Maggie Charles
University of Oxford, Great Britain

DO-IT-YOURSELF CORPORA FOR LSP: DEMYSTIFYING THE PROCESS AND ILLUSTRATING THE PRACTICE

Abstract

This paper argues for an approach to LSP in which teachers and students compile their own do-it-yourself corpora using specialised texts in their area of study and teaching. I show that it is relatively easy to build a rough and ready corpus of this type and illustrate how it can be used both indirectly in the construction of materials for a German reading course and directly by EAP students taking a course on editing doctoral theses. It is suggested that both LSP teachers and students can benefit from such a corpus due its specificity and relevance to their needs. For teachers, a DIY corpus provides a means of familiarising themselves with the discourse of a specialised area and a source of authentic examples for materials production, while students particularly value the corpus as a lexico-grammatical reference resource.

Keywords: data-driven learning; English for academic purposes; do-it-yourself corpora; German for academic purposes; corpus pedagogy; materials development

1. Introduction

A corpus can be defined as an electronic collection of naturally occurring texts that is assembled according to set criteria for a specific purpose and is accessed through the use of software designed for linguistic analysis. In this paper I show that it is relatively easy to compile a rough and ready corpus and that there are many benefits of being able to do so, particularly for teachers of languages for specific purposes (LSP). I use the term 'do-it-yourself (DIY) corpora' to refer to corpora created locally by teachers or students for their own individual use, whether as an input to materials production or for consultation during the language learning and teaching process (Charles, 2012, 2015, 2017).

A distinction has often been made between 'direct' and 'indirect' uses of corpora (Leech, 1997). A 'direct' use of corpora is one that involves the students in having access to the corpus data themselves. This may be through hands-on corpus consultation, but may also occur when students are given print-outs of corpus data to examine on paper. In both cases the corpus results are often investigated with a view to answering specific lexico-grammatical queries. This approach is termed 'data-driven learning' or DDL and was pioneered in the teaching of English for academic purposes (EAP) by Johns (1991a, b). It has been argued that DDL offers many benefits to the learner; thus O'Sullivan (2007) points out that corpus consultation involves the exercise of mental or cognitive skills such as predicting, observing, noticing, analysing, interpreting and making inferences. She argues that DDL increases the mental activity of the learner and helps develop mental and cognitive processes. Moreover, there is evidence in support of the efficacy of DDL from a meta-analysis of 64 studies in a range of difference pedagogical contexts, which concludes that "DDL works pretty well in almost any context where it has been extensively tried" (Boulton & Cobb, 2017, p. 386). Thus it would seem that the direct use of corpora may well have a valuable contribution to make to the LSP classroom.

A use is termed 'indirect' when the corpus data form the basis for the construction of teaching or reference materials, i.e. when students do not have access to the corpus itself, but rather encounter corpus data indirectly, through pedagogic or reference resources. As Leńko-Szymańska and Boulton (2015) point out, corpora are currently used in the production of learner dictionaries, reference grammars, usage manuals, courses, course manuals and supplementary materials. Given that so many types of learning resources are now informed by corpus data, indirect uses are likely to be the most frequent way in which corpora appear in the LSP class, even though neither teachers nor students may be consciously aware of this.

For LSP teachers, there are two factors that should be taken into account when considering the feasibility of incorporating direct or indirect corpus use into their classes: the availability and specificity of the corpus. The corpora used in commercial materials production are often the property of publishers and may well be inaccessible to teachers. However, there are a large number of open-source corpora freely available on-line. English is the language best served in this regard, with large general corpora like the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA) available at http://corpus.byu.edu/. Smaller more specialised collections that could be of great interest to ESP teachers include the corpora of professional discourse held at the Hong Kong Polytechnic, http://rcpce.engl.polyu.edu.hk/index.html, and the Michigan Corpus of Upper-level Student Papers (MICUSP)

at http://micusp.elicorpora.info/. The website of The Compleat Lexical Tutor, https://www.lextutor.ca/, gives access to several corpora of English, German, French and Spanish. In addition, much larger general corpora of German are available at https://www.dwds.de/r, while Spanish is catered for by the Corpus del espagnol at http://www.corpusdelespanol.org/ and Italian by the PAISA corpus at http://www.corpusitaliano.it/en/. However, although these corpora provide a good basis for general language queries, they are much less likely to provide adequate coverage for highly specialised LSP courses.

