EPiC Series in Language and Linguistics

Volume 1, 2016, Pages 79–87

CILC2016. 8th International
Conference on Corpus Linguistics

An Introduction to Using Corpora with EFL Learners

Mark Donnellan
Kwansei Gakuin University, Nishinomiya, Japan
mjdonnellan@kwansei@ac.jp

Abstract

The use of corpora in the classroom represents an innovative way to enable English language learners to undertake independent study of lexical and grammatical patterns; however, only a limited amount of investigation into the use of corpora with students exists. This paper will first briefly introduce pertinent literature, which will give a basic overview of corpus linguistics. The paper will report on the use of the British National Corpus (BNC) and other corpus tools with students in a semester long course in a Japanese university with advanced EFL learners. These students undertook a series of tasks and projects, which allowed them to achieve the overall course goal of being able to conduct independent research into lexical and grammatical patterns. In order to assess student progress and to gather student opinions about the course and the use of corpora, data was collected in two ways: pre and post CEFR style student self-assessments, and a course reflection and evaluation survey. The results indicated that the students had progressed throughout the semester and that they had a largely positive opinion of the course. In concluding the paper, suggestions for teachers wishing to use corpora and corpus tools with their students will be offered.

1 Introduction

While the field of corpus linguistics is by no means a new area of study, it has gained prominence in recent years. This can be attributed to two main factors: 1) advances in technology which allow for the collection of larger amounts of text and for analysis using various kinds of software, 2) advances in research which have recognized the importance of the study of lexis. With these advances it has become more popular and practical to use corpora with EFL learners. This paper reports on the use of corpora in with advanced EFL learners in a semester long course at a Japanese university. The overall goal of the course was to enable the students to conduct independent research and investigation into lexical and grammatical patterns using corpora. The corpus that students use for this course was the British National Corpus (BNC). Specifically, this paper aims to investigate two research questions:

A.Moreno Ortiz and C.Pérez-Hernández (eds.), CILC2016 (EPiC Series in Language and Linguistics, vol. 1), pp. 79–87

1. To what extent did students improve their practical and theoretical knowledge of corpus linguistics as a result of their participation in this course?

2. At the end of this course, what were the students' opinions about the use of corpora and about the specific components of the course?

The next section of this paper will review the literature. Following on from that, an overview of the course that was conducted will be given. Next, the survey data that was collected will be described and the results will be presented. This paper will conclude by commenting on the data collected and offering suggestions for teachers wishing to undertake similar courses with their students.

2 Literature Review

2.1 Overview of Corpus

A corpus is a collection of text, either written, spoken or both, which is gathered to reflect the state of a language in general or in in specific socio-pragmatic situation. Flowerdew lists four defining features of a corpus: it consists of authentic data; it has explicit design criteria; it is representative of a particular language or culture; it is designed for a specific linguistic or socio-pragmatic purpose (2012, p. 3). The authentic text contained within a corpus combined with the fact that a corpus can be built for specific purposes make corpus a powerful tool for studying and researching lexis, grammar and other features of language.

In terms of the benefits of corpus, it can be a significant help in alleviating student frustration with teacher explanations such as 'it just sounds better' (Hunston, 2002, p.20). This 'it just sounds better' rationale could more specifically be referred to as native speaker intuition. Hunston isolates four areas in which native speaker intuition may not suffice and asserts that corpus is more reliable in these areas. These areas are: judgments about collocations; judgments about frequency; making semantic prosody (see Louw, 1993) and pragmatic meaning clear; and revealing details of phraseology (2002, p.20-22). Put simply, using corpus data allows for in-depth investigation and research of lexis including collocations, phraseology and meaning.

Historically, corpus does have limitations though, most significantly is that many of the corpora that have been built avoid dealing with spoken language (Sinclair, 1991) since it is significantly easier to gather written text than spoken text. Advances in technology have negated this problem to a fair extent, but it does still exist. In addition to this significant limitation, Hunston also lists four. Firstly, corpus only tell us if something is frequent, it will not tell us if it is possible, for that we must use intuition. Secondly, a corpus can only show its own contents and can only be "treated as deductions and not facts". That being said, general corpora of English such as the BNC can claim to be representative of English or at least of British English. Thirdly, a corpus only offers evidence, but this evidence still needs to be interpreted by the researcher. Finally, a corpus presents language out of context because it does not present features such as intonation and body language (2002, p.22-23). While these limitations are not insignificant, corpus data combined with the researcher's interpretation can provide significant insights and research into language.

