Developing a Pilot Learner Corpus: New Directions in Writing Instruction

Anthony Diaz[1]

Abstract

The collection of learners' writings or speech for comprehensive analysis can serve as an indispensable resource for a language program and shed light on the areas in which students are excelling or struggling (Nesselhauf, 2004). Furthermore, when successfully implemented, a learners' corpus can provide a source of insight for the content and sequencing of material being taught to students, or as a consciousness-raising tool for teachers, affording them a better grasp of the language use of their students (Kennedy, 1998, p. 281). A learners' corpus can also highlight which features of the target language students tend to over or underuse (Leech, 1998). The aims of this article are threefold: to refer to relevant literature in regard to the designing and implementation of learner corpora, to detail how the benefits of this kind of inquiry can be applied to not only inform but also enhance current pedagogical practices for tertiary EFL programs such as that at Ritsumeikan Asia Pacific University (APU), and finally, to outline a corpus design that can be implemented to evaluate the writing of students studying in the APU English program.

Key terms: Corpus Linguistics, Learner Corpus, Learner Corpora, First Language (L1), EFL, Word Frequency, Learner Errors, Error Analysis, Cross-Linguistic Influence, Contrastive Interlanguage Analysis (CIA)

1. Introduction

Sinclair (1991, p.171) describes a corpus as "a collection of naturally occurring language text, chosen to characterize a state or variety of language". A corpus can consist of written or spoken text gathered from a variety of contexts. While several corpora focus on spoken text, the majority of corpora are comprised of written text (De Cock, 2010; Granger, 2004). Until the present date, corpora have been made for a range of general and specific purposes, some of which include investigating the language used by certain discourse communities, analyzing texts in specific academic fields, such as science and engineering, and evaluating the process of language acquisition in non-native speakers (Kennedy, 1998). One would be hard pressed to find a contemporary English textbook or dictionary that has not been influenced at least in part by corpus linguistics (Hunston, 2002). This influence can range from which words are included in textbooks according to their frequency of use to what word collocations are the most useful to cover in a new dictionary for non-native speakers. As technology progresses and access to it improves, the ability to use a corpus to examine various aspects of language and its acquisition has shifted from a practice that was once open only to a few knowledgeable and dedicated linguists to one that can be carried out by any

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

motivated researcher (Hunston, 2002). Moreover, as technology continues to become more prevalent in daily life, the use of corpus linguistic methods to explore language acquisition should be expected to increase in the realm of language research and teaching. The endeavor of compiling and analyzing a corpus, which used to be an extremely labor-intensive task, has become a process that any academic with access to a computer can do in a fraction of the time that it used to take (Kennedy, 1998; Tono, 2000). A particularly promising avenue of corpus research lies in its application to the collection and study of learner-generated text, also known as learner corpora. When applied to the context of an English program, the richness of information that a learners' corpus can provide to the language instructor and or researcher can be utilized to pinpoint the specific needs of students and create more relevant materials and curricula (Granger, 2002). The aim of the current article is to propose the creation of a Ritsumeikan Asia Pacific University learners' corpus with the primary goals of collecting examples of the kinds of errors Japanese students make in their writings and as a resource for instructors to be better informed about the areas of students' writing that need special instruction.

