# The APU Corpus of Japanese Learner English

Anthony Diaz[1]

**Abstract**

The application of corpus linguistics to the collection and study of learner-produced texts offers a wealth of possibilities for linguistic inquiry that are limited only by the inquisitiveness of the language researcher. Furthermore, the field of learner corpus research has the capacity to shed light on largely unknown aspects of the process by which language learners acquire a target language. This article will describe the design, collection, and initial analysis of a pilot corpus of writing produced by Japanese EFL learners studying in a tertiary English program in Japan. The writings that comprise the first phase of the corpus were collected in the Fall semester of 2018 and the Spring semester of 2019 at Ritsumeikan Asia Pacific University (APU) in Japan and consist of all the writing assignments that a selected group of students were required to complete throughout each semester. The primary objectives of this project are to investigate the process of language acquisition by Japanese students through the exploitation of a learners' corpus (Kennedy, 1998) and to produce a resource for investigating gaps in the students' interlanguage in order to address them in course curricula (Nesselhauf, 2004, Granger, 2002). The topics of discussion will include a description of the process of corpus design and text collection, a discussion of word frequencies in comparison with a native English corpus (LOCHNESS), and a discussion of the word coverage of the New General Service List (NGSL) and the New Academic Word List (NAWL). The findings from the coverage data of the two word lists suggest that students might benefit from curriculum that focuses on words from the NGSL that are outside of its most frequent 1000 words, and an increased emphasis on the teaching of the academic vocabulary included in the NAWL.

**Key terms:** Corpus Linguistics, Learner Corpus, Word Frequency, Louvain Corpus of Native English Essays (LOCHNESS), New General Service List (NGSL), Academic Word List (AWL)

## 1. Introduction

The main goal of the Ritsumeikan Asia Pacific University (APU) Corpus of Japanese Learner English (JLEC) is to investigate the language use of Japanese EFL students in the context of writing assignments produced for classes in the APU English Program. Motivation for the undertaking of this project came from the realization that particular mistakes Japanese students made in their writing occurred frequently across proficiency levels. This suggests that these types of mistakes are deeply ingrained in students' interlanguage and may also be influenced by their first language. It was noticing these mistakes that led to the collection of student-produced writing for linguistic analysis.

[1] English Lecturer at Ritsumeikan Asia Pacific University (APU), Beppu City, Oita, Japan. Email: adiaz@apu.ac.jp

Furthermore, aside from the use of writing textbooks, much of what teachers do in order to address student needs tends to be based on intuition and personal experience regarding what skills or knowledge students lack. Nesselhauf (2004) speaks to this idea stating that native corpora are useful for language teaching because they can reveal better than teacher intuition how a language is used by native speakers (p. 125). In addition, Nesselhauf adds that learner corpora can inform teachers of the difficulties certain groups of students have with acquiring English and assist in the process of developing materials (p. 126). While there is no doubt that personal experience and intuition are important factors in determining what to teach students, the implementation of a learner's corpus makes it possible to empirically identify issues occurring in the language acquisition process and devise program and curriculum goals which address those issues. Further support for the necessity of this kind of inquiry is that an L1 learners' corpus can contribute to the field of Second Language Acquisition by providing a snapshot of a group of students from the same L1 background which can be used for linguistic inquiry and contribute to the understanding of the process by which leaners acquire English (Granger, 2015). This article details the work done on compiling the APU JLEC until the time of the writing of this article and outlines plans for its future expansion. The article also discusses analysis of the data according to word frequencies and investigates how many words from the academic word list and New General Service List occur in the students' writings. These two lists are relevant to the context of the APU English program for the following reasons: (1) a discussion of the coverage of the NGSL has the potential to shed light on possible gaps in the general vocabulary of the students, (2) a discussion of the coverage of the NAWL may reveal to what extent students are capable of using academic vocabulary in their writing.

## 2. Description of the Corpus

Considerations for the design of the corpus were taken from Nesselhauf's (2004) criteria for the collection of learner texts to be used in a learners' corpus. These criteria include the levels of learners, the L1 of learners, type of language acquisition (instructed vs. naturalistic), and task setting (timed vs. untimed writing, use of reference tools) (p. 130). Texts for the JLEC were collected to reflect several of these criteria, namely students' L1, a description of the task types, and the proficiency levels of the students. Figure 1 illustrates the metadata based on these criteria for the JLEC as of Fall 2019.

Being that Ritsumeikan Asia Pacific University (APU) is an international university in Japan, most students enrolled in the compulsory English courses are from a Japanese L1 background. Therefore, a fundamental decision regarding the design of the JLEC was made to only collect writings that were produced by Japanese leaners of English studying in the compulsory levels of the APU English program. This decision was made so that any trends discovered in the data could possibly be explained with reference to the influence of the students' first language (i.e. Japanese). Furthermore, by having a corpus which consists of texts produced by learners from the same first language background, a method utilized in learner corpus research known as Contrastive Interlanguage Analysis (CIA), can be applied to the data (Gilquin, & Granger, 2015). CIA is a