One way to address this situation is to build an individual DIY corpus using specialised texts that represent the academic or professional area targeted by the course. The advantage of compiling such a corpus is that the user has control over its contents and can ensure that only relevant and appropriate texts are included. Used indirectly as a reference resource by the teacher, it enables the analysis of the specialist discourse, the provision of authentic examples and offers a reliable resource for checking students' work. Used directly by students, such a DIY corpus allows LSP learners to consult the work of expert writers, to pose their own queries and to gain answers that reflect the discourse characteristics of their own field. For both teachers and students, then, using the data from a specialised DIY corpus can lead to increased knowledge of and confidence in the use of the target discourse.

2. Building a DIY corpus: Process and problems

The process of compiling a DIY corpus for local or individual use is fairly straightforward, involving four, or optionally five, steps. This section describes each of these steps in detail, indicates some of the problems that may arise and how to deal with them.

The first step in the process is one of the most important, but at the same time one of the most familiar, as it requires the compiler to select the texts that will form the contents of the corpus. Before beginning the construction itself, it is necessary to think carefully about the purposes for which the corpus will be used, to decide on the type of information to be retrieved and who is going to use the corpus. For example, will it be accessed only by the teacher for the creation of student materials, is it intended for use by the students directly, or for both purposes? The answers to these questions will have an influence on decisions concerning the texts to be included and the way in which the corpus is organised.

For example, a corpus of research articles in a relevant discipline would be particularly useful for an advanced academic writing class, while at lower levels, it might be more appropriate to build a corpus from other web-based resources or from the students' textbooks. Care should be taken to comply with copyright restrictions, but there is a wealth of open access material available. For example in EAP, a good source of discipline-specific texts can be found in the directory of open access journals (https://doaj.org/), which contains research articles in a range of science, technology, social science and humanities disciplines and allows browsing by subject.

The purpose for which the corpus will be used will determine which genre(s) to include. Thus, if the course focuses specifically on writing abstracts, then a corpus of abstracts alone will give more relevant results than a larger corpus of complete research articles. If the course covers a range of genres, then the corpus can be organised into several sub-folders each

containing a single genre, thus offering, for example, sub-corpora of abstracts, research proposals, laboratory reports and so on. The user can then decide to access either a single genre or the whole corpus, depending on the information required.

Two other issues that often arise in regard to DIY corpora concern the size of the corpus and the quality of the contents. It has been suggested that 50,000 to 250,000 words is enough for a highly specific corpus (Flowerdew, 2012). However, good results have also been achieved with smaller corpora and the length of the target genre will affect the overall size of the corpus. In practice, when building corpora of research articles a useful guideline is to aim for about 50 individual papers. This seems to provide adequate size, while still being an achievable target for both students and teachers.

Of more concern to some users is the quality of the language contained in the corpus. This is particularly pertinent when compiling a corpus of research articles for teaching EAP. Academic publication is now a truly global endeavour, which means that many contributions are written by non-native speakers. Although the corpus compiler can select publications by authors with affiliation to institutions in English-speaking countries, this is by no means a guarantee of native-speaker status; nor, of course, is native-speaker status a guarantee of good quality language use. Here it is important to stress that interpreting corpus data always involves attention to the relative frequency of the items retrieved. Thus the user should focus on the most frequent occurrences and ignore instances that are not part of a recurring pattern. Given that examples of standard usage are likely to be much more frequent than non-standard examples, increasing the size of the corpus will render the disparity more apparent and thus easier to take into account when interpreting the data.