2.2 Use of Corpora with Students

Beyond teacher or researcher investigations, corpora can also be used in the classroom with students (Flowerdew, 2012; Sripicharn, 2010; Hunston, 2002). There are several possible areas that students can investigate including hypothesis testing, error correction, learning about genre, self-study in specialized areas and contrastive studies (Sripicharn, 2010, p. 372-374). The option of doctoring concordance lines to produce teaching materials to train students can be a useful way to get students used to looking at concordance lines (Sripicharn, 2010, p. 380). These doctored concordance lines can be accompanied by questions or consciousness raising activities. When it comes to actual interpreting of concordance lines though, Sripicharn cautions against altering the concordance lines saying that students need to deal with ‘messy’ lines, which are irrelevant. More specifically, these doctored concordance lines might be described as data-driven-learning. The term data-driven learning (DDL) was coined by Johns (1990) and refers to students using concordance lines from a corpus to conduct their own independent investigations, as Johns puts it, "the language learner is also, essentially, a research worker whose learning needs to be driven by access to linguistic data (p. 2)". This kind of investigation is at the heart of what the course discussed in this paper set out to achieve. Römier (2010) subdivides DDL into computer-based (hands-on) where the students use computers to directly search corpora, and paper-based (hands-off) where teachers search corpora and prepare materials for students, with the overall goal stated in the introduction of having the students conduct independent research, a hands-on approach was preferred for this course.

3 Overview of the Course

The course in question was conducted with three classes of undergraduate advanced EFL learners at a Japanese university. The first group was comprised of five students, the second group consisted of nine students, and the third group consisted of eight students. The students' TOEIC scores ranged from 625-930, and this cross faculty course included students from 11 faculties with major such as Science and Technology, Literature, and Economics. The course was designated as a content-based course in the broad general area of language and communication, and within this broad general area the instructor had the freedom to choose a subtitle/specific subject area, in this corpus linguistics. Thus, the course title became *Language and Communication - An Introduction to Corpus Linguistics*. Ninety-minute classes were held once a week for fourteen weeks. One group used a classroom equipped with desktop computers using Windows 7, the other two groups used MacBook Pro computers, and neither group encountered any issues in using the corpus tools with either OS X or Windows. The course consisted of five main components.

3.1 Theory

In weeks 1 and 2 of the course, the teacher gave lectures to familiarize the students with key concepts and principles of corpus linguistics. Subsequent classes included small amounts of theory as appropriate.

3.2 Corpus Tasks

A series of tasks were designed to give the students hands on and practice with the BNC. The University of Lancaster's version of the BNC, BNCweb (Hoffmann & Evert, 2013) was used since it provided a free and user-friendly interface. Three sample tasks were:

1. Hoffmann, Evert, Smith, Lee and Belglund Prytz's investigation of the word *wicked* (2008, p.47-65). In this investigation, the investigator searches the corpus for instances of *wicked* in the non traditional (negative) sense.
2. Error correction passages designed with errors chosen from An A-Z of common English Errors for Japanese Learners (Barker, 2010).
3. Contrastive analysis tasks designed to highlight differences in collocations between Japanese, the students' L1 and English. Nation proposes that *take medicine* may often be difficult for students to predict (2001, p. 328) and Donnellan confirms that this is the case for Japanese students since their L1 equivalent, *kusuri wo nomu*, directly translates to *drink medicine* (2015, p.228). These tasks were designed to highlight these differences between the students' L1 and L2.

These tasks were mainly carried out during weeks 2-5 of the course, but continued intermittently throughout the course.

3.3 Teaching Materials Project

O'Keeffe, McCarthy and Carter (2007) outline the use of corpus data in designing textbooks. A brief summary the theory behind this was presented to the students who were then asked to prepare teaching materials that would be suitable for beginner Japanese students of English. They were asked to draw on their own experiences as students, and the main criteria were that they use BNCweb to confirm any hypothesis they may have had, and that they include examples from the BNC in their materials. The students carried out this project in groups/pairs. When they had competed their teaching materials they taught their materials with their classmates acting as their students. This project ran from weeks 5-8 of the course.