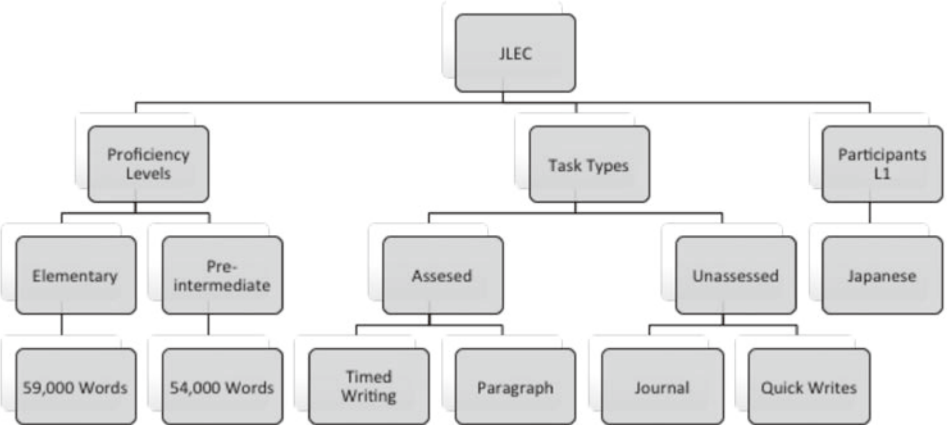

*Figure 1:* Metadata of the texts collected for the JLEC as of fall 2019

method developed for comparing L1 learner corpora with either native corpora or other L1 corpora. In order to apply CIA techniques, the JLEC was compared with a corpus of native produced texts called the Louvain Corpus of Native English Essays (LOCHNESS) compiled by the Center for English Corpus Linguistics at the University of Louvain in Belgium. The LOCHNESS corpus consists of a approximately 324,000 words and is comprised of British A level essays, British university students' essays, and American university students' essays (Center for English Corpus Linguistics, n.d.) The main objective of this analysis is to contrast word frequency data between both corpora in order to examine differences in the data and theorize possible reasons for why they occur. One important consideration regarding the two corpora is that it is necessary to mention that the JLEC does not include any essays but is a mixture of several different writing task types. Therefore, the difference in task types may have a possible influence on this study's findings. An ideal comparison would be either a Japanese learner's corpus consisting of essays that are similar to the essays found in the LOCHNESS or a native corpus of writing tasks that reflect the tasks in the JLEC more closely, but at the time of this writing no such resources were available. In addition to the discussion of word frequencies, the coverage of two word lists is examined. The two lists are the New General Service List and the New Academic Word List (Browne, Culligan, & Phillips, 2013).

Writings were collected from the elementary and pre-intermediate levels of the APU English program during the Fall 2018 and Spring 2019 semesters. In order to obtain wide coverage of student writing, all writing assignments that students produced during each semester were collected. In the case of the submission of multiple drafts, which may have been edited to address teacher feedback on mistakes, the first drafts were collected in order to obtain unrevised examples of student writing. The texts collected for the JLEC were also largely unaltered so as to preserve any data regarding mechanics or spelling errors. The writing tasks consist of several different types of assessed and unassessed tasks including paragraph assignments, timed writings, journal writings, and quick writes, which served as practice for the course's progress test component. All participants whose writings were collected for the corpus signed a release form as part of the data collection and all efforts were made to protect the identity of the writers.

The initial goal of the project was to collect 50,000 words of student-produced writing from each of the four levels of the APU core curriculum for a total size of 200,000 words. Currently, a total of approximately 113,000 words have been collected for the JLEC from the lower two levels of the compulsory English courses at APU and the collection of writing from the upper two levels of the course will be carried out in Fall 2019 and Spring 2020. All texts were produced on computer and collected electronically through Microsoft Office 365 and the website Turnitin, which is the online submission system and plagiarism checker that the university uses in its English courses.

### 3. A Discussion of Error Analysis and its Challenges

One of the most linguistically interesting applications of learner corpora is a discussion of the different kinds of errors that students make in their writing and how often they occur in the data. By categorizing the different kinds of errors made by a group of students from the same L1 and ranking their frequency it is possible to make pedagogical decisions regarding students' writing issues and possible shortcomings in the class writing curriculum. Error frequency data can also reveal information regarding the process of the development of students' interlanguage throughout the language acquisition process. Furthermore, in terms of dealing with a learners' corpus of students from one L1 background, it may be possible to make generalizations about errors that are persistent across the data or result from the influence of the L1. However, regarding learner corpus research, there are several challenges involved with developing a system for identifying and tagging learner errors. One of the most prominent issues that may be apparent to anyone who has had experience teaching writing in an EFL context is that it can be extremely difficult to identify errors in the poorly written or ungrammatical sentences that are sometimes produced by learners. This is due in part to learners' lack of grammatical control, but also due to learners' tendencies to directly translate their English sentences from their native languages. This can be especially problematic in the case of Japanese because the syntax of Japanese is so different from that of English; it is sometimes quite obvious when students are translating their writing directly from Japanese. When this occurs, sentences can contain so many errors that it is difficult to understand what the student may have been trying to say in their writing, which makes the process of categorizing student errors a time-consuming task. A further complication is that currently, most of the corpus tagging schemes that have been developed are for use with native corpora. Nagata, Whittaker & Sheinman (2011) note that there are many issues found in student writing that existing tagging schemes simply do not cover. For instance, students often make spelling, grammar, and mechanics errors which complicate the application of part of speech tags.

