THE JOURNAL OF TEACHING ENGLISH FOR SPECIFIC AND ACADEMIC PURPOSES
Vol. 6, N° 2, Special Issue, 2018, pp. 217–224

UDC (811.111'243:654.922):378.147 https://doi.org/10.22190/JTESAP1802217C

USING DO-IT-YOURSELF CORPORA IN EAP: A TAILOR-MADE RESOURCE FOR TEACHERS AND STUDENTS

Maggie Charles

Oxford University
Phone: +44(0)1491651723, E-Mail: maggiecharles_oxford@yahoo.com

Abstract. *Do-it-yourself (DIY) corpora can be defined as small-scale databases of electronic texts built by users for specific, limited and local purposes. Such corpora can be of great benefit to both teachers and students of English for academic purposes (EAP), who, with recent software advances, can now construct DIY corpora to their own specifications relatively easily. For teachers, who may be tasked with giving courses related to disciplinary areas in which they have little or no expertise, the specialist DIY corpus provides an opportunity to examine a body of texts that they have selected as relevant to the target course and thus enables them to familiarize themselves with the discourse of the discipline in question. Such corpus-based investigations not only facilitate teachers' understanding of disciplinary norms and practices, but also provide examples for use in class or in course materials and give frequency information on lexis and phraseology so that instructional decisions can be made on a sound evidential basis. For graduate students, who have needs that are highly specific, it is also valuable to construct DIY corpora from material within their own field. This tailor-made resource provides individual, precisely-targeted information, which can be drawn upon to answer lexicogrammatical queries both at the composing and editing stages of the writing process. This paper makes the case for the use of DIY corpora in EAP contexts, illustrating the argument with two sets of examples: first, a teacher's use of a corpus of theses in creating a course for doctoral students in materials science and second, students' uses of DIY corpora for individual writing support.*

Key words*: do-it-yourself corpora, corpora in EAP, academic writing, discipline-specific discourse*

1. INTRODUCTION

The use of corpora in English for academic purposes (EAP) has grown substantially over the last couple of decades and is now widespread, covering what Leech (1997) termed 'indirect' and 'direct' uses. Indirect uses are those in which only the instructor or course developer has access to the corpus data and acts as an intermediary between the corpus and the students. The corpus data are used to inform the content of courses and the production of materials, but the end users may not be aware of this and are not themselves confronted with the data. Direct uses are those in which students have access to corpus data themselves, either by means of paper-based tasks which present material from the corpus or through their own hands-on investigations using concordance software on computers. Boulton and Cobb (2017) provide a recent review of this work.

Submitted April 5th, 2018, accepted for publication October 3rd, 2018

Many studies reporting corpus use in EAP are based on corpora that have been pre-constructed and are available, usually on-line, for the teacher to access. For example, Flowerdew (2012) reports on a course for business students using a 1-million word corpus of business letters (available at www.someya-net.com/concordancer), while Chen and Flowerdew (2018) describe workshops in research writing for graduate students, which access the British National Corpus (BNC). Large on-line corpora that can be used for teaching academic writing include the academic components of the BNC and the Corpus of Contemporary American English (COCA) (Davies 2008), both available at http://corpus.byu.edu/. Such corpora have several points to recommend them: they are likely to be representative and reliable as they are compiled by professionals; they may also be very large and thus give access to a wide range of data, but a major advantage for the teacher is that they are immediately available, although the use of the corpus query software still has to be mastered.

Despite the existence of such resources, however, many teachers still find it necessary to build their own corpora. Such corpora, called here 'do-it-yourself (DIY) corpora', can be defined as small-scale databases of electronic texts built by users for specific, limited and local purposes. Although such corpora take time to compile and may not be as large as pre-constructed corpora, their key advantage is that they can be built to correspond exactly to the discourse needs of a specific group of learners. The teacher/course developer has complete control over the content of the corpus and can adjust it at will, adding or deleting texts to reflect the ongoing needs of their students.