3.4 Corpus Building Project

The students were given an overview of the theory behind the use of specialized corpora and asked to come up with a list of topics related to their hobbies and/or subjects related to their majors (the students were from 6 different faculties within the university). They were grouped with students who had listed similar topics or subjects, some of these were: Disney movies, American literature and Song lyrics. The groups were asked to make a specific list of questions they would like to investigate. Following on from this, the groups collected text related to their topic/subject and complied it into text (.txt) files. These text files were then uploaded to AntConc, which Anthony (2014) describes as "a freeware corpus analysis toolkit". The groups used AntConc to investigate their research questions. Finally, they presented their results to the class. This project was carried in weeks 9-11 of the course.

3.5 Individual Research Project

This project was introduced before the corpus-building project, but could be considered the final project of the course since much of the work was carried out after the corpus-building project had concluded. The project was introduced in week 9 and students were given 2 weeks to formulate research questions. These questions were checked and refined in week eleven. From the end of week eleven to the start of week fourteen the students worked on their individual research in class with teacher support, and alone for homework. They were given freedom to choose a research topic that they wished under the guidance of the teacher. Topic chosen included the difference between *sweet* and *cute*, and translations of the Japanese verb *jistugen suru,* these included *achieve, realize, fulfill, materialize, come true and accomplish*. They were required to write a short essay on their research and to present the results of their research in week fourteen.

4 Data Collection

4.1 Can-do Self-assessment

Students were required to complete a CEFR style can-do (see Verhelst, Van Avermaet, Takala, Figueras & North, 2009) pre-course self-evaluation in the second class, and the same self-evaluation in the final class. The students were asked to self-evaluate themselves in terms of can-do statements with on a scale of 1-5 with 1 indicating that they could not do this and 5 indicating that the could do this with no problems. The evaluation contained a total of thirteen can-do statements, seven of which related directly to corpus; these seven are listed in in the results section of the paper below.

4.2 Survey

In the final class, students were asked to complete an online survey. The aim of this survey was to gather student opinions about the course. The survey was comprised of six sections, one section each to gather student opinions about each of the five main components of the course outlined above, and one section to gather student opinions about the course as a whole. The survey was comprised of twenty-five statements (see the results section) which students were asked to respond to on 4-point Likert scales. Each of the first five sections had an optional comment space at the end of the section and the final section had a required comment section where students were asked to write freely any comments they had about the course.

5 Results

5.1 Can-do statements

The results in Table 1 show the pre and post course means of the self-assessment, which was on a 5-point scale.

Can-do Statements	Pre-course mean	Post-course mean
I can use BNCweb for research and investigation	2.07	4
I can formulate appropriate research questions relating to vocabulary and grammar.	1.92	4.07
I can make teaching materials from corpus data.	1.69	3.92
I can build a specialized corpus.	1.23	3.23
I can carry out individual corpus research.	1.76	4.07
I can understand the theories behind corpus research.	1.84	3.54
I can use other research tools to support corpus research.	2.23	4.08

Table 1: Can-do Statements Pre and Post Means (n=21)

The can-do statements indicated that the students' felt they had progressed significantly as a result of the course in all areas. However, it must be said that with an area of study such as corpus linguistics, where all the students admitted to having almost no prior knowledge of the subject, it would be unlikely that the student would feel they had not progressed.

5.2 Survey - Likert Scale items

Table 2-7 show the results of the survey in which the answers were on a 4-point Likert scale with 1 being strongly disagree and 4 being strongly agree.

Survey Item	Mean	Standard Deviation
1. I can understand the theories behind the study and use of corpus	3.1904	0.602
2. Learning about the theory of corpus in this class was useful and interesting	3.1904	0.981
3. This course should include more theory	2.238	0.7

Table 2: Survey Results, Section 1 - Theory (n=21)

The student responses indicated that the amount of theory was adequate, but that perhaps students had some difficulty in understanding all the theory, this was echoed in the student comments where two students said they thought the course was too difficult at first.

Survey Item	Mean	Standard Deviation
4. The corpus tasks helped me to learn how to use BNCweb	3.524	0.512
5. The corpus tasks helped me to develop ideas for my subsequent projects and research	3.571	0.507
6. The corpus tasks were relevant and interesting	3.3	0.865
7. The corpus tasks were appropriately challenging (not to difficult, not too easy)	3.238	0.889
8. The course should include more corpus tasks	2.571	0.811

Table 3: Survey Results, Section 2 - Corpus Tasks (n=21)

The results in this section indicated that the students found the corpus tasks to be at an appropriate level and that they found them to be beneficial in terms of becoming more proficient users of BNCweb.