In their book "Language Two" Dulay, Bert & Krashen (1982) define two purposes for the study of learner errors: "(1) it provides data from which inferences about the nature of the language learning process can be made; and (2) it indicates to teachers and curriculum developers which part of the target language students have most difficulty producing correctly and which error types detract most from a learner's ability to communicate effectively (p. 138)." If a learners' corpus is to be exploited for the study of learner errors, then there is a need for an adequate system to be developed in order to classify each type of error that is found within the corpus. In reference to this, Dulay, Bert & Krashen (1982) provide two distinct methods of error taxonomy: one based on linguistic categories such as morphology and syntax, and one that focuses on how learners alter surface structures of the language, including omission of necessary items, addition of unnecessary ones, misformations and misorderings (p. 150). Dulay, Bert & Krashen state that evaluating errors from this perspective is significant because it reveals information about the "cognitive processes that underlie the learner's reconstruction of the new language (p. 150)." Granger (2003) made efforts to apply and adapt Dulay, Bert & Krashen's error taxonomies in her work to devise an error tagging system for the French learners' corpus, FRIDA. Granger developed a hybridized taxonomy that combines aspects of the linguistic categories and surface structures taxonomy that Dulay, Bert & Krashen outline. Furthermore, Granger (2003) added a third level of categorization to her system with *word category*, which includes 54 grammatical subcategories. The future proposed error categorization system for the JLEC will most likely be an adaptation of Granger's system developed for the FRIDA corpus but with special considerations and or changes in order to better suit the kinds of errors Japanese learners make in their English writing.

## 4. Analysis

The following sections discuss the analysis of the initial 113,000 words collected for the corpus. The first analysis section focuses on a comparison of word frequencies in the JLEC and a native corpus (LOCHNESS). This method of analysis provides information regarding students' over and underuse of vocabulary in relation to native English writers. The word frequency analysis and comparison with a native corpus in this article is modeled after Ringbom's (1998) chapter in *Learner English on Computer*. The second analysis section focuses on a discussion of the percentage of the JLEC that is covered by the NGSWL and NAWL word lists. This analysis provides information about areas of vocabulary that may be worthwhile for students to learn. Readers may refer to Appendix A for a full table that compares the most frequent 100 words in the JLEC and LOCHNESS.

### 4.1 JLEC Word Frequencies in Comparison with Native Corpus (LOCHNESS)

One immediate possibility for linguistic investigation using a learner corpus is examining the frequencies that words occur in the data and comparing them with a native corpus. This type of comparison provides evidence for which words in the learner corpus may be over or underused by the learners (Ringbom, 1998). In addition to providing information regarding over and underuse of words, certain conclusions can also be made about the level of students' grammar. The Louvain Corpus of Native English Essays (LOCHNESS) was used to compare word frequencies and a discussion of the findings will follow. The reason for choosing the LOCHNESS corpus is that it is freely available online, and it is a corpus of academic English produced by native English speakers from the United States and the UK. The total size of all the texts that comprise the LOCHNESS corpus is around 324,000 words compared with the approximately 113,000 words which currently make up the JLEC. However, as previously mentioned in the description section, the LOCHNESS corpus and JLEC corpus do not contain corresponding task types, therefore it is necessary to consider this in light of any findings discussed in this paper.

Due to the differing sizes of the two corpora, the word frequencies have been normalized to frequencies per 10,000 words using the formula FN = FO($10^4$)/C where FN represents the normalized frequency, FO represents the observed frequency, and C represents corpus size. Appendix A presents the 100 most frequent words per 10,000 words in the JLEC in comparison with the LOCHNESS corpus.

### 4.2 General Observations in Word Frequency data

This section will discuss the ten most frequent words in each corpus and other general observations. The most frequent words in the two corpora are *I* in the JLEC and *the* in the LOCHNESS. In contrast, *I* is ranked as the 40th most frequent word in the LOCHNESS and *the* is ranked as the 5th most frequent word in the JLEC. Out of the ten most frequent words in both corpora, eight of the words are the same but with slightly different rankings (see Table 1). The dissimilar words are *I* and *my* ranked first and eighth most frequent in the JLEC, and *that* and *be* ranked as the eighth and tenth most frequent words in the LOCHNESS. In contrast, in the JLEC, *that* is ranked at 17 and *be* is ranked 50. While the Japanese writers' overuse of the words *I* and *my* is most likely influenced by the types of writing tasks included in the JLEC, their frequencies would suggest that they tend to be overused in comparison with native writing. Conversely, Japanese writers tend to underuse the multi-functional word *that* and the verb *be* in comparison with the native writers. When inspecting the use of *that* in the JLEC with concordancing software, it is difficult to make any inferences about why this underuse is occurring, however, the apparent underuse of *that* could possibly be due to a lack of relative clauses in the Japanese students writing. In contrast to the difficulty explaining why students may be underusing *that,* when analyzing the data for examples of the use of *be* in the students' writing, there is a clear trend that can be observed. In the JLEC data 306 of the 402 occurrences of *be* consist of only a few specific constructions that use *be*. Overall, there were a total of 102 occurrences of *will be* or *would be*, 98 occurrences of *to be*, 63 occurrences of *want to be*, and 43 occurrences of *can be*. What are nearly absent from the JLEC data are constructions with other common English modals, such as *may be* with 15 occurrences, *must be* with 8 occurrences, and both *could be* and *might be* with 6 occurrences each. As one might expect, the frequency of constructions of *be* with a modal verb in the LOCHNESS is comparatively much higher. Therefore, the lack of English modals in the JLEC as an explanation for the students' underuse of *be* has some validity. Considering this observation, it is also worth mentioning that this could also be influenced by the task types included in the JLEC and LOCHNESS respectively and the tendency for native writers to use hedging when they make claims in their writing. The higher use of hedging in the native essays could explain this underuse in the JLEC, and perhaps if essays were included in the JLEC data, then the data might change.