For teachers, this ability to respond precisely to the needs of the learners is particularly important when they are required to develop a course or materials in a discipline or genre that is not well-served by pre-constructed corpora. Section 2 describes the compilation and use of a DIY corpus of doctoral theses in materials science, illustrating a genre that is not often included in on-line corpora and a discipline that is also somewhat under-represented. However, it is not just teachers who can build DIY corpora; graduate students or researchers who are working in highly specialized fields have found that an individual DIY corpus can be constructed quite easily from the texts in their bibliography. This provides a topic-specific resource which they can use to supplement the information available in dictionaries or reference grammars and adapt to their developing research needs (Charles 2012, 2017; Lee and Swales 2006). The use of individual student DIY corpora is exemplified in section 3.

2. DIY Corpus Use for Teachers and Course Developers

When teaching EAP, it often happens that teachers are asked, sometimes with very little notice, to prepare a new course in an area of which they have very little knowledge. A DIY corpus is of considerable benefit in enabling teachers to discover some of the discoursal characteristics of an unfamiliar discipline.

In response to a request from the materials science department for a discipline-specific course on writing a doctoral thesis, a corpus of eight successful theses (about 300,000 words) written by native-speakers of English was compiled. The theses were divided into chapters, with one chapter per file, and were saved in plain text, the format necessary for using corpus software. The texts contained many equations, tables and other graphics, which were deleted to achieve a 'clean' corpus consisting only of the

running text. The procedure of cleaning is rather time-consuming and is probably only necessary if the corpus is intended for long-term use. Thus, while it was advisable for the teacher in this case, for students who build their own DIY corpora, it is not considered necessary (see section 3). The freeware AntConc (Anthony 2014) was used to examine the corpus.

Since the department had requested a genre-based course, the first corpus searches were devoted to finding out how the theses and their individual chapters were structured. Research has shown that there is substantial variation in doctoral theses with regard to the placing of the literature review (Swales 2004). Thus a search on the noun *review* was made and revealed that only one thesis contained a chapter headed *Review of the literature*. The other writers tended to include review sections within several chapters and to use content-based headings such as *Powder in Tube Developments*, rather than the term *review* or *literature review*. Although it might have been possible to ascertain these tendencies by reading the theses and using the search facility in Word, the use of the DIY corpus made the process much quicker and easier. Thus the corpus enabled the teacher to discover some generic features of an unfamiliar discipline and provided evidence to underpin course material on the structure of the thesis.

A further benefit of the DIY corpus is that research findings on genre analysis can be verified and exemplified within a given discipline. Using search terms found to be associated with a specific part-genre of the thesis, the teacher/course developer can find relevant passages to form the basis of course work. For example, Bunton (2005) notes the prevalence of discussions of future research in the conclusions to science and technology theses. A search on *future* not only allowed retrieval of these sections in the materials science thesis corpus, but also provided many instances of phraseology for students to learn and practise, including e.g. *in the future; suggestions of future work; future developments in…; point the way to the future*. Similarly, occurrences of Swales' 'Create a Research Space' model (Bunton 2002; Swales 1990, 2004) were located in the introduction chapters of the theses and selected extracts presented for analysis and discussion as part of the course.

Tasks using concordance lines were also incorporated into the course materials. Examination of individual chapters in the body of the thesis revealed that their introduction sections contained a step not reported in the literature: 'referring back to earlier chapters'. The following concordance lines, obtained from searches on *in chapter* and *previous chapter* were used as part of a task which asked students to note down two different ways of referring to earlier chapters, to report on the tense used (often past simple) and to explain why that tense was appropriate for realising this genre step.

1. HARACTERISATION 1. Introduction As discussed in Chapter II, it was suggested that the Gas R
2. : BEHAVIOUR OF CARBON BLACKS 1. Introduction In Chapter IV, graphite was investigated in t
3. . This was briefly illustrated with graphite in Chapter IV, primarily for comparison with t
4. r conductive polymer composites. As outlined in Chapter II, bulk treatments of particulate
5. bserved here. The creep strength was defined in the previous chapter as the stress required
6. l Particles Overview The apparatus described in the previous chapter was originally built o

Since research has reported the need for more of such metadiscoursal references in L2 academic writers' texts (Bunton 1999), students were encouraged to discuss the reason for the inclusion of this step, particularly in the context of a lengthy text such as a thesis. Thus the aim of the task was not only to promote noticing of the step and its lexicogrammatical realisations, but also to increase students' awareness of the needs of their readers and the use of metadiscourse markers to facilitate understanding.