2. Learner Corpora

A commonly-held belief among linguists is that linguistics should be concerned with the realities of language acquisition and describing those phenomena rather than prescribing what language should be. Yallop (2004) affirms that "A common slogan of linguists [is] that linguistics is descriptive not prescriptive" (p. 36), and "There is an indisputable obligation to aim to describe what is there, rather than to describe what you would like to be there [in language]" (p. 36). In keeping with this tradition, compiling a learners' corpus is a method to describe the interlanguage of learners as it exists in reality and does not treat language ability as an idealized construct, which may or may not be prescribed by institutional standards or other external means. Learner corpora are a specific application of corpus linguistics and can be defined as "electronic collections of spoken or written texts produced by foreign or second-language learners" (Granger, 2004, p. 124). In addition, Granger (2004) points out that while the focus on learner corpora is a relatively new development, the collection of English text for the purpose of investigation has been in practice since the founding of the discipline of SLA. The primary objective of a learners' corpus is "to contribute to a better understanding of the second-language acquisition process" (Granger, 2015 p. 486). By examining learner-produced texts, it is possible to acquire evidence of how the interlanguage of a group of learners is developing. It is also possible to observe which specific aspects of language they are excelling in and which they are struggling to acquire. This in turn can lead to the development of materials that are specifically tailored to address the problems found within a specific group of L1 learners as well as the possibility of comparing groups of learners from different L1 backgrounds (Mendiokoetxea, Murcia Bielsa & Rollinson, 2010; Gilquin & Granger, 2015).

3. Methods and Applications of Learner Corpora

According to the literature concerning learner corpora research, there are three methods developed specifically for analyzing learners' corpus data. These three methods include contrastive interlanguage analysis (CIA), the integrated contrastive model, and computer-aided error analysis (Gilquin, & Granger, 2015). CIA consists of two procedures: comparing a learners' corpus with a native English corpus and comparing a learners' corpus with different learner corpora produced by learners from different L1 backgrounds (Gilquin, & Granger, 2015). This type of comparison can be useful for accentuating differences between the language of non-native speakers and natives. This can lead researchers to discover areas of their students' interlanguage that are fossilized. Furthermore, by comparing learners' corpus data of different groups of learners, for example Japanese L1 speakers and Spanish L1 speakers, evidence of errors directly related to a learner's L1 can be observed and errors that are a result of L1 transfer can be identified. A belief held by a number of language researchers is that learner corpora have a particularly important role to play in examining the process by which Japanese L1 learners acquire English. Tono, Kaneko, Isahara, Saiga, and Izumi (2001) point out that it is necessary to document how Japanese learners acquire English. This is particularly important to the authors because they argue that when comparing Japanese learners to those from other L1 backgrounds, the Japanese learners tend to lack communicative competence in English despite studying the language for a period of six years in the Japanese education system. Other work carried out by Tono (2000) compared the order of acquisition of grammatical morphemes by Japanese students with the order of acquisition observed in native English-speaking children as outlined by Brown (1973). Tono discovered that Japanese L1 students acquire the grammatical morphemes in a different order, suggesting that further study comparing corpus data from different groups of L1 learners should be explored in order to uncover more information about the interlanguage of Japanese learners. Next, similar to the CIA method, the integrated contrastive model is also concerned with comparison (Gilquin, & Granger, 2015). This method compares learner corpus data with a corpus from the group of learners' native language in order to "get a clear picture of the characteristics of the learner's L1 and the differences it presents with the target language" (Gilquin, & Granger, 2015, p. 426). This information can then be used to predict what kind of transfer errors will occur in the learners' output and attempt to remedy them (Gilquin, & Granger, 2015). The last method is known as computer-aided error analysis. This method involves identifying errors in learner corpus data and thoroughly tagging them using a "system of error tags" (Gilquin, & Granger, 2015, p. 427). Computer-aided error analysis can be applied to small sets of data and has the potential to be used in exploratory learning. By having students examine their own writings in a corpus format, it is possible to raise their awareness of exactly what kinds of mistakes they are making. Furthermore, teachers can use the data from the class to develop material that addresses the most frequent errors of their students and devote class time to the most salient mistakes. As the initial aim of this project was to quantify the most frequent errors that the Japanese learners make at the university, computer-aided error analysis will be the primary focus of the pilot APU learners' corpus.

4. The Proposed APU Learners' Corpus

4.1 Design

In the planning stages of a learner's corpus, it is important to have a clear idea of the design and what information will be included along with the texts. The more information included with a learners' corpus, the more aspects of the learners' language acquisition one can investigate. Nesselhauf (2004) outlines some specific 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, etc.) (p. 130).

