16
CORPUS STUDIES IN EAP

Hilary Nesi

Introduction

Corpora are collections of naturally occurring language data, stored in electronic form, designed to be representative of particular types of text and analysed with the aid of computer software tools. Corpora are now common in English for academic purposes (EAP) research and practice, both to provide quantitative information about discourse, and to corroborate insights derived from more qualitative studies. They also play an increasingly important role in EAP pedagogy, providing syllabus items, examples to illustrate accepted usage, and opportunities for data-driven learning.

Most EAP corpus-related activity includes a comparative element, for example differentiating texts belonging to different disciplinary domains (e.g. Hyland, 2008), or cultural contexts (e.g. Gardezi & Nesi, 2009), exploring variation between EAP materials and the texts students encounter in the disciplines (e.g. Chen, 2010), noting signs of progress in texts produced across stages of study (e.g. Issitt, 2011), or comparing texts produced by novices and experts (e.g. Cortes, 2004; Gilquin & Paquot, 2008) and/or by L1 and L2 users (e.g. Chen & Baker, 2010; Ädel & Erman, 2012; Carrió-Pastor, 2013; Pérez-Llantada, 2014).

It is easy to get started on a corpus investigation; a simple yet discipline-specific corpus can be compiled quite quickly using students' files or resources from the web, and lists of lexical items, singly and in combination, can be generated almost instantly, using free downloadable corpus query software. Corpora of spoken academic language or handwritten academic texts require more resources, and more elaborate searches require more elaborate annotation of corpus contents, but all types of corpus investigation have their place in EAP practice; the appeal is that almost any search of any academic corpus can reveal information that is genuinely new to even the most experienced EAP practitioner.

This chapter will provide an overview of the types of corpora most relevant to EAP practitioners and their students, and will consider some of the many ways in which corpora can inform understanding of academic discourse, from lexical, phraseological, grammatical, and genre perspectives.

Types of EAP corpora

The earliest corpora of interest to the EAP community were created by publishers, at considerable cost, for lexicographical purposes. For example, the Bank of English was created at Birmingham University in the 1980s to provide information for the *Collins COBUILD English Language Dictionary* (Sinclair, 1987). Such corpora were large enough to provide thousands of examples of the more frequent lexical items, but were not designed for EAP researchers who wanted to explore the features of specific types of academic text. Since then, many specialist corpora have been created quite cheaply by individuals and small teams, and it is probably safe to say that most of the corpora now referred to in EAP research are relatively small and for private use only.

Corpora can only be shared publicly if the owners of the texts have given their permission for their use in this way, and most academics and teachers do not have the time and the resources to arrange this, or to create the documentation that a publicly available package would require. However, specialist mini-corpora are often more useful than large general corpora for investigating a specific domain, as Tribble (1997) points out. Baker (2006, p. 25), Walsh (2013, p. 40), and others have argued that the process of creating one's own corpus, rather than working with one that is ready-made, helps to give the researcher a 'feel' for the data, a sense of context, and the means to generate initial hypotheses as the first stage of corpus research. Small private corpora have been used to reveal patterns of academic language use in a wide range of genres, for example doctoral theses (Thompson, 2000; Charles, 2003, 2006), student–tutor interactions (Farr, 2003), textbooks (Mudraya, 2006; Bondi, 2012, Wood & Appel, 2014), academic bios (Tse, 2012), and research abstracts (Cava, 2011; Cutting, 2012).

Copyright restrictions make it difficult to share corpora created from academic journal content, but nevertheless a large amount of EAP corpus research and practice has focussed on research articles, because they are easy to collect in electronic form, can be selected to represent highly specific research domains, and can yield findings of great relevance to certain EAP contexts. One way of introducing corpora to the EAP classroom is to ask students to make collections of published research articles and their own unpublished writing. Students can compare research articles with their summaries of these articles, for example (Seidlhofer, 2000), or their own research reports with articles they have selected to suit their disciplinary areas (Lee and Swales, 2006). Some collections of research articles are widely known and have acquired status despite the fact that they are not publicly available: the 1.4 million word 'Hyland Corpus', for example, is referenced in many of Ken Hyland's investigations into the nature of disciplinary discourse (as in Hyland, 2005, 2009, 2012), but is only available to Hyland and his close associates.