One of the most striking differences in the word frequencies between the two corpora is the possessive pronoun *their*, which is ranked 26 in the LOCHNESS corpus and 89 in the JLEC (see Appendix A). When inspecting this facet of the Japanese learners' writings with concordancing software, almost all of the 183 instances of *their* are used in constructions that contain plural nouns related to people. In contrast, when performing the same inspection on the LOCHNESS corpus, the concordance data for *their* reveals a different use of the word not observed in the JLEC data. This additional usage of *their* by the native writers is in constructions with plural nouns unrelated to people. This would suggest that in addition to the Japanese learners' underuse of the possessive plural pronoun *their* in their writing, Japanese learners might be unaware that *their* is not only restricted to use with nouns related to people but also used so show possession of any plural noun. Another consideration is this underuse of the possessive pronoun *their* could likely be explained due to the Japanese learners' struggle with the English system of pronouns to refer to other nouns and their tendency to omit them in their writing.

Table 1:
*Top Ten Most Frequent Words in JLEC and LOCHNESS*

| **JLEC** | | | **LOCHNESS** | | |
|---|---|---|---|---|---|
| **Rank** | **Word** | **Frequency** | **Rank** | **Word** | **Frequency** |
| **1** | I | 488 | **1** | the | 651 |
| **2** | is | 355 | **2** | to | 332 |
| **3** | to | 335 | **3** | of | 331 |
| **4** | and | 258 | **4** | and | 257 |
| **5** | the | 224 | **5** | a | 211 |
| **6** | in | 191 | **6** | in | 196 |
| **7** | a | 173 | **7** | is | 194 |
| **8** | my | 149 | **8** | that | 152 |
| **9** | of | 144 | **9** | it | 99 |
| **10** | it | 133 | **10** | be | 99 |

### 4.3 Discussion of the Overuse of Words in the JLEC

After the ten most frequent words in each corpus, the word frequencies per 10,000 are consistently higher in the JLEC compared with the LOCHNESS corpus (see Appendix A). What this data indicates is that while the order that words are ranked is different between the two corpora, the majority of the 100 most frequent words in the JLEC occur at a higher frequency than the 100 most frequent words in the LOCHNESS. In some cases, this would suggest that the Japanese students tend to overuse several of the 100 most frequent words in their writing in comparison to native English-speaking writers. Table 2 displays a selection of these overused words.

Table 2:
*Overuse of Words in the JLEC*

| Word | JLEC Rank | JLEC Frequency | LOCHNESS Rank | LOCHNESS Frequency |
|---|---|---|---|---|
| **so** | 11 | 102 | 53 | 22 |
| **can** | 13 | 89 | 34 | 34 |
| **have** | 15 | 86 | 18 | 63 |
| **because** | 16 | 79 | 46 | 26 |
| **people** | 18 | 77 | 22 | 48 |
| **about** | 21 | 63 | 64 | 18 |
| **but** | 22 | 62 | 28 | 40 |
| **we** | 23 | 62 | 43 | 28 |
| **like** | 24 | 62 | 98 | 11 |
| **very** | 26 | 59 | 80 | 14 |

A possible explanation for the overuse of the words in Table 2 is the Japanese students' lack of lexical resource. For example, whereas native writers would likely have more ways of indicating cause and effect relationships in their writing (e.g. due to, as a result, since), the Japanese writers may be compensating for this lack in their overuse of the words *so* and *because*. In addition to this overuse, there are relatively few occurrences of alternate expressions which show cause and effect in the Japanese students' writing. Table 3 displays the frequencies of different cause and effect expressions across both corpora.

Table 3:
*Normalized frequencies of common cause and effect expressions in both corpora*

| Word | JLEC Frequency | LOCHNESS Frequency |
|---|---|---|
| **so** | 102 | 22 |
| **because** | 79 | 26 |
| **as** | 22 | 87 |
| **since** | 5 | 6 |
| **as a result** | 0.5 | 1.4 |
| **due to** | 0.04 | 5 |

In addition to the much higher frequency of *so* and *because* in the JLEC, as Table 3 clearly indicates, the frequencies of other cause and effect expressions are higher in the LOCHNESS. Furthermore, while the frequencies for *since* are almost the same in both corpora, upon closer examination with concordancing software, the majority of constructions that include *since* in the JLEC are used to indicate time and not cause and effect. This evidence clearly indicates that the Japanese students are overusing the words *so* and *because* in their writing to indicate cause and effect relationships.

### 4.4 Discussion of the Underuse of Words in the JLEC

In addition to evidence of the overuse of some words in the JLEC, there are also some instances where words in the LOCHNESS have a significantly higher frequency than in the JLEC. This suggests that the Japanese students are underusing these words. Table 4 displays a selection of words that are possibly underused by Japanese writers considering the corpus data.