The APU learners' corpus will have two main objectives: to collect data on the most prevalent learner errors in the Japanese student population and to apply those findings to the creation of materials, specifically for writing. Kaszubski (1998) argues that writing textbooks for EFL/ESL students are not usually suited to different populations of students since there is typically little room given to stylistic features of writing. He claims that textbooks on writing should be enhanced with learner corpus data that is tailored to a specific group of students. Kaszubski (1998) adds that there are two major ways learner corpus data can supplement writing material: by providing frequency information on overuse/underuse/misuse, and by supplying new content (e.g., collocations and phraseology) (p. 180). A pilot corpus, which focuses on several classes in the lower levels of the APU English program, will be the foundation for the corpus, and the scope of the project will be expanded from the initial findings. Student-produced writings will be collected electronically, and a corpus will be compiled to look at salient errors in the following areas: collocations, the overuse/underuse/misuse of words, lexical and grammatical errors. The data will be tagged with a part of speech tagger, such as TagAnt (Anthony, 2015), and a system of error tags will be added to the data. These tags will focus on some of the specific errors Japanese learners are prone to make which include inconsistent pronoun use, subject-verb agreement, underuse of present perfect tense, misuse of passive tense, and misuse of vocabulary. This information will be used to create sample materials that focus on these areas. The free program AntConc (Anthony, 2018) will be used to analyze the data and investigate the frequency of which tagged errors occur in the data. A total of 50,000 words will be the initial target size to be collected by the end of the Spring academic semester at the university. This number was chosen because it is a fourth of the size of the monolingual sub-corpora which make up the International Corpus of Learner English (Granger, 2003) and because it is a reasonable target for the pilot. Data collection will be limited to writing produced by Japanese L1 students in order to keep the data consistent. The rationale behind this decision is that in order to better serve the largely Japanese population studying in the English program at APU, it is necessary to explore their particular habits in writing.

4.2 Applications

The APU learners' corpus can be utilized in a number of ways. The first possible use is to categorize the most frequent errors in students' writing. By utilizing this process, a profile of the learners can

be made that includes information about the most frequent errors in the learners' writing. Once prepared, this information can be made available to all instructors in the program and serve as a form of faculty development. A similar study was conducted by Díaz-Negrillo and Valera (2010) when they analyzed and error-tagged a Spanish L1 learners' corpus for errors. They discovered several categories of errors that were persistent across the corpus, with comma use being the most widespread (p. 82). The students were also found to have issues with vocabulary use, which included selecting the wrong nouns for corresponding verbs in relation to their meaning (p. 82). Mendikoetxea, Murcia Bielsa, and Rollinson (2010) outline how they compiled a small corpus of their students' writings and subsequently developed activities based on errors found in the data. This same principle could be applied to the APU learners' corpus as a source of teaching materials that focus on the errors that the predominant group of Japanese L1 speakers make in their writing. Next, CIA can be utilized to examine the difference in word frequencies when compared with a native corpus. As previously stated, this method involves comparing learner language with native corpora in order to note differences or inconsistencies. Gilquin and Granger (2015) elaborate that "The exploitation of a native corpus, used in combination with the learner corpora, makes it possible to see how the learner data are situated in relation to a certain reference norm" and "the inclusion of the L1 variable gives a glimpse of the possible influence of the mother tongue" (p. 434). Data gathered from this can shed light on the differences between the interlanguage of the Japanese L1 students and that of native speaker interlanguage. A further application of CIA is to compare the word frequencies of the learners' corpus with a word list, such as the Academic Word List, to examine the variety of words the students are capable of using and which ones they are lacking. Finally, the use of collocations, phrasal verbs, and verbs that require specific prepositions can be examined in student-produced texts.