In addition to countless home-made corpora of research articles, there exist some corpora of professional academic writing that have resolved copyright issues and are available in the public domain. The most notable of these is the academic component of the Corpus of Contemporary American English (COCA-A) (Davies, 2011; Gardner & Davies, 2014) which is made up of research articles (about 85 million words), magazines (about 31.5 million words), and newspaper finance sections (about 7.5 million words) but does not permit searches limited to a specific source, in order not to 'infringe on the domain of other resources ….which are oriented toward searching specific journals' (Davies, 2014, p. 164). The 17 million word Professional English Research Consortium (PERC) corpus of science and technology research articles, developed by a consortium with members in Japan, the UK, and the USA, contains output from nearly 300 journals; it seems that it was possible to make the texts available because the corpus does not contain complete articles, just

samples, and so science and technology researchers are unlikely to use it as a free alternative to journal subscription. Cooke and Birch-Becaas (2008) adopted another strategy to bypass publishers' copyright restrictions; their corpus was made up of drafts donated by the authors (francophone researchers in the health and life sciences) rather than the final published documents, and they describe how one version indicated what the authors originally wrote, alongside the corrections, editing, and reformulations that they later made in consultation with an L1 advisor, while other versions drew attention to lexicogrammatical features and the move structure of research articles. For obvious reasons, discourse structure is best investigated in corpora which contain entire texts.

Ädel (2006, pp. 206–207) debates whether corpora of professional writing should be the reference point for EAP activities, as although professional writing 'represents the norm that advanced foreign learner writers try to reach and their teachers try to promote', the writing of proficient L1 students may constitute a more realistic model because it is at a comparable educational level. It should be borne in mind that in many contexts students are not expected to write in the same way as established writers, for example with regard to the confidence of their claims or the strength of their conclusions (Lee and Swales, 2006, p. 68). On the whole, the findings of research article corpora seem most relevant in cases where EAP is taught to graduates preparing for research careers; at lower levels the writing normally expected of students may have a different communicative purpose, and evidence from corpora of research articles needs to be applied with caution.

Private corpora

For commercial reasons, most EAP practitioners will never be able to access some of the largest academic corpora, as they have been developed by testing and publishing companies and are only available to writers and researchers working on authorised projects. Varying amounts of information are available for these corpora. Some, such as the TOEFL 2000 Spoken and Written Academic Language (T2K-SWAL) corpus, are described in detail in publicly available reports. T2K-SWAL is the property of the Educational Testing Service, but Biber et al. (2004b) and Biber (2006) provide a full account of its design, construction, and analysis. Other large commercial corpora are introduced in conference proceedings or in research articles. The Pearson International Corpus of Academic English (PICAE), for example, is described in Ackermann et al. (2011), and The Cambridge and Nottingham Corpus of Academic Discourse (CANCAD) features in an article by Evison (2012). Still others are known only in outline: very little has been publicly revealed about the Cambridge Academic English Corpus, for example, other than that it contains more than 400 million words of published writing and teaching discourse, and is, according to the Cambridge University Press website, 'the largest and most extensive corpus of Academic English to date'. Currently, Cambridge University Press is inviting writers to contribute to a new academic writing component of the Cambridge Academic English Corpus, but so far no account of this venture seems to have been published.