Table 4:
*Underuse of Words in the JLEC*

| Word | JLEC Rank | JLEC Frequency | LOCHNESS Rank | LOCHNESS Frequency |
|---|---|---|---|---|
| **the** | 5 | 224 | 1 | 651 |
| **a** | 7 | 173 | 5 | 211 |
| **of** | 9 | 144 | 3 | 331 |
| **that** | 17 | 77 | 8 | 152 |
| **this** | 20 | 65 | 13 | 86 |
| **not** | 29 | 54 | 15 | 74 |
| **they** | 33 | 51 | 17 | 64 |
| **has** | 44 | 40 | 24 | 48 |
| **be** | 50 | 36 | 10 | 99 |
| **on** | 52 | 34 | 20 | 55 |
| **would** | 100 | 15 | 27 | 45 |

The underuse data from the JLEC sheds light on the gaps in students' interlanguage and can be used to make inferences about what grammar points students might benefit from the most. One of the most striking discrepancies in the underuse data is the word *would* which is ranked 27 in the LOCHNESS and 100 in the JLEC. Upon further inspection, a majority of the students' use of *would* in the JLEC occurs in the construction *would like + infinitive*. These instances account for 97 out of the 167 total occurrences of *would* in the data. In comparison, the same construction of *would like + infinitive* only accounts for 38 instances out of a total of 1461 occurrences of *would* in the LOCHNESS corpus data. This suggests that most of the Japanese students' grasp of the word *would* is limited to a single construction. Therefore, these students might benefit from learning more uses of *would*. Table 5 presents a selection of sentences from the JLEC that include *would like + infinitive* that have been unaltered and transcribed as they appear in the corpus data.

Table 5:
*Selected Sentences from the JLEC that Contain the Construction would like + infinitive*

| |
|---|
| In 2024, I ***would like to contribute*** Japan using English and international relationship that I build at APU and outside of it. |
| That's why I ***would like to cook*** dishes. |
| So, I ***would like to coutinue*** watching “ the ItteQ to the end of the world, which is my favorite TV show. |
| I ***would like to describe*** my favorite memory. |
| I ***would like to do*** my own business. |
| I ***would like to eat*** authentic Sausage with drinking beer. |
| I think that I ***would like to eat*** various Irish dishes. |

While the scope of this article prohibits a full discussion of all of the word frequency data, it is clear that the process of comparing a learner corpus to a native corpus in terms of frequency data of over or underused words, has the potential to illuminate areas of vocabulary and grammar in which students lack proficiency.

### 4.5 Coverage of the New General Service List (NGSL)

In order to reveal data regarding students' productive vocabulary knowledge along with possible gaps, the JLEC was analyzed for its coverage of two word lists. The first word list that was used for this analysis was the New General Service List (NGSL). The NGSL is a relatively new word list that was created by Browne, Culligan, & Phillips (2013). The NGSL is an effort to continue the work of Michael West with his General Service Word List published in 1953 (West, 1953). The concept behind this word list was to provide learners with a list of the most frequent English words, i.e. the words that would be of the most service to any English learner. The major difference between Brown, Culligan, & Philips' NGSL and the GSL is that the NGSL was developed using a much larger corpus of 273 million words. The list contains a total of 2368 word families and 2818 inflected word forms (lemmas) (Brown, Culligan, & Phillips, 2013). Table 6 shows the coverage of the words in the NGSL found in the JLEC. For this analysis the freeware program AntWordProfiler was used to check for coverage (Anthony, 2014).

In the table, the percentage columns refer to the percentage of words from the NGSL that make up the JLEC according to each category. The token column refers to the number of occurrences of all words including inflected forms from the NGSL in the JLEC and includes multiple occurrences. The token% column refers to the percentage of the JLEC corpus made up of all occurrences of the tokens. The type column refers to how many inflected words or word types that occur in the JLEC. For example, *accents* and *accented* are types of the word group or word family *accent*. Unlike the token column's frequency, the type's frequency only counts one occurrence of each word type. The type% column is the percentage of all the occurrences of the word types included in the type column frequency found in the JLEC.

Table 6:
*Coverage of NGSL in the JLEC*

| | **Token** | **Token%** | **Type** | **Type%** |
|---|---|---|---|---|
| NGSL (1st 1000) | 94,413 | 82.8% | 1,679 | 28.95% |
| NGSL (2nd 1000) | 5,544 | 4.86% | 848 | 14.62% |
| NGSL (3rd 800) | 2,111 | 1.85% | 383 | 6.6% |
| NGSL Total | 102,068 | 89.51% | 2,910 | 50.17% |
| Off List Words | 11,951 | 10.48% | 2,889 | 49.82% |
| **JLEC Total** | 114,019 | 100% | 5,799 | 100% |

As table 6 indicates, the first 1,000 most frequent words of the NGSL comprise 82.8% of the words in the JLEC. The second most frequent 1,000 words from the NGSL account for just 4.86% of the words in the JLEC, and the third most frequent 800 words of the NGSL account for 1.85% of the words found in the JLEC. Referring to Table 6, in total, words from the NGSL account for 89.51% of the words in the JLEC.