Once the texts have been selected, step 2 is to convert them into plain text format so that they will be readable by the corpus software. This is best carried out using a batch file converter, which is a piece of software that converts multiple files at a single command. A freeware tool of this type is the AntFileConverter (Anthony, 2015), a screen shot of which is provided in Figure 1. Files in pdf or Word format are uploaded into the left-hand window under 'Input Files' and the user presses the 'start' button. The converted files appear one by one in the right-hand window under 'Output Files' and the software reports either 'created' or 'failed' for each file. The files are saved with the same names as the originals but in a separate folder called 'txt', which the software creates in the same folder as the original files.

As a third step, it is advisable to check that the files have actually converted successfully. This can be done simply by scrolling through, scanning the text quickly. Very occasionally the software signals that a text has been created, when the content is missing. This can easily be verified by looking at the size of the text file; if it is less than around 1KB, the file is probably empty. There are two other types of issues that might arise. First, there can be problems with line and word breaks; sometimes the conversion produces a text with a single word on each line. In this case, it is probably best to delete the file from the corpus, since it is likely to prove difficult to interpret the text. The lack of word breaks, also known as 'clumping', produces instances such as *Platformmeasurementandreporting.* Here again, if the issue is widespread, it is likely to disrupt normal usage and the file should be deleted. Other difficulties that might arise at the conversion stage are those affecting the recognition of certain characters. In particular, the combinations 'ff' and 'fi' in words such as *different* or *definition* may not be converted properly, giving forms such as *de□nitions*. This is annoying, but may not pose a

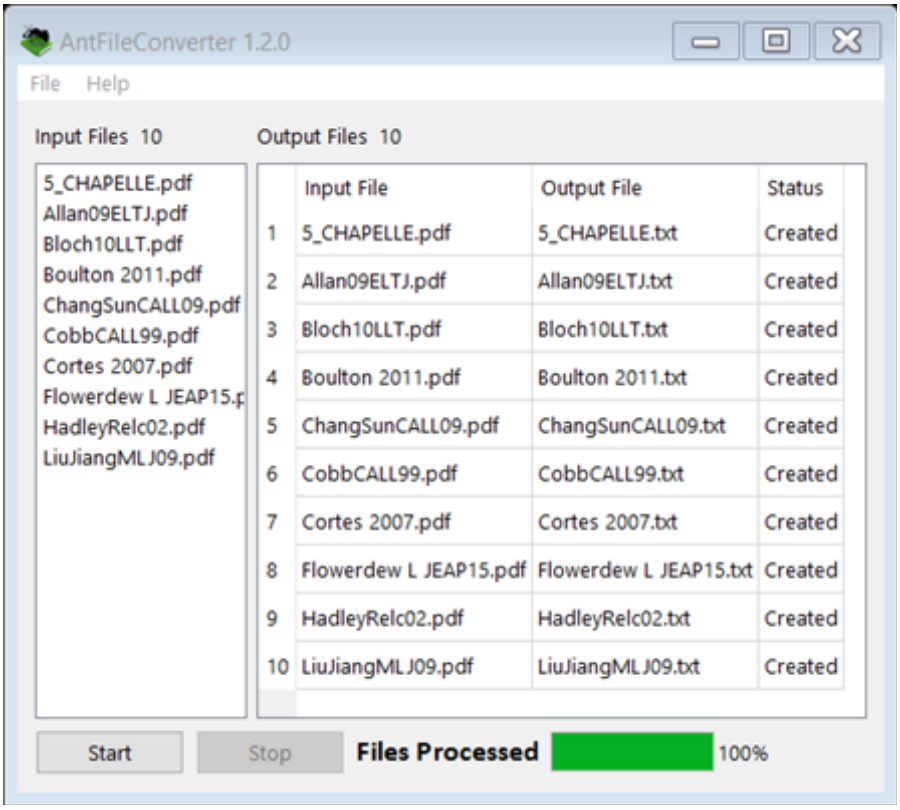

Figure 1. Screenshot from the AntFileConverter.

serious problem, providing the user can readily understand the meaning of the words. In all cases where the conversion has not completely succeeded, it is more time-efficient simply to delete the file and replace it with another one rather than to spend time trying to correct the ill-converted text. It should be stressed that, although I have spent some time detailing these potential problems, the AntFileConverter does convert the vast majority of files without any problem.

