and tested in the examinations. Therefore, this paper introduces and describes the Exam Corpus, along with its design, data, metadata collection, annotation, and application.

3. Corpus Design

3.1 Background of the Exam Corpus

The Exam Corpus is a specialised corpus which the author has been creating to inform, validate, and analyse the English presented in English proficiency examinations. It originated from the University Entrance Exam in Japan (UEEJ) Corpus, which the author began compiling in 2004. As of 2015, the UEEJ Corpus contained approximately 133,000 words of 5,038 multiple-choice vocabulary and grammar questions taken from the 2002–2007 edition of the test book called *Zenkoku Daigaku Nyushi Mondai Seikai Eigo – Shiritsu Dai Hen* (Obunsha, 2002–2007). Since 2018, the author of the current paper has added more questions to the corpus, and the UEEJ Corpus was renamed the Japanese University Entrance Exam Corpus (JUEEC), containing 15,160 multiple-choice vocabulary and grammar questions taken from the 2001–2016 edition of the test book called *Zenkoku Daigaku Nyushi Mondai Seikai Eigo – Shiritsu Dai Hen* (Obunsha, 2001–2016). As a representative sample of the multiple-choice vocabulary and grammar questions in Japanese university entrance examinations, the UEEJ Corpus and the JUEEC included questions presented in private universities in Japan with a relatively wide variety of geographic locations, departments, and proficiency levels and involving the officially authorised National Centre Test for University Admissions. However, these corpora excluded questions presented in public and national universities in Japan, because these universities do not often present multiple-choice vocabulary or grammar questions in their entrance examinations.

```
<RF>20001
<NU>Sapporo
<YE>02
Do you know by any chance what has become zzz Yamada?
XXA)of     B)for    C)with    D)in
```

Figure 1. Sample file in the UEEJ Corpus (taken from Usami, 2015)

In the UEEJ Corpus and the JUEEC, each vocabulary and grammar question is stored as one plain text file and tagged with meta-information: 1) RF for a 5-digit reference number, 2) NU for the name of the university which provided that question, and 3) YE for the year in which the question was administered (e.g., 02 for 2002). The missing part of each question is labelled *zzz*, instead of brackets, as a placeholder for the actual word in cluster and collocational tables. In addition, the correct option in each case is prefaced with the code *XX* to distinguish between correct and incorrect answers, or distractors (see Figure 1). A suite of corpus tools called WordSmith Tools version 3 (Scott, 1999) was used to analyse the data in the UEEJ Corpus and the JUEEC.

3.2 Design of the Exam Corpus

The compilation of the Exam Corpus began in 2019, including the collection of data from the UEEJ Corpus and the JUEEC. Currently, the Exam Corpus contains 1,191,850 words presented in 23,837 multiple-choice vocabulary and grammar questions. However, in order to also examine and analyse other types of English proficiency examinations, the Exam Corpus will include reading, listening, and conversation questions as well as grammar and vocabulary questions (see section 3.4). Therefore, the Exam Corpus functions as a monitor corpus, adding approximately 1,000 new questions from different English proficiency examinations every year and updating the existing files by adding meta-information. In the following sections, the Exam Corpus is described in more detail in terms of the types of examinations, skills and question types, annotation and sample files, and application.

3.3 Types of Examinations Contained in the Exam Corpus

The Exam Corpus has been expanded to include the original UEEJ Corpus and the JUEEC, and was constructed to examine and analyse a wide variety of English proficiency examinations by adding English proficiency tests from around the world, including from Japan. Currently, the Exam Corpus includes the following four categories of English proficiency examinations: 1) university entrance examinations in Japan in 2001–2021, 2) English proficiency examinations held in Japan (e.g., Test of English for Academic Purposes (TEAP) and EIKEN–1st, pre 1st, 2nd, pre 2nd, and 3rd grades in 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)® in 2005–2020), and 4) English proficiency examinations created by Cambridge Assessment English in the U.K. (e.g., Cambridge English–Key English Test (KET), Preliminary English Test (PET), and First Certificate in English (FCE) in 2006–2014). Many reference books based on these English proficiency examinations have been published. Therefore, only the actual questions in the official test books published after 2000 by the examination boards or the publishers are contained in the Exam Corpus.

3.4 Skills and Question Types

The UEEJ Corpus and the JUEEC contained only multiple-choice questions, focusing only on vocabulary and grammar skills. However, to examine and analyse overall English proficiency examinations, the Exam Corpus will include reading, listening, and conversation questions as well as grammar and vocabulary questions. Table 1 shows the number of files across each skill and examination in the Exam Corpus. The questions related to reading, listening, and conversation skills will be added in the future, though most vocabulary and grammar questions are already included in the Exam Corpus. Approximately 1,000 new questions, mainly from Japanese university entrance examinations and EIKEN, will be added, because these past examination books are published annually.