Table 7:
*Total NGSL Word Groups Found in JLEC Data*

| | **Group** | **Group% out of 5036** |
|---|---|---|
| NGSL (1st 1,000) | 916/1,000 | 19.17% |
| NGSL (2nd 1,000) | 644/1,000 | 13.48% |
| NGSL (3rd 800) | 329/800 | 6.89% |
| NGSL Total | 1,889 | 39.54% |
| Off List Words | 2,889 | 60.46% |
| **JLEC Total** | 5,036 | 100% |

In table 7, the group column refers to how many headwords or uninflected words from the NGSL are found in the JLEC. The group frequency only accounts for one occurrence of each headword and doesn't include subsequent occurrences. Group% accounts for the percentage of word groups in the JLEC corpus. For example, the number 916 in the first row of the group column indicates a 19.17% coverage of the total word groups in the JLEC. The data contained in table 7 provides us with information regarding the number of word groups or headwords found in the JLEC data. In reference to the group column, 916 out of 1,000 headwords in the most frequent 1,000 words of the NGSL are present in the JLEC.

Referring to table 6, words from the first 1,000 most frequent words of the NGSL make up 82.8% of the JLEC. This data suggests that overall students have productive knowledge of these words from the NGSL because they make up a large frequency of student produced texts. While task type may have an influence on the lower frequency of the less frequent 1,800 words of the NGSL, these lower frequencies might suggest that students could benefit from an emphasis on the teaching of vocabulary contained in the less frequent 1,800 words families (groups) of the NGSL.

### 4.6 The New Academic Word List (NAWL)

Because the APU English curriculum is primarily concerned with instructing students in academic English, one word list relevant to the context of the program is the New Academic Word List (Browne, Culligan, & Phillips, 2013). The NAWL is a list of 963 word families or groups which occur outside of West's (1953) first 2000 General Service List of the most common English words. These words were selected from a large corpus of 288 million words gathered from academic texts (Browne, Culligan, & Phillips, 2013). Since the data collected so far for the JLEC is from the lower two levels of the standard track of the APU English curriculum, one might expect that students lack the productive knowledge of quite a few of the words found within the NAWL. For this analysis the freeware program AntWordProfiler was also used to check for coverage (Anthony, 2014). Table 8 shows the coverage of words in the NAWL that the JLEC covers.

The data for the NAWL in Table 8 contains the same set of data as the discussion in the previous section. The token column refers to the number of occurrences of all words from the NAWL in the JLEC and the Token% column refers to the percentage of the JLEC corpus made up of those tokens. The type column refers to how many inflected words or word types that occur in the JELC. The Type% column is the percentage of word types that make up the JLEC corpus. Referring to table 8, all occurrences of words from the NAWL make up less than one percent of the words in the JLEC.

Table 8:

*Coverage of the NAWL in the JLEC*

| | **Token** | **Token%** | **Type** | **Type%** |
|---|---|---|---|---|
| NAWL | 1,085 | 0.95% | 200 | 3.45% |
| Off List Words | 112,934 | 99.05% | 5,599 | 96.55% |
| **JLEC Total** | 114,019 | 100% | 5,799 | 100% |

In table 9, the group column refers to how many headwords or uninflected words from the NAWL are found in the JLEC and group% is the percentage of word groups that make up the total word groups in the JLEC. This table indicates that out of the 963 word groups (headwords) contained in the NAWL, only 177 are found in the JLEC data.

Table 9:

*Total NAWL Word Groups Found in JLEC Data*

| | **Group** | **Group% of 5776** |
|---|---|---|
| NAWL | 177/963 | 3.06% |
| Off List Words | 5,599 | 96.94% |
| **JLEC Total** | 5,776 | 100% |

While the level of students and task types most likely have an influence on the lack of academic vocabulary in the students' writing, the data in Tables 8 and 9 would still suggest that if one of the aims of the APU English program is to teach students academic English in order to prepare them to enroll in English medium lecture courses, then a greater emphasis should be placed on the teaching of academic vocabulary to students in the APU English program. However, there is one crucial piece of information that is impossible to surmise from a learners' corpus and that is gauging the students' receptive knowledge of the vocabulary contained in the NAWL. One possible way to mitigate the lack of this data is to combine the learners' corpus data with a measure that tests for the students' receptive knowledge of the NAWL words. This could be done in the form of a vocabulary test given to students studying in the various levels of the APU English program and might be of interest for future research into this area.

## 5. Limitations

While there are many possible applications for a learners' corpus such as the JLEC, there are also limitations. Considering discussions of the breadth of vocabulary in student-produced writing, the evaluation of a learners' corpus is only able to provide information about students' productive knowledge of vocabulary and receptive vocabulary skills require additional measures to assess. Furthermore, it could be argued that receptive knowledge of academic vocabulary is enough for most Japanese students in the APU English program and an emphasis on the productive knowledge of vocabulary would provide minimal benefits for students.

Another limitation of the current study is the proficiency levels of students. Presently, the writings that comprise the JLEC were only collected from the Elementary and Pre-intermediate levels of the APU English program. This has the potential to influence findings and may not be an accurate representation of student ability across all levels of the APU English program. One possible way to mitigate this is to add writings from the other levels of the course and to have the ability to compare JLEC texts across the program's proficiency levels.