Having checked the file conversions, the fourth step is simply to re-name the 'txt' folder with the name of the corpus and the creation of a 'quick and dirty' corpus is complete. At this point, it is possible to load the corpus folder into corpus analysis software such as the freely available AntConc (Anthony, 2014) and proceed to make searches to investigate the data. This 'quick and dirty' corpus is adequate to answer most of the queries that teachers and students will want to pose. However, it is worthwhile understanding why it is called a 'dirty' corpus and how this may affect the data that is retrieved. By a 'dirty' corpus we mean that the files have not been cleaned of material that is not part of the running text. This includes, for example, the title of the text, the author's name, page numbers and graphics such as tables or charts. These will therefore appear when certain searches are carried out. For example, tables may appear as lists of numbers or there may be an unexpectedly large number of hits for certain terms. When teachers or students have built the corpus themselves, they are likely to be familiar with the contents and therefore it is usually quite easy for them to recognise instances that are not part of the running text. In fact, in a study of 40 students who used DIY corpora over a twelve-month period, only half carried out any cleaning (Charles, 2014).

The fifth and optional step is to clean the corpus, which involves going through each file and manually deleting all the material that is not part of the running text, a time-consuming

and tedious process, which probably does not justify the effort involved. Before embarking on any cleaning, it is therefore advisable to try out the corpus in its dirty version to see how well it performs. It should be borne in mind, however, that a quick and dirty corpus is not designed to provide statistically valid generalisations; rather, it is a rough and ready aid to allow teachers and students to find data and examples to use in their teaching and learning.

3. Using DIY corpora

3.1 Indirect use: Materials design for German for specific purposes

I have argued above that DIY corpora are a valuable resource when LSP teachers are faced with the problem of creating materials, especially in fields that are outside their area of expertise. Under these circumstances, creating a corpus of relevant texts and using it as the basis for materials ensures authenticity and relevance for students.

One example of the successful use of a corpus as an input to materials development is provided by the work of Eckhard-Black (2016) at X University Language Centre. This corpus initiative was undertaken in order to support the work of students who were taking her Accelerated German Reading Course for Classicists. This course consisted of three terms lasting eight weeks each, with four class hours per week. Students were also expected to devote another four hours per week to self-study. The course was designed for absolute beginners of German and by the end of the course the aim was for students to be able to understand with a dictionary academic texts from specialist journals, to extract important information and to decide on its relevance to their own work. Using the CEFR descriptors, the goal was a reading comprehension level of C1/C2. The participants were graduate students of Greek or Roman history or Greek or Latin language and literature. At the end of the course an examination required German to English translations of three passages from academic journals.

These passages for translation formed the content of the corpus, offering practice material which was both authentic and highly relevant to the students. Passages from past examination papers (set by Eckhard-Black) were downloaded from the examinations website and converted into plain text files to form a corpus consisting of 136 files (33,554 words). Although relatively small in size, this corpus is highly specialised, representing a genre and register that reflects the aims of the course and corresponds to the needs of the participants. Hence it can be considered adequate in size to achieve its purpose.

Eckhard-Black used the corpus as the basis for creating a range of materials that students could access for self-study and reference. Using the Wordlist tool in AntConc, she compiled several word lists, including a list of the 24 most frequent functional words for reading academic texts, as illustrated in Table 1. One of the advantages of using a corpus to compile such a list is that the software gives frequency counts, which makes it easy to present the words in frequency order, as here. This means that students can gain an appreciation of the most important functional words to focus on, especially at the early stages of language learning.

Table 1

Extract from List of Top 24 Functional Words.