Survey Item	Mean	Standard Deviation
9. The teaching materials project helped me to develop ideas for my subsequent projects and research	3.45	0.686
10. The teaching materials project helped me to improve my skills using BNCweb	3.45	0.759
11. The teaching materials project was relevant and interesting	3.4	0.681
12. The teaching materials project was appropriately challenging (not to difficult, not too easy)	3.15	0.988

Table 4: Survey Results, Section 3 - Teaching Materials (n=21)

Again, the student responses indicated that they found the teaching materials project to be useful. However, there was some deviation in item twelve where some students felt the project was not challenging enough.

Survey Item	Mean	Standard Deviation
13. The corpus building project helped me to improve my skills using various corpus tools	3.048	0.74
14. The corpus building project helped me to develop ideas for my subsequent research	3	0.858
15. The corpus building project was relevant and interesting	3.238	0.768
16. The corpus building project was appropriately challenging (not to difficult, not too easy)	3.238	0.7

Table 5: Survey Results, Section 4 - Corpus Building Project (n=21)

This section received the lowest ratings from the respondents. This may be due to the fact that they were learning to use a new corpus tool, AntConc, or to the time consuming nature of collecting the texts required to build the corpus.

Survey Item	Mean	Standard Deviation
17. I had a clear research topic	3.552	0.759
18. The theory, tasks and projects helped me to prepare for my research	3.551	0.51
19. I could use other tools (dictionaries, surveys, etc.) to support my research	3.4	0.598
20. I had interesting and significant results	3.351	0.745
21. The individual research was appropriately challenging (not to difficult, not too easy)	3.1	0.641
22. The presentation component of the individual research was appropriately challenging (not to difficult, not too easy)	3.15	0.875
23. The essay writing component of the individual research was appropriately challenging (not to difficult, not too easy)	2.95	0.945

Table 6: Survey Results, Section 5 - Individual Research Project (n=21)

The individual research project was mostly well received by students. However the lower scores for items twenty-two and twenty-three and the larger deviation reflected the differences in the students' previous EFL study with some students having studied academic writing and presentation, while others had not.

Survey Item	Mean	Standard Deviation
24. Overall, I benefitted from this course	3.611	0.503
25. I will use corpus in the future for my research and/or English study	3.462	0.681

Table 7: Survey Results, Section 6 - Overall Impression of the Course (n=21)

The last two items on the survey were designed to elicit the students' overall impression of the course. They indicate that the students felt that they benefitted from the course, and that they were likely to utilize corpus data again in the future.

5.3 Survey - Student Comments

Each of the first 5 sections had an optional comment section for students to comment on that section of the course, however only three comments were made in sections 1 to 4. One student reported that they had found the teaching materials project to be effective and had applied what they learned in another class. The other two comments were about the corpus-building project. One student simply commented that they had found the project difficult while another commented that the project was extremely time consuming.

There were three student comments about the individual research project. One student commented that they had gradually learned to use corpus throughout the course and that the individual research project was a good way of self-checking that they had achieved some kind of proficiency in using corpus. Another commented that the essay writing was difficult. Finally, one student commented that this was the biggest project, and that it helped them the most to understand how to use corpus.

The comments section at the end of section 6 was required and as such all twenty one students responded. Many of the responses were not direct evaluations of the course or reflections on the course, but rather general comments thanking the teacher and saying they had enjoyed the course. There were comments that were more specific, two students commented that they had felt this course was too difficult at the beginning, but by the end they had gotten to grips with it. Another student commented that they had difficulty understanding the need for corpus when they could check words in a dictionary, but through the tasks and projects they had seen the usefulness of corpus. Two students commented on having used or planning to use corpus again outside of this class. In addition to the two student comments about the course being difficult at the start, the only other negative comment was in a similar vein, a student commented that having only previously experienced standard EFL classes where their teachers spoke slowly, they had difficulty with understanding the teacher's instructions. Finally, one student commented that it was good that this course did not require students to do too much homework, this may however indicate that students were not investigating the concordance lines in enough depth, this will be commented on further in the conclusion below.