5. Limitations

While learner corpus research is an exciting prospect for the discovery of new information regarding the language acquisition process, it should not be seen as a miracle or catch-all strategy. For example, trends that are discovered in a certain population of L1 speakers are not necessarily representative of all learners from the same L1 background. Furthermore, there are potentially many factors to take into account that can have an effect on how students perform on the types of texts collected for a corpus, such as the motivation level of the students, the pressure of time constraints, the extent of control over the writing task, and the access to dictionaries or the internet. Another aspect to consider is the fact that many of the discoveries that have come from learners' corpus research have resulted from a researcher's initial hypothesis or intuition. It can be difficult to utilize a learners' corpus without having a general idea of what one is looking for in the data, which can lead to confirmation bias in the results (i.e., Researchers may look for what they want or expect to find in the data). However, despite these considerations, the empirical nature of corpus-based research is, at the very least, ample reason to collect as much data as possible in order to contribute to the burgeoning field of learner corpus research.

6. Conclusion

It is likely that corpus linguistics will have an increasingly important role to play in the future direction of SLA and ELT research. In particular, the specialized field of learner corpus research is ripe with potential to have pedagogical as well as theoretical influence on how curricula are designed and implemented. When applied to the context of an academic English program, a learners' corpus can be used as a source of material that targets deficiencies in learners' interlanguage and can help shape future university curricula.

References

Anthony, L. (2015). TagAnt (Version 1.1.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://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

Brown, R. (1973). A first language: The early stages. Cambridge, MA: Harvard University Press.

Carlson, C. (2012). Proficiency level—A fuzzy variable in computer learner corpora. *Applied Linguistics*, 33(2), 161-183.

De Cock, S. (2010). Spoken learner corpora and EFL teaching. In M. Campoy-Cubillo, B. Bellés-Fortuño, & M. Gea-Valor (Eds.), *Corpus based approaches to English language teaching* (pp. 123-137). New York, NY: Continuum International Publishing Group.

Díaz-Negrillo, A., & Valera, S. (2010). A learner corpus-based study on error associations. *Procedia Social and Behavioral Sciences*, (3), 72-82.

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.

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.

Granger, S. (2003). The international corpus of learner English: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly*, 37(3), 538-546.

Granger, S. (2004). Computer learner corpus research: Current status and future prospects. In Connor, U & Upton, T. A. (Eds.), *Applied corpus linguistics: A multidimensional perspective* (pp. 123-146). New York, NY: Rodopi.

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.

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

Kaszubski, P. (1998). Enhancing a writing textbook: A national perspective. In S. Granger (Ed.), *Learner English on computer* (pp. 172-185). New York, NY: Longman.

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

Leech, G. (1998). Preface. In S. Granger (Ed.), *Learner English on computer* (pp. Xiv-Xx). New York, NY: Addison Wesley Longman.

Mendiokoetxea, A., Murcia Bielsa, S., & Rollinson, P. (2010). Focus on errors: Learner corpora as pedagogical tools. In M. Campoy-Cubillo, B. Bellés-Fortuño, & M. Gea-Valor (Eds.), *Corpus-based approaches to English language teaching* (pp. 180-194). New York, NY: Continuum International Publishing Group.

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.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford Univ. Press.

Tono, Y. (2000). A computer learner corpus-based analysis of the acquisition order of English grammatical morphemes. In L. Burnard & T. McEnery, (Eds.), *Rethinking language pedagogy from a corpus perspective: Papers from the Third International Conference on Teaching and Language Corpora* (pp. 123-132). New York, NY: Peter Lang.

Tono, Y., Kaneko, T., Isahara, H., Saiga, T., Izumi, E., Narita, M., & Kaneko, E. (2001). The standard speaking test (SST) corpus: A 1 million-word spoken corpus of Japanese learners of English and its implications for L2 lexicography. In S. Lee, (Ed.), *ASIALEX 2001 Proceedings: Asian Bilingualism and the Dictionary: The Second Asialex International Congress* (pp. 8-10). Korea: Yonsei University. doi:10.1.1.120.5827&rep=rep1&type=pdf

Yallop, C. (2004). Words and meaning. In *Lexicology and corpus linguistics: An introduction* (pp. 23-71). New York, NY: Continuum.