German	English	Grammatical Description
den	the	definite article [masc: Acc // plural: Dat]
der	the	definite article [masc: Nom // fem: Dat + Gen // plural: Gen]
die	the	definite article [fem: Nom + Acc // plural: Nom + Acc]
im	in the	= in + dem (Dat) contraction = preposition + definite article
in	into // in	two-way preposition (+ Acc // Dat)
und	and	co-ordinating conjunction

Eckhard-Black also created a much longer word list, which gave the most frequent words for classicists and historians and presented them in alphabetical order. It consisted of over 200 entries and included all lemmas that occurred at least seven times per 20,000 words. This material was available to students as a printable list, as well as providing the input to ANKI flashcards for learning and revising the vocabulary. For further details of the ANKI system, see https://apps.ankiweb.net/. A short extract from Eckhard-Black's list of frequent words is given in Table 2.

Table 2

Extract from List of the Most Frequent German Words for Classicists and Historians

German	English
ägyptisch	Egyptian adj
allerdings	However adv
allgemein	general adj & adv
alt	old adj
der Anfang ("e)	beginning noun
die Antike	antiquity noun

From the entries in Table 2, it can readily be seen that this is a highly specialised list with a wealth of useful detail and Eckhard-Black makes the important point that the translations of the German words are not necessarily their base meaning, but rather the meaning they have in academic texts.

In addition to providing lists of vocabulary, the corpus also allows the construction of material designed to help students with tricky translation problems, exemplified by the explanatory sheet entitled "Translating *wohl*: Adverb or Modal Particle?" Again, the utility of the corpus is demonstrated particularly by its ability to provide frequency information on the various meaning and translation options for *wohl*, which enables students to assess how likely they are to come across each individual meaning. For example, the material states that *wohl* occurs fifteen times in the corpus. There are four instances of *sehr wohl,* which is often translated by *indeed* or *very much so* and nine instances of *wohl* with *bestätigt* or *tatsächlich*, which implies that the writer is making a guess. The material provides authentic examples derived from the examination passages, with translations to assist the student in understanding the differences between the meanings, as seen in the following examples:

"*Dabei ist **sehr wohl** denkbar…*"

*"In this context it is **very** plausible…"*

"*…der Limes im Westen **wohl tatsächlich** als Schutz vor den einfallenden Germanenstämmen gedient haben mag…*"

*"The Limes in the west **may well** have **actually** served as protection against the invading Germanic tribes…"*

The use of the corpus greatly facilitates the selection of appropriate examples and enables the materials writer to have authentic data to support their conclusions, rather than having to rely on their own intuitions. Building a small DIY corpus such as that employed by Eckhard-Black is a relatively simple procedure, as indicated in section 2, and has a substantial pay-off in terms of its relevance, specificity and ease of use.

3.2. Direct use: Data-driven learning for EAP

As noted earlier, DIY corpora are a valuable resource not only when the data are used indirectly by teachers, but also when students interact directly with the corpus in a data-driven learning approach. To illustrate the direct use of DIY corpora, I will take some examples from a course on editing for EAP doctoral students. On this course, each student built an individual corpus of research articles in their own field, which they used to explore the characteristics of expert writing in their discipline; they also built a corpus consisting of chapters of their own thesis, which enabled them to compare their own writing against published papers. Students were introduced to the AntConc software (Anthony, 2014) and were shown how to apply the individual tools to answer their own editing queries. For each of these queries, students were asked to fill out a worksheet with details of the question they wanted to address, the corpus searches they made, the results they obtained and the conclusions they drew from their data.

As noted by Lee and Swales (2006), advanced graduate students are likely to face issues of lexico-grammatical 'fine-tuning' and it is precisely in this area that corpora can be especially helpful. One of the most valuable corpus tools for editing and improving text is the concordancer. This tool creates a set of concordance lines for a search item (a word or phrase) that the user enters. The search term appears in the centre of the line, with several words of context either side of it. These data enable the user to discover how the search item is used and to notice the patterns that are associated with it, thereby providing information that can be used to improve their writing. Figure 2 shows a screenshot from AntConc with a set of concordance lines retrieved by Malee, a student who was investigating the use of the verb *indicate*.