The downside to all private corpora is that they can lead to duplication of effort: the same kinds of academic texts are collected and stored in different institutions, only to be analysed by a handful of researchers in each case. Lack of public access also means that many published corpus studies which could be strengthened by replication cannot be replicated, and there is no way for readers to check reported findings. However, of course, some EAP corpora are only really relevant to their own specific context, and were not compiled with the intention of deriving general insights about academic language. Issitt (2011), for example, created a

small private corpus of international student writing at a single institution, and compared the work the students produced at the beginning and end of their pre-sessional programme in order to gauge the programme's success. In this case, although the corpus findings were of concern to the people involved with the particular pre-sessional programme, it is the corpus design and the procedures of analysis that other EAP practitioners will wish to replicate.

Publicly accessible corpora

By far the largest EAP corpus in the public domain is COCA-A, which is divided into nine broad disciplinary areas: business and finance, education, history, humanities, law/political science, medicine and health, philosophy/religion/psychology, science and technology, and social science. However, although the powerful COCA interface enables comparisons across these areas, and across all the COCA domains, it is not possible to drill down to specific disciplines or subtypes of academic writing.

Other smaller but more specific academic corpora include the British Academic Written English (BAWE) corpus (Alsop & Nesi, 2009; Nesi & Gardner, 2012) and the British Academic Spoken English (BASE) corpus (Thompson & Nesi, 2001), both of which can be downloaded from the Oxford Text Archive or searched via the open-access SketchEngine interface; and the Michigan Corpus of Upper-level Student Papers (MICUSP) (Römer & Swales, 2010; Römer & O'Donnell, 2011) and the Michigan Corpus of Spoken Academic English (MICASE) (Simpson et al., 2002), both of which have interfaces at the University of Michigan.

BAWE and MICUSP are corpora of proficient student assignments, awarded high grades by subject tutors, although BAWE is bigger (6.5 million words) and focusses on the writing produced by first to final year undergraduates and postgraduates on taught Masters courses, whereas MICUSP (2.6 million words) focusses on the writing produced in the final undergraduate year and three levels of postgraduate study. This design difference is a reflection of higher education practices in the two countries: British undergraduates produce more written assessed work in the early years, but there are few formal written assessment tasks for British postgraduates beyond Masters level; doctoral students in the USA are given more written assignments which are independent of their final dissertation. The two corpora are also not entirely comparable in terms of text distribution. The BAWE corpus draws on more than twice as many disciplines as MICUSP (34+ and 16, respectively), and identifies almost twice as many writing genres, including some that are intended to prepare undergraduates for the world of work as opposed to further study ('preparing for professional practice', as described by Nesi and Gardner, 2012).

Nevertheless, the MICUSP categories of argumentative essay, creative writing, critique/evaluation, proposal, report, and research paper are not dissimilar to the essay, narrative recount, critique, proposal, methodology recount, and research report genre families in BAWE, and the two corpora provide scope for useful comparisons of student writing conventions. Whereas international journals have their own specific requirements which researchers must conform to wherever they are based, there is considerable regional as well as disciplinary variation in the requirements for student writing. These differences are highly relevant for EAP learners, who in order to be successful have to write in a style that suits the local context. For this reason, subcomponents of MICUSP and/or BAWE have also been compared with small corpora representing discipline-specific student writing in English from other regions of the world. Such corpora include a collection of writing by Pakistani economics students (Gardezi & Nesi, 2009), a corpus of Chinese undergraduate dissertations (Lee & Chen, 2009),

the Hanken Corpus of assignments by Finnish- and Swedish-speaking economics students (Hiltunen & Mäkinen, 2014), and the French and Norwegian components of the Varieties of English for Specific Purposes database (VESPA) (Paquot et al., 2013).

Although the majority of assignments in BAWE and MICUSP were written by users of English as a first language, both corpora also contain a substantial proportion of texts written by users of English as an academic lingua franca (18 per cent in the case of MICUSP and about 30 per cent in the case of BAWE). This makes it possible to compare L1 and L2 writing within these corpora; Chen and Baker (2010) and Leedham (2014) examined BAWE assignments produced by Chinese and L1 English students, for example. Neither corpus was specifically designed to compare L1 and L2 output so care has to be taken to ensure that if such comparisons are made, the L1 and L2 samples belong to the same disciplines, genres, and levels of study.