As these examples show, DIY corpus analyses can reveal the discoursal characteristics of a given discipline and allow multiple authentic examples to be sourced. A DIY corpus also provides a number of options for presenting the data for different pedagogical purposes. As seen above, concordance lines focus students' attention on repeated regularities in the data and AntConc also enables gapped concordance lines to be obtained to create gap-fill tasks. However, sentence examples can be used if extended context is needed, or paragraphs can be extracted to illustrate genre moves. The word list tool provides a list of all the words in the corpus with their frequencies, enabling the teacher to present e.g. the most frequent reporting verbs or linking adverbs used in the discipline. The corpus is particularly helpful when marking highly specialized student writing, since it offers evidence of disciplinary usage, which may differ significantly from that of general English. For indirect use by teachers, then, a DIY corpus can readily provide access to a wealth of information which is vital to devising and teaching a course in an unfamiliar discipline or genre.

3. DIY Corpus Use for Students

When considering the direct use of a DIY corpus, it is necessary to take into account the make-up of the class in terms of the disciplines represented. Some classes are homogeneous, with all students taking a single course and following the same syllabus; this means that it is possible to compile a single DIY corpus which addresses the needs of all students, as shown for example by Bianchi and Pazzaglia in economics (2007). However, in many cases, groups may either cover a restricted range of disciplines (e.g. medical sciences or social sciences) or may be completely multi-disciplinary, including students working in any field. Given that there are substantial differences in the discourse of disciplines (see e.g. Hyland 2005), it is more useful in both these cases, for students to compile their own DIY corpora, which can correspond exactly to their own writing needs. Using current software, it is now relatively quick and easy to build such a corpus and it provides a tailor-made resource for students to consult throughout their career (Charles 2014).

This approach has been used on courses designed to help doctoral students edit their theses. The students' DIY corpora consisted of research articles (RAs) that they had downloaded to their bibliographies and they used the freeware AntFileConverter (Anthony 2014) to convert batches of these pdfs to plain text format. Cleaning was not carried out, as it is a lengthy process and most students found their DIY corpora adequate to their needs without cleaning.

Students searched their corpora using the AntConc software (Anthony 2014) and made queries relevant to their own individual concerns. Data is available for 90 students from a wide range of disciplines, who recorded their searches on worksheets, which were collected and analyzed by the teacher. The purpose of most student searches was to check lexicogrammar and extend phraseological knowledge. Although the distinction between 'discipline-specific' and 'general academic' discourse is not always easy to determine, roughly a third of the student queries were clearly discipline-specific.

The work of Francesca[1], an Italian student of chemistry, provides an example of this discipline-specific work. She built a corpus of 198,678 words, consisting of 51 RAs. Francesca wanted to find a synonym for the adjective *harsh* in the phrase *under harsh reaction conditions*. By searching on *conditions* and sorting the concordance lines to the left, she was able to find the options *forcing conditions* and *harder conditions*, and decided to use *under forcing conditions* in her text. Francesca went on to make two further searches to extend her knowledge of this phraseology. First, she checked the singular form *condition*, which returned no hits, showing that in her field, the phrase *under* + adjective + (*reaction) conditions* was generally used with the plural form of the noun *conditions*. Her third search, using the preposition *under* and sorting the lines to the right, retrieved *under harder conditions* and *under harsher conditions*, confirming the main options as *under + forcing/harsh(er)/hard(er) + conditions*.