6 Conclusion

This paper aimed to investigate two research questions:

1. To what extent did students improve their practical and theoretical knowledge of corpus linguistics as a result of their participation in this course?

2. At the end of this course, what were the students' opinions about the use of corpora and about the specific components of the course?

In answer to question 1, while it is clear from the student self-assessments that there was significant improvement, this is hardly groundbreaking since all of the students admitted to having virtually no knowledge of corpus linguistics before the course began. However, the extent to which they felt they improved is encouraging.

The survey designed to gather data to answer to question 2 provided results of more consequence. In terms of the course design the student responses indicated that the corpus-building project might be an area for improvement in future iterations of designing the course. Also, the student comments and the responses to items twenty-two and twenty-three on the survey, pertaining to academic writing and presentation, indicated some complications. The comments demonstrate that bringing undergraduate students from 11 different faculties with a variety of EFL learning experiences together in a course such as this may present difficulties in relation to the English level and the academic skills required. I would however contend that a teacher who is aware of this and supportive of the students could overcome these difficulties. Additionally, there were some indications from the survey that students felt that certain aspects of the course were time consuming and also a comment that there the low amount of homework was a good thing. I would propose that these comments mean that students lack awareness of two things, firstly the amount of time that a researcher would have to invest to collect text and build a corpus, or the amount of time that any researcher must invest in any research project. This lack of awareness is understandable for undergraduates many of whom were taking their first steps in academic research, but it remains an area that student need to be more conscious of. Secondly, I would suggest that the student comment about the low amount of homework relates to students not spending enough time reading concordance lines when asked to do so outside of class. Earlier in the paper Sripicharn's assertion that students need to engage with the corpus and get used to reading 'messy' concordance was discussed. Perhaps the students who took this course would benefit from more explicit training in reading concordance lines in future iterations of the course.

Having discussed the areas for improvement in this course, I would like to conclude the paper by reviewing the positive aspects of the course. The student comments reflected that they felt they had benefitted from and enjoyed the course, which is crucial for any teaching situation. They also indicated that they had gained confidence in investigating corpus data, and that they would be able to use corpora such as the BNC to aid their future studies and that they had realized the usefulness of corpora through this course.

References

Anthony, L. (2014). AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Barker, D. (2010). *An A - Z of Common English Errors for Japanese Learners* (2nd ed.). Tokyo: BTB Press.

Campoy-Cubillo, M. C., Bellés Fortuño, B. and Gea-Valor, M. L. (Eds.). (2010). *Corpus-based approaches to English language teaching.* London: Continuum.

Donnellan, M. (2015). An Analysis of Collocations in an Authentic Text. *Kwansei Gakuin University Humanities Review, 19,* 227-243.

Flowerdew, L. (2012). *Corpora and Language Education.* London: Palgrave MacMillan.

Hoffmann, S., & Evert, S. (2013). BNCweb (Version 4.3) [Computer Software]. Lancaster, UK: Lancaster University. Available from http://bncweb.lancs.ac.uk/

Hoffmann, S., Evert, S., Smith, N., Lee, D., & Belglund Prytz, Y. (2008). *Corpus Linguistics with BNCweb - a Practical Guide.* Frankfurt: Peter Lang.

Hunston, S. (2002). *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

Johns, T. (1990) From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *CALL Austria*, 10: 14–34.

Louw, B. (1993). Irony in the text or insincerity in the writer? the diagnostic potential of semantic prosodies. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 157-176). Amsterdam: John Benjamins Publishing.

O'Keeffe, A., & McCarthy, M. (Eds.). (2010). *The Routledge Handbook of Corpus Linguistics*. Abingdon/New York: Routledge.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From Corpus to Classroom: Language Use and Language Teaching.* Cambridge: Cambridge University Press.

Nation, P. (2001). *Learning Vocabulary in Another Language.* Cambridge: Cambridge University Press.

Römer, U. (2010). Using general and specialized corpora in English language teaching: Past, present and future. In: Campoy-Cubillo, M. C., Bellés Fortuño, B. and Gea-Valor, M. L. (Eds.), *Corpus-based approaches to English language teaching* (pp.18-38). London: Continuum.

Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sripicharn, P. (2010). How can we prepare learners for using language corpora? In A.O'Keeffe & M.McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 371- 384). Abingdon/New York: Routledge

Verhelst, N., Van Avermaet, P., Takala, S., Figueras, N., & North, B. (2009). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.