Because not all their contributors were users of English as a first language, BAWE and MICUSP might qualify for classification as corpora of academic English as a lingua franca (ELF), like VESPA, a growing collection of university student assignments produced in different national contexts (see Paquot et al., 2013), and also the written academic ELF (WrELFA) corpus, containing proficient academic ELF in the form of examiners' statements and science blogs (Carey, 2013). WrELFA is rather unusual in the field of academic ELF because corpora designed to examine the language of expert ELF users tend to focus on speech, which is more likely to retain evidence of regional variation, lost in edited written texts. In contrast, many academic corpora of L2 writing are 'learner corpora', containing assignments produced for English language courses, and intended primarily to aid the identification of learner language deficit rather than the linguistic features of academic genres and disciplines (see for example the HKUST Corpus of Learner English (Milton, 1998), the International Corpus of Learner English (Gilquin et al., 2007), and the Corpus of Academic Learner English (CALE) (Callies & Zaytseva, 2011)).

The best known ELF corpora are the Vienna-Oxford International Corpus of English (VOICE) (Seidlhofer, 2001, 2012) and the English as a Lingua Franca in Academic Settings (ELFA) corpus (Mauranen, 2003, 2012), both containing about 1 million words of naturally occurring, non-scripted, face-to-face interactions between speakers from a variety of first-language backgrounds. VOICE concentrates on dialogic speech events from three domains (professional, educational, and leisure), while ELFA contains solely academic speech events, roughly two-thirds dialogic (seminars, thesis defences, conference discussions, etc.) and one third monologic (lectures and presentations).

MICASE and BASE also have dialogic and monologic components. They are roughly equal in size (about 1.7 and 1.5 million words respectively), but MICASE captures a wide variety of spoken academic genres such as lectures, dissertation defences, office-hour interactions, and small peer-led study group sessions, whereas BASE focusses on lectures and seminars, the former led by academics and the latter by students. They have been examined, together and separately, in a number of studies, for example in terms of modifiers (Poos & Simpson, 2002; Swales & Burke, 2003; Lin, 2012), formulaic language and lexical bundles (Nesi & Basturkmen, 2009; Simpson-Vlach & Ellis, 2010), and lexicogrammatical patterns (Deroey, 2011; Deroey & Taverniers, 2012).

Perhaps because of difficulties in obtaining recording and transcription rights, there are few publicly available corpora of conference presentations, although the John Swales Conference Corpus (JSCC) contains transcripts of papers on applied linguistics topics (see Wulff et al., 2009). Doctoral students and research-active university staff would certainly benefit from access to more corpora of this kind, relating to more academic disciplines.

Methods of analysing academic corpora

Academic corpora are used to inform our understanding of academic discourse, from lexical, grammatical, phraseological, and genre perspectives. However, although corpus linguists often make qualitative judgements about meaning and communicative functions, the software tools they use are essentially quantitative, performing calculations based on the frequency of specified lexicogrammatical items.

Frequency counts can be used to generate various kinds of academic wordlists, ranging from simple itemisations of every word form to more complex comparisons within and between corpora, and lists of terms restricted to specific domains. The best known of these is still the Academic Wordlist (AWL) (Coxhead, 2000, 2011). This was a successor to the University Word List (Xue & Nation, 1984), and was generated from a corpus of texts from the arts, commerce, law, and science using the Range corpus analysis programme (Heatley & Nation, 1996). The AWL groups words into 570 'word families' which occurred in all domains and in 15 or more disciplines, excluding the most frequent 2000 words in the General Service List (West, 1953). The corpus was small by modern standards (3.5 million words) and only represented reading material as opposed to speech or student writing. Nevertheless, Coxhead's technique of eliminating both frequent and discipline-specific words identified a 'subtechnical' vocabulary set which covers about 10 per cent of the tokens in many other written academic corpora (Cobb & Horst, 2004; Ward, 2009; Chen & Ge, 2007; Li & Qian, 2010), even if coverage across disciplines has been found to vary (Hyland & Tse, 2007), and it accounts for only 4.41 per cent of tokens in the BASE corpus (Dang & Webb, 2014). A large number of resources have drawn on the AWL to describe and teach academic vocabulary, including Haywood's AWL Highlighter and Gapmaker (n.d.), and the AWL tools on Cobb's Compleat Lexical Tutor website (n.d.).