Task types are also a possible limitation on the study due to the less formal writing tasks, such as journal writings, that were included in the JLEC. One possible solution to this issue is the exclusion the journal writing tasks from the JLEC data or balancing the LOCHNESS with similar native texts. If this step were to be taken, it might provide a more accurate representation of the over and underuse of vocabulary in the students' writing. Nevertheless, based on the discussion of findings in the previous sections of this article, it is possible to make some inferences about the language use of Japanese students based on the data in the JLEC.

## 6. Conclusion

This article has detailed information regarding how this project was undertaken and described the process by which the corpus was planned and collected. The analysis of the word frequencies between the JLEC and a native corpus (LOCHNESS) provided evidence for over and underuse of words in the student-produced writings. With this data, it is possible to make inferences about the sequencing and scope of specific grammar points in which students would benefit from receiving further instruction. An examination of the word frequency and concordance data reveals that there are several grammatical constructions that students could benefit from. Considering this data, Japanese students tend to underuse several words including *their*, *that*, and *be*. This could have possible implications for the sequencing or focus of grammar points that are covered in course curriculum, specifically an increased emphasis on the English system of pronouns, relative clauses, and teaching the use of modals at an earlier stage in the curriculum. Furthermore, the data regarding the overuse of words, such as *so* and *because,* suggests that students lack the lexical resources for expressing cause and effect in their writing, and since this is typically an important skill in writing, a consideration of teaching students more ways to express cause and effect in their writing may be necessary.

The discussion of the coverage of the NGSL and NAWL provided information about potential areas for the development of vocabulary curriculum, namely the less frequent 1800 words of the NGSL and the majority of the NAWL. In addition, the coverage data could be combined with a measure of students' receptive knowledge of vocabulary found in each word list to bolster understanding of students' weak areas. As previously mentioned in the article, the JLEC is an ongoing project that includes plans to add writings from the upper two levels of the compulsory English courses in APU's English program and to devise a system for error tagging the corpus so student-produced errors can be categorized and evaluated for frequency. In addition to these future plans, one hope for this project is to make the data available online for other researchers to access and to serve as a contribution to the field of Second Language Acquisition by providing a resource of the process by which Japanese English learners develop their interlanguage.

## References

Anthony, L. (2014). AntWordProfiler (Version 1.4.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Anthony, L. (2018). AntConc (Version 3.5.7) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software

Browne, C., Culligan, B. & Phillips, J. (2013). The New Academic Word List. Retrieved from http://www.newgeneralservicelist.org

Browne, C., Culligan, B. & Phillips, J. (2013). The New General Service List. Retrieved from http://www.newgeneralservicelist.org

Centre for English Corpus Linguistics (CECL) (n.d.). The Louvain Corpus of Native English Essays (LOCNESS). Université catholique de Louvain, Belgium. Retrieved from https://www.learnercorpusassociation.org/resources/tools/locness-corpus/

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, *34*(2), 213. https://doi.org/10.2307/3587951

Dulay, H., Burt, M., & Krashen, S. (1982). *Language Two.* New York: Oxford University Press.

Gilquin, G., & Granger, S. (2015). Learner language. In D. Biber, R. Reppen (Eds.), *The Cambridge* *handbook of English corpus linguistics* (pp. 418-435). Cambridge, United Kingdom: Cambridge University Press. https://doi.org/10.1017/CBO9781139764377.024

Granger, S. (2002). A bird's Eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign Language teaching* (pp. 3-33). Philadelphia, PA: John Benjamins Publishing Company. https://doi.org/10.1075/lllt.6.04gra

Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy. CALICO Journal. 20. 465-480.

Granger, S. (2015). The contribution of learner corpora to reference and instructional materials design. In Granger, S., Gilquin, G. & Meunier, F. (eds.) *The Cambridge handbook of learner corpus research* (pp. 486-510). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.022

Kennedy, G. (1998). *An introduction to corpus linguistics.* New York, NY: Addison Wesley Longman Limited.

Nagata, R., Whittaker, E., & Sheinman, V. (2011). Creating a manually error-tagged and shallow-parsed learner corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 1210–1219). Portland, OR: Association for Computational Linguistics.

Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In J. M. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 125-152). Philadelphia, PA: John Benjamins Publishing Company. https://doi.org/10.1075/scl.12.11nes

Ringbom, H. (1998). Vocabulary Frequencies in Advanced Learner English: A cross-linguistic approach. In S. Granger (Ed.), *Learner English on computer* (pp. 172-185). New York, NY: Longman.

West, M. (1953). A general service list of English words. London: Longman, Green & Co.