Malee was a Thai doctoral student of chemistry, who built a corpus of 50 research articles (almost 270,000 words). She gave her question on the worksheet as follows: *"Can we use 'is indicative of' instead of 'indicate'?"* A search on *indicative* found 13 hits, while 192 instances appeared with the search term *indicate**. The asterisk * is a wild card, standing for any character(s), so this search retrieved the verb forms: *indicate/indicates/indicated*. The corpus data showed that both *indicative of* and the verb *indicate* can be used, but that the verb form was much more frequent. After examining the concordance lines, Malee summed up her

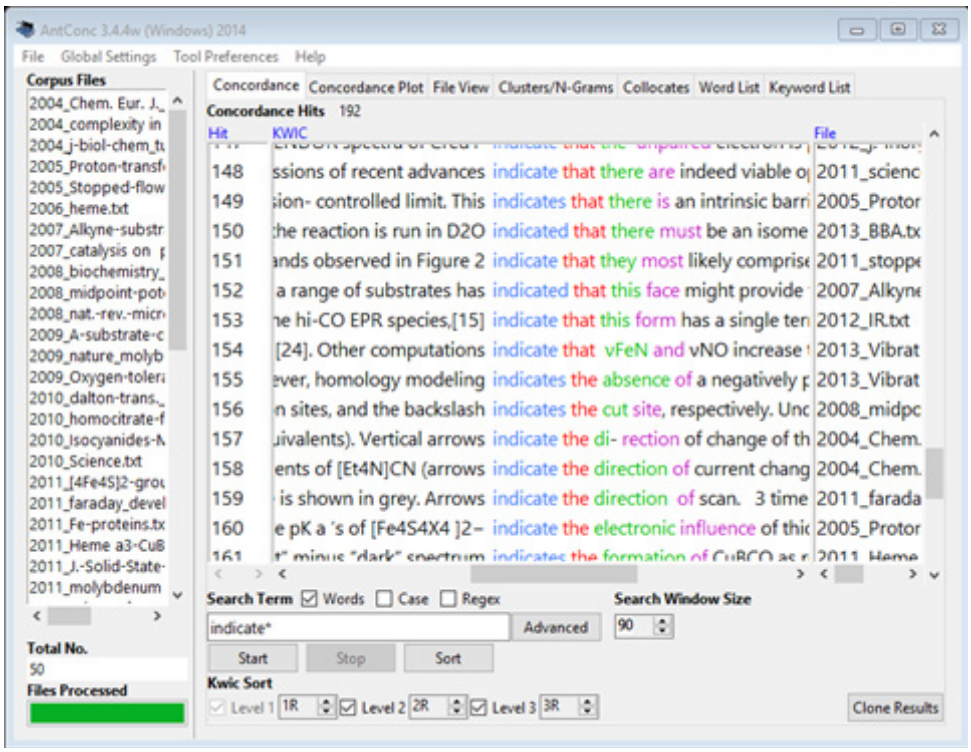

Figure 2. Screenshot from AntConc of an extract from a concordance on the search term *indicate**.

conclusions as follows: *"indicate and indicative of follow by noun; indicate that follow by sentence/phrase."* Although the student did not use the standard grammatical term 'clause', it is clear that in addition to answering her original question, she was able to distinguish three patterns of use: 1) *indicative of* + noun; 2) *indicate* + noun; 3) *indicate that* + clause. This information should enable her to employ these items with more confidence in her own writing.