More recent wordlists have applied more sophisticated corpus analysis techniques to identify core academic vocabulary. Paquot (2010), for example, created an Academic Keyword List (AKL) of 930 items, which were significantly more frequent in a corpus of academic writing (consisting of professional texts and student writing from the Louvain Corpus of Native Speaker Essays and the BAWE Pilot Corpus) when compared to a reference corpus of fiction writing. The Academic Vocabulary List (AVL) (Gardner & Davies, 2014) was compiled using a similar method, comparing COCA-A with COCA as a whole. Likewise, the Academic Formulas List (AFL) (Simpson-Vlach and Ellis, 2010) extracted multi-word units (lexical bundles) from a corpus including MICASE, and a selection of research articles and academic texts from the Briish National Corpus (BNC). Simpson-Vlach and Ellis adapted the procedure for identifying bundles pioneered by Douglas Biber and his colleagues (e.g. Biber & Conrad, 1999; Biber et al., 2004a), but in order to ensure that their academic formulas were meaningful and pedagogically relevant, they used not only statistical measures of frequency but also mutual information scores, to identify collocational strength, and the qualitative judgements of experienced EAP instructors and language testers.

Unlike the Academic Wordlist, all these later wordlists included lexical items that were frequent in non-academic corpora, as long as they occurred significantly more often in the academic texts. They used the log-likelihood statistic to measure significant differences in an item's frequency, comparing their academic corpora with their non-academic ('reference') corpora, but they also took into account the fact that log-likelihood does not distinguish between items occurring fairly evenly across corpus subsections, and items occurring very frequently, in only a few subsections. These wordlists only consisted of items that were distributed across a range of disciplinary areas. In the case of AKL and AVL, they also required items to be well dispersed across individual texts in each area.

AKL, AVL, and AFL focus on core academic vocabulary, but there are strong arguments for a greater EAP focus on discipline-specific language, both because EAP learners experience difficulty with technical terms (Liu & Nesi, 1999; Evans & Morrison, 2011), and because words behave differently in different disciplinary domains, in terms of frequency and range, and also in terms of meaning and collocation (Hyland & Tse, 2007). Many recent corpus-derived wordlists have taken a discipline-specific approach, concentrating on texts from areas such as engineering (Mudraya, 2006; Ward, 2009), agriculture (Martinez et al., 2009), science (Coxhead & Hirsh, 2007), medicine (Hsu, 2013), and chemistry (Valipouri & Nassaji, 2013). Comparisons between disciplinary vocabulary requirements have also been made by means of corpus techniques. Cortes (2004) examined lexical bundles in history and biology, for example, and Hyland (2008) examined lexical bundles in applied linguistics, biology, business, and electrical engineering. Durrant (2014) used Gries's deviation of proportions (DP) statistic (Gries, 2008) and hierarchical cluster-analysis (Durrant, 2009) to capture variation in the distribution of words in the BAWE corpus, finding that although it is possible to identify clusters of students with similar vocabulary needs across different disciplines and levels, 'only around half of the words that are important for particular groups of students are generic to the writing of other groups' (Durrant, 2014, p. 14).

Many EAP-relevant corpus studies start by identifying lexical items or lexicogrammatical patterns, and then examine these in corpus output. Hyland (2004), for example, describes a process of manually coding sample texts to identify metadiscursive items, and then searching for them in a corpus of masters and doctoral dissertations. Alternatively, corpus studies may take as a starting point lists of potentially productive items developed by earlier researchers, such as Hyland (2005), who looked for interactive features in research articles, and Liu (2012), who looked for core academic bundles in sub-corpora of COCA and the BNC.