***Appendix A***

*Comparison of the 100 most frequent words per 10,000 words in both corpora*

| JLEC | | | LOCHNESS | | |
|---|---|---|---|---|---|
| **Rank** | **Word** | **Frequency** | **Rank** | **Word** | **Frequency** |
| **1** | I | 488 | **1** | the | 651 |
| **2** | is | 355 | **2** | to | 332 |
| **3** | to | 335 | **3** | of | 331 |
| **4** | and | 258 | **4** | and | 257 |
| **5** | the | 224 | **5** | a | 211 |
| **6** | in | 191 | **6** | in | 196 |
| **7** | a | 173 | **7** | is | 194 |
| **8** | my | 149 | **8** | that | 152 |
| **9** | of | 144 | **9** | it | 99 |
| **10** | it | 133 | **10** | be | 99 |
| **11** | so | 102 | **11** | for | 97 |
| **12** | for | 99 | **12** | as | 87 |
| **13** | can | 89 | **13** | this | 86 |
| **14** | are | 88 | **14** | are | 79 |
| **15** | have | 86 | **15** | not | 74 |
| **16** | because | 79 | **16** | he | 67 |
| **17** | that | 77 | **17** | they | 64 |
| **18** | people | 77 | **18** | have | 63 |
| **19** | want | 66 | **19** | with | 59 |
| **20** | this | 65 | **20** | on | 55 |
| **21** | about | 63 | **21** | by | 52 |
| **22** | but | 62 | **22** | people | 48 |
| **23** | we | 62 | **23** | his | 48 |
| **24** | like | 62 | **24** | has | 48 |
| **25** | he | 59 | **25** | was | 48 |
| **26** | very | 59 | **26** | their | 47 |
| **27** | English | 57 | **27** | would | 45 |
| **28** | think | 55 | **28** | but | 40 |
| **29** | not | 54 | **29** | or | 40 |
| **30** | there | 53 | **30** | an | 38 |
| **31** | with | 53 | **31** | from | 37 |
| **32** | was | 51 | **32** | more | 36 |
| **33** | they | 51 | **33** | which | 35 |
| **34** | if | 48 | **34** | can | 34 |
| **35** | time | 47 | **35** | will | 34 |
| **36** | me | 46 | **36** | there | 33 |
| **37** | lot | 45 | **37** | if | 33 |
| **38** | many | 43 | **38** | one | 33 |
| **39** | you | 43 | **39** | at | 32 |
| **40** | she | 42 | **40** | I | 30 |
| **41** | when | 42 | **41** | all | 29 |
| **42** | at | 41 | **42** | many | 29 |

| | | | | | |
|---|---|---|---|---|---|
| **43** | good | 40 | **43** | we | 28 |
| **44** | has | 40 | **44** | who | 28 |
| **45** | job | 39 | **45** | also | 27 |
| **46** | festival | 39 | **46** | because | 26 |
| **47** | do | 39 | **47** | when | 26 |
| **48** | will | 38 | **48** | these | 26 |
| **49** | go | 37 | **49** | been | 24 |
| **50** | be | 36 | **50** | should | 24 |
| **51** | eat | 36 | **51** | only | 23 |
| **52** | on | 34 | **52** | were | 23 |
| **53** | food | 34 | **53** | so | 22 |
| **54** | friends | 33 | **54** | other | 22 |
| **55** | first | 33 | **55** | do | 22 |
| **56** | work | 31 | **56** | what | 21 |
| **57** | from | 31 | **57** | life | 21 |
| **58** | also | 31 | **58** | no | 21 |
| **59** | use | 30 | **59** | could | 20 |
| **60** | am | 27 | **60** | however | 18 |
| **61** | school | 27 | **61** | had | 18 |
| **62** | get | 26 | **62** | our | 18 |
| **63** | class | 26 | **63** | them | 18 |
| **64** | two | 25 | **64** | about | 18 |
| **65** | more | 25 | **65** | being | 18 |
| **66** | however | 25 | **66** | out | 17 |
| **67** | don't | 25 | **67** | some | 17 |
| **68** | make | 24 | **68** | you | 17 |
| **69** | by | 23 | **69** | such | 16 |
| **70** | than | 23 | **70** | time | 15 |
| **71** | as | 22 | **71** | her | 15 |
| **72** | other | 22 | **72** | into | 15 |
| **73** | his | 22 | **73** | may | 15 |
| **74** | second | 21 | **74** | up | 15 |
| **75** | important | 20 | **75** | than | 15 |
| **76** | money | 20 | **76** | way | 15 |
| **77** | example | 20 | **77** | then | 14 |
| **78** | her | 20 | **78** | even | 14 |
| **79** | or | 19 | **79** | most | 14 |
| **80** | high | 19 | **80** | very | 14 |
| **81** | day | 19 | **81** | its | 13 |
| **82** | every | 18 | **82** | she | 13 |
| **83** | most | 17 | **83** | how | 13 |
| **84** | one | 17 | **84** | make | 13 |
| **85** | them | 17 | **85** | world | 13 |
| **86** | Japanese | 16 | **86** | society | 13 |
| **87** | some | 16 | **87** | us | 13 |
| **88** | things | 16 | **88** | does | 13 |

| | | | | | |
|---|---|---|---|---|---|
| **89** | their | 16 | **89** | him | 13 |
| **90** | fun | 16 | **90** | money | 12 |
| **91** | who | 16 | **91** | any | 12 |
| **92** | meat | 16 | **92** | children | 12 |
| **93** | much | 16 | **93** | women | 11 |
| **94** | family | 15 | **94** | much | 11 |
| **95** | all | 15 | **95** | use | 11 |
| **96** | famous | 15 | **96** | just | 11 |
| **97** | talk | 15 | **97** | own | 11 |
| **98** | both | 15 | **98** | like | 11 |
| **99** | study | 15 | **99** | over | 10 |
| **100** | would | 15 | **100** | made | 10 |