Another tool, called 'n-grams', can be useful in helping students identify the typical phraseology of their field. An n-gram is a sequence of n words. For example, in the phrase *students from diverse backgrounds*, there are four 1-grams or words, three 2-grams: *students from, from diverse* and *diverse backgrounds* and two 3-grams: *students from diverse* and *from diverse backgrounds*. When applying the n-grams corpus tool, the user determines the length of the sequences they wish to obtain (3-gram, 4-gram etc.); the software then retrieves all the n-grams of that length in the corpus and presents them in a list with information on their frequency. The advantage for the user is that the tool is fully automatic. In contrast to the concordancer, which retrieves only the words or phrases requested by the user, the n-grams tool retrieves all possible sequences. The automaticity of the procedure means that the n-grams tool can reveal fixed or semi-fixed phrases which may not be familiar to the student. Thus, for example, Yoko, a Japanese doctoral student of medicine, noted that the top 3-grams retrieved from her research article corpus of over 2 million words (273 files) contained specialist terminology such as "*the auditory cortex*" and "*medial geniculate body*". Further examination of these 3-grams, using the concordancer, revealed the wider context in which the terms occurred, thereby providing the student with rich and authentic examples to facilitate learning, e.g. "*most recent investigators have shifted to using electrical current to activate the auditory cortex*".

For students, the benefits of having direct access to a tailor-made DIY corpus derive from the fact that, as corpus builders, they are familiar with the content of the corpus and have control over it. Thus students can be sure that the examples they retrieve occur in their specialist area and have been written by experts in the field. This, in turn, gives them the confidence they need when making linguistic decisions about their own texts.

4. Conclusion

In this paper, I have sought to show that the construction of DIY corpora is well within the capabilities of most LSP teachers and that the benefits far outweigh the initial input of work needed. For LSP teachers faced with teaching a course in a new and unfamiliar area, constructing a DIY corpus can enable them to familiarise themselves with the necessary specialised discourse, as well as providing a reference resource for responding to students' questions about linguistic usage in the field. For many specialist courses where no suitable teaching materials are available, the data from such a corpus can offer examples and lexico-grammatical support for creating relevant and appropriate materials. At an advanced level, students can create their own corpora and often prove to be enthusiastic users of corpus data, appreciating the autonomy that access to their own corpora brings, since they need no longer be entirely reliant on native-speaker teachers, proof-readers and editors to improve their texts. As one student advised with regard to DIY corpus-building, *"Don't be afraid of using this technique. Very soon you will find its worth."*

References

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

Anthony, L. (2015). AntFileConverter (Version 1.2.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/.

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, *67*(2).

Charles, M. (2012). 'Proper vocabulary and juicy collocations': EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes*, *31*(2), 93–102.

Charles, M. (2014). Getting the corpus habit: EAP students' long-term use of personal corpora. *English for Specific Purposes*, *35*, 30–40.

Charles, M. (2015). Same task, different corpus: The role of personal corpora in EAP classes. In A. Boulton & A. Leńko-Szymańska (Eds.), *Multiple Affordances of Language Corpora for Data-driven Learning* (pp. 131–153). Amsterdam: Benjamins.

Charles, M. (2017). Do-it-yourself corpora in the classroom: Views of students and teachers. In K. Hyland & L. Wong, (Eds.), *Faces of English education: Students, teachers and pedagogy* (pp. 107–123). Abingdon: Routledge.

Eckhard-Black, C. (2016). Corpora and Materials for Accelerated German Reading for Classicists.

Flowerdew, L. (2012). *Corpora and Language Education*. Basingstoke, UK: Palgrave Macmillan.

Johns, T. (1991a). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Johns & P. King (Eds.), *Classroom Concordancing* (pp. 27–37). Birmingham: ELR University of Birmingham.

Johns, T. (1991b). Should you be persuaded: Two samples of data-driven learning materials. In T. Johns & P. King (Eds.), *Classroom Concordancing* (pp. 1–16). Birmingham: ELR University of Birmingham.

Lee, D., & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes*, *25*(1), 56–75.

Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichman, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and Language Corpora* (pp. 1–23). London: Longman.

Leńko-Szymańska, A., & Boulton, A. (2015). *Multiple Affordances of Language Corpora for Data-driven Learning*. Amsterdam: Benjamins.

O'Sullivan, Í. (2007). Enhancing a process-oriented approach to literacy and language learning: The role of corpus consultation literacy. *ReCALL*, *19*(3), 269-286.