Table 1.

The number of files across each skill and examination in the Exam Corpus

Skill / Exam	UEEJ	TEAP	EIKEN	TOEFL®	TOEIC®	Cambridge
Reading	✓	✓	✓	✓	✓	✓
Listening	✓	✓	✓	✓	✓	✓
Conversation	✓	-	✓	-	-	-
Grammar	7,901	0	250	176	823	0
Vocabulary	12,154	60	1,670	0	423	380

Note. "UEEJ" indicates Japanese university entrance examinations. "✓" indicates that questions will be added in the future, while "-" indicates that no questions will be presented in the examinations.

In the Exam Corpus, only receptive skills such as reading and listening and enabling skills such as grammar and vocabulary are included, while productive skills such as writing and speaking are excluded. Among receptive skills, only discrete points, focused on multiple-choice questions are included because they are widely used in testing receptive skills in English proficiency examinations. For the reading questions, the Exam Corpus will include reading passages, stems, and options with the correct answer and the distractors; for listening questions, the corpus will include scripts, stems, and options again with the correct answer and the distractors. In addition, conversation questions are included; they can be categorised into longer conversation questions (usually more than four-turn conversations) and shorter conversation questions (usually two- to four-turn conversations), as follows:

Shorter conversation question:
A: Brian Ford is the man for the job, don't you think?
B: (   )
A: Who do you suggest, then?
A) I'm afraid I don't agree.
B) I'm sorry, I didn't hear you.
C) In my opinion, you're right.
D) That's exactly what I think.

(Obunsha, 2002)

It is sometimes difficult to clearly distinguish between vocabulary questions and grammar questions, depending on what is actually tested. However, vocabulary questions are defined as questions that test the respondent's understanding of the meanings of vocabulary or idiomatic expressions, while grammar questions are defined as questions testing their understanding of how to use targeted grammar. Vocabulary questions include both questions on synonyms—where test takers are required to choose the answer that is most similar in meaning to the underlined word—and gap-filling questions, where test takers are required to select the correct answer to fill in the blank. This is because both types of questions are frequently presented in university entrance examinations in Japan.

Synonyms 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)

Comparing the number of vocabulary and grammar questions across the examinations, TEAP and Cambridge English do not provide questions related to grammar as multiple-choice questions. In TOEFL Institutional Testing Program (ITP)®, only structure questions from Section 2 (Structure and Written Expression) of Level 1 are included in the Exam Corpus. In contrast, Japanese university entrance examinations and EIKEN seem to have far more multiple-choice questions related to vocabulary than to grammar. Moreover, TOEIC Part 5 tends to provide more questions related to grammar than to vocabulary. Unfortunately, there are some discrepancies in the number of vocabulary and grammar questions; however, this does not present a major issue, because vocabulary and grammar questions will be examined and analysed separately in terms of their frequency and collocation.

3.5 Annotation and Sample File

In the Exam Corpus, each vocabulary and grammar question is stored as one plain text file, the same as in the UEEJ Corpus and the JUEEC. In each file, the following meta-information is added: 1) RF for a 9-digit reference number (in which the 1st–3rd digits represent the question number, the 4th digit the question skill, the 5th–8th digits the examination grade and/or year, and the 9th digit the test name), 2) TN for a test name (e.g., UEEJ, TEAP, EIKEN, TOEFL®, TOEIC®, or Cambridge English), 3) TD for test details (e.g., the names of reference books, the grades for EIKEN and Cambridge

Examination, or the year when the question was administered), 4) QS for a question skill (e.g., VC for vocabulary or GR for grammar), 5) QC for a big–small question category (e.g., gerund–verb), 6) CR for a CEFR level of the targeted vocabulary or grammar (e.g., A1, A2, B1, B2, C1, or C2), 7) IF for item facility, and 8) DI for discrimination index. The Exam Corpus also functions as an item bank. Therefore, future additions will include aspects such as item facility, indicating how difficult the item is; discrimination index, indicating how well discriminated the item is, and distractor analysis.

Below <DI> in the following Figure 2, the stem and the 3 or 4 options for each vocabulary or grammar question are presented. The missing part of each stem is labelled *zzz*, instead of brackets, as a placeholder for one word presented in collocational tables, the same as in the UEEJ and the JUEEC. Each option is prefaced with A), B), C), and D), rather than 1), 2), 3), and 4), and the correct option is prefaced with the code * to distinguish it from incorrect answers, or distractors (see Figure 2). In order to analyse questions in plain text files in the Exam Corpus, a free concordancer, LancsBox (see http://corpora.lancs.ac.uk/lancsbox/) will be used, in order to obtain wordlists, keyword lists, concordances, and collocations.