I would argue that for students working in highly specialized areas, such detailed phraseological work is best performed on a tailor-made corpus of specialist texts. The collocation *under forcing conditions* is unlikely to occur in a general corpus; indeed there are no hits in the BNC (100 million words) and only one in COCA (560 million words); moreover a doctoral student like Francesca would not be able to have confidence in the appropriateness of that instance for her own purpose of writing a thesis in chemistry, since the example occurs in 'Physics Today', a publication which is a magazine rather than an academic journal and is devoted to physics rather than chemistry. By contrast, Francesca's corpus provided eight instances of *forcing conditions* despite being far smaller at just under 200,000 words. These eight examples provided Francesca with sufficient information to make a well-motivated choice and she could be confident that this collocation was indeed used in her own field. It should also be noted that, unless they had specialist training in chemistry, a teacher of EAP would not be able to suggest the adjective *forcing* as a collocate of *conditions*; nor would they be able to verify its applicability in this context.

Similar discipline-specific searches included a student of zoology who wanted to know whether *fauna* was used as a singular or plural noun, a student of psychology who needed to distinguish between *visuoauditory* and *audiovisual* and a student of art history who wanted to extend her knowledge of adjectives used with the noun *brushwork*.

A DIY corpus also provides a wealth of examples of general academic discourse and it is estimated that about two-thirds of the student searches concerned issues that are frequent across a wide range of disciplines. It is revealing, therefore, to examine how queries about the same issue carried out by students from contrasting disciplines can differ considerably, in terms of the search itself, the findings obtained and the conclusions drawn by the student. An example of this variation is provided by the queries of three students in different disciplines, who investigated the phraseology of the verb *compare*.

Mitsuko, a Japanese student of sociology, asked the following question: 'which is common to say, compared with or comparing with?'. She searched both options in her corpus of 523,427 words (52 RAs) and found fifteen instances of *compared with*, but no hits for *comparing with*. From these results she concluded, 'comparing with is not used (incorrect)'.

[1] Student names are pseudonyms to preserve their anonymity.

Jialing, a Chinese student of medicine, initially focused on the preposition following *compare*, asking: 'compare with vs. compare to' and making a search on *compar**, which returned 105 hits from her corpus of 506,361 words (56 RAs). The asterisk * is a wild card which stands for one or more characters, so that this search retrieves all forms that begin *compar*, including *comparing*, *compared*, *comparison* etc. Thus Jialing's investigation revealed the more detailed phraseological patterns of a wider range of forms, which she noted in five categories as follows: '1. *comparable to* 2. *compare with*; *compare to* 3. noun + *compared to* (more often) noun + *compared with* (less often but still common) 4. *comparison of* and 5. i*n comparison with*; *in comparison to*.' Although she did not give the relative frequencies of these patterns, it is clear that this investigation gave Jialing the information she needed to extend and refine her options for expressing comparison.

The third example comes from the work of Ana, a Portuguese student of psychology, who framed her query in terms of the sentence that she wanted to check: 'Movement was reduced comparing to the incongruent condition.' Using her corpus of 377,294 words (43 RAs), she made a search on *compare**, found *compared to* and corrected her sentence as follows: 'Movement was reduced compared to the incongruent condition.' Unlike Jialing, she put the wild card after the form *compare* (with a final 'e'); thus her search was not able to retrieve the form *comparing* (without 'e'); nonetheless, although flawed, her search did enable Ana to find the appropriate form and to correct her sentence.

What these examples show is that individual students are driven by very different concerns and that this may lead them to take different approaches, even when asking the same or similar questions; they may phrase their queries in different ways, make different searches and come up with a range of different answers to satisfy their own purposes at that time. Thus Mitsuko needed to check the form of the verb, while Jialing was interested in the preposition and Ana wanted to correct her sentence. Mitsuko answered her question with two precisely targeted searches and concluded that one of her potential forms was incorrect. Jialing's search, however, was designed to capture a much wider range of phraseology and her results were thus more detailed in terms of identifying frequent patterns. For Ana, the important point was to check her sentence and correct it if necessary. Although she achieved this aim, her search was the least satisfactory of the three in that it could not reveal to her that the form *comparing to* was incorrect.

It should be noted, however, that these were the first searches that the students carried out using their DIY corpus to answer their own queries. Thus, even at an early stage of corpus use, students can make valuable discoveries; at the same time these examples underline the fact that practice is needed in formulating queries to ensure that the information retrieved really does answer the student's question. In this regard, further practice in searching should lead to better search techniques and improved outcomes.