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

Figure 2. A sample file in the Exam Corpus

Furthermore, each question stored in each plain text file in the Exam Corpus is also stored in each row in a separate sub-sheet across the different kinds of examinations in one Excel file. This is so that item writers who are not familiar with analysing the Exam Corpus stored in plain text files using LancsBox can search test items using Excel. They can search items across the meta-information mentioned above such as the test names, the question skills, the question category, the CEFR levels, or the correct answer using the filter function in Excel.

3.6 Application

Once a sufficient number of vocabulary and grammar questions are stored, the Exam Corpus can be applied to language testing in particular ways. Simply put, using the filter function in Excel, item writers can search or refer to the specific questions across particular meta-information such as examination types, the targeted vocabulary and grammar, the CEFR levels, or the values of the item facility or the discrimination index annotated in the Exam Corpus.

Furthermore, the Exam Corpus can be applied to inform and validate the English proficiency examinations. Usami (2005) examined the English presented in multiple-choice vocabulary and grammar questions in Japanese university entrance examinations with respect to the collocation used in the stems, the targeted grammar, the answers, and the distractors using the UEEJ Corpus, the Longman Learner Corpus, and the British National Corpus. Following and advancing her research, the English presented in the stems, the answers, and the distractors in each English proficiency examination can be examined; for example, whether the CEFR level of the English used in the stems, the answers, and/or the distractors is appropriate to the targeted proficiency level of the examination. If the question targets CEFR A2 level, the English used in the stem, the answer, and the distractors should be A2 or below A2. In addition, the collocations presented in the stems can be examined to check whether they are appropriate and authentic, obtaining the n-grams, and for concordance or for collocation with LancsBox.

In addition, each question can be examined and improved across the meta-information annotated in the Exam Corpus. For example, the targeted vocabulary and grammar can be analysed to ascertain whether the question tests the targeted CEFR level of vocabulary and grammar. For another example, CEFR A2-level examinations should test the CEFR A2 level of vocabulary or grammar questions. In addition, frequently targeted vocabulary and grammar can be examined across the different English proficiency examinations, analysing and improving the stems and the distractors. By using the item facility and discrimination indices, questions rated as too easy or difficult or as not well discriminated can be identified for improvement.

4. Conclusion and Future Research

As one of the applications of specialised corpora to language testing, this paper has described a specialised corpus which has been created by the author, the Exam Corpus, in terms of its design, data, metadata collection, annotation, and application. The Exam Corpus was originally the UEEJ Corpus, which the author began compiling in 2004; then, more data was added to it, and it was renamed the JUEEC in 2018. Currently, it contains 1,191,850 words presented in 23,837 multiple-choice vocabulary and grammar questions. The Exam Corpus will be expanded to include more questions related to other skills such as reading, listening, and conversation, and will function as a monitor corpus. Each vocabulary and grammar question is stored as one plain text file and a line of the Excel file and analysed using LancsBox and the Excel file, respectively. Meta-information is also annotated; it is comprised of the reference number, the test name, the test details, the question skill, the question category, a CEFR level, the item facility, and the discrimination index, so that each question can be searched and analysed across the annotated information. A substantial number of vocabulary and grammar questions have already been stored in the Exam Corpus; therefore, it can be utilised by item writers to simply search and refer to the questions they want to look at. Furthermore, the Exam Corpus can be applied to inform and validate the questions, and examine and analyse the English used in the stems, the distractors, and the answers used in the examinations. Each question can also be analysed and improved across the meta-information annotated in the Exam Corpus.

For future research, the Exam Corpus will be expanded to include questions for reading, listening, and conversation skills by adding more vocabulary and grammar questions. In the future, the Exam Corpus can be applied to language testing in various ways. First, the English presented in the stems, the answers, or the distractors in each English proficiency examination can be examined and analysed in terms of whether it tests the targeted English proficiency level or its collocations reflect authentic English. Next, and more simply, new test items can be created and improved using the results of the examination and analysis.

Acknowledgements

The present study was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number (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). London: Longman.

Baker, P., Hardie, A., & McEnery, T. (2006). *A glossary of corpus linguistics*. Edinburgh: 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. (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). London: 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). Ann Arbor: 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. Princeton, NJ: ETS.

Chambers, S. (2010). What is data-driven learning? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 345-358). London: 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). London: Routledge.

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 waters. *Language Testing, 34*(4), 555-564.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London: Routledge.

Obunsha. (2001–2016). *Zenkoku daigaku nyushi mondai seikai eigo—shiritsu dai hen*. [Compilation of English past exam papers – private university exam version]. Tokyo: 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.

Scott, M. (1999). *WordSmith Tools version 3*. Oxford: Oxford University Press.

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. J*ournal 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. (2015). *The application of corpus linguistics to language testing—improving multiple choice questions*. Saarbrücken, Germany: 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.

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). London: Routledge.

Xi, X. (2017). What does corpus linguistics have to offer to language assessment? *Language Testing, 34*(4), 565-577.