Previous research on the use of DIY corpora has shown that students consider the main advantage of the DIY corpus to be its high level of discipline-specificity, which enables their individual needs to be addressed with greater precision and accuracy (Charles 2017). Although the students on this course performed more general academic than discipline-specific searches, their examination of the corpus data constantly exposed them to the specialized discourse of their own discipline. Moreover, as they had control over the contents of their DIY corpora, students gained a sense of ownership over the resource, which is not possible to achieve using a pre-constructed corpus.

4. CONCLUSION

This paper has argued that DIY corpora provide a viable alternative to pre-constructed on-line corpora, especially for those who work in highly specialized areas. While the compilation of a DIY corpus requires an initial input of time, recent improvements in software mean that this factor has become less significant and a commitment of time at the corpus construction stage is likely to save time in the long run. For indirect use by teachers and course developers, a DIY corpus can provide access to discipline-specific texts in a way that makes it easy to investigate the discourse and create appropriate materials for a given class. For direct use by individual graduate students or researchers, the DIY corpus can correspond exactly to their writing needs, providing the information they require to compose discipline-specific texts that are acceptable at the highest academic level. I would suggest that once users have compiled a DIY corpus and experienced its benefits, they are equipped with a skill which can stand them in good stead as they progress throughout their career, enabling them to adapt their corpora or build new ones to reflect their changing research or teaching concerns.

REFERENCES

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

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

Bianchi, Francesca, and Roberto Pazzaglia. 2007. 'Student Writing of Research Articles in a Foreign Language: Metacognition and Corpora'. In *Corpus Linguistics 25 Years On*, edited by Roberta Facchinetti, 261–87. Amsterdam: Rodopi.

Boulton, Alex, and Tom Cobb. 2017. 'Corpus Use in Language Learning: A Meta-Analysis'. *Language Learning* 67 (2): 348–93.

Bunton, David. 1999. 'The Use of Higher Level Metatext in PhD. Theses'. *English for Specific Purposes* 18 (S): S41–56.

———. 2002. 'Generic Moves in PhD Thesis Introductions'. In *Academic Discourse*, edited by John Flowerdew, 57–75. London: Longman.

———. 2005. 'The Structure of PhD Conclusion Chapters'. *Journal of English for Academic Purposes* 4 (3): 207–24.

Charles, Maggie. 2012. '"Proper Vocabulary and Juicy Collocations": EAP Students Evaluate Do-It-Yourself Corpus-Building'. *English for Specific Purposes* 31 (2): 93–102.

———. 2014. 'Getting the Corpus Habit: EAP Students' Long-Term Use of Personal Corpora'. *English for Specific Purposes* 35: 30–40.

———. 2017. 'Do-It-Yourself Corpora in the Classroom: Views of Students and Teachers'. In *Faces of English Education: Students, Teachers and Pedagogy*, edited by Lillian Wong, and Ken Hyland, 107–23. Abingdon: Routledge.

Chen, Meilin, and John Flowerdew. 2018. 'Introducing Data-Driven Learning to PhD Students for Research Writing Purposes: A Territory-Wide Project in Hong Kong'. *English for Specific Purposes* 50: 97–112.

Davies, Mark. 2008. *The Corpus of Contemporary American English (COCA): 560 Million Words, 1990-Present.* Available online at https://corpus.byu.edu/coca/.

Flowerdew, Lynne. 2012. ‘Exploiting a Corpus of Business Letters from a Phraseological, Functional Perspective’. *ReCALL* 24 (2): 152–68.

Hyland, Ken. 2005. *Metadiscourse*. London: Continuum.

Leech, Geoffrey. 1997. ‘Teaching and Language Corpora: A Convergence’. In *Teaching and Language Corpora*, edited by Anne Wichman, Steven Fligelstone, Tony McEnery, and Gerry Knowles, 1–23. London: Longman.

Swales, John M. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

———. 2004. *Research Genres: Exploration and Applications*. Cambridge: Cambridge University Press.