Multidimensional analysis (MDA) takes rather a different approach: instead of analysing the behaviour of specified linguistic features, it focusses on groups of texts representing registers or language varieties, and examines how they are characterised by combinations of these features. The MDA methodology, pioneered by Douglas Biber, was first used to distinguish between broad domains of use, most notably speech and writing (Biber, 1988), but has since been applied to academic corpora such as the T2K-SWAL (Biber & Gray, 2010), BAWE (Nesi & Gardner, 2012), MICUSP (Hardy & Römer, 2013), and collections of research articles (Gray, 2013). These investigations have revealed variation in the clustering patterns of linguistic features across genres, disciplines, and levels of study, with many implications for EAP. In MICUSP, for example, MDA revealed linguistic differences between writing in the sciences and the humanities (Hardy & Römer, 2013), and in the BAWE corpus, students were found to write in a less context-dependent and more informational way as they progressed through their degree programmes, with a gradual decrease in narrative and persuasive features (Nesi & Gardner, 2012, pp. 13–14).

Biber's method of MDA requires the use of an elaborate automatic annotation system to identify all the linguistic features that might contribute to the factor analysis (see Biber, 1988). For other types of analysis, corpora can be marked-up to varying degrees. Annotation can be undertaken by hand, automatically, or with some assistance from a computer programme, and can tag features relating to phonology (pronunciation or prosodic features), lexis (lemmas,[1] parts-of-speech (POS) tags, or semantic characteristics), syntax (parsing), and/or discourse features (co-reference relations, functions, or stylistic characteristics). In EAP studies, automatic lexical annotation is common: lemmatised forms are used for the creation of wordlists, and POS and semantic tagging allows researchers to search more widely for words representing specified parts of speech or semantic fields. A search for

proper noun + number within brackets + lexical or modal verb using the CLAWS POS system (constituent likelihood automatic word-tagging system; Garside & Smith, 1997), for example, will find integral citations in written academic texts, enabling the investigation of academic referencing practices (Nesi, 2014). A search for semantic tags using the USAS category system (UCREL semantic analysis system; Archer et al., 2002) reveals that words relating to DECIDING, WANTING, PLANNING, and CHOOSING are particularly common in BAWE corpus genres which prepare students for the world of work (Nesi & Gardner, 2012, p. 203).

Although some superficial features beyond sentence level, such as pronominal reference, can be identified automatically, annotation at the level of discourse is generally less advanced and has received less attention from EAP researchers. Annotation of this kind must still for the most part be done manually, and is thus too time-consuming for most EAP practitioners to undertake on a large scale. Moreover, some corpus linguists, such as Sinclair (2004, p. 191), have argued against annotating functions of discourse, on the grounds that this imposes particular interpretations of the text on the corpus user. However, in order to create useful teaching materials, it is necessary for EAP practitioners to interpret speakers' and writers' communicative intent, and for this reason discourse-level annotation has great potential within EAP, for example in the area of automated writing evaluation. Cotos (2011) describes how annotating research article introductions according to Swales' 'Create a Research Space' (CARS) model led to the development of the Intelligent Academic Discourse Evaluator, a programme that evaluates draft articles and provides feedback to novice writers.

We can expect more complex corpus-based pedagogical applications in the future, but also more revelations from simple, plain text corpora, which will continue to provide EAP-relevant information without the use of any form of annotation, for example relating to collocations, domain-specific words, and phraseology. The EAP materials created by Tim Johns in the 1980s, using text files and a simple concordancing programme, are still relevant today (see Johns, 1991, and http://lexically.net/TimJohns/). Despite advances in technology and corpus query techniques, the insights of the human analyst are likely to remain the top resource in EAP for a long time to come.

Further reading

Biber (2006); Hyland (2012); Nesi & Gardner (2012)

Related chapters

14 Acquiring academic and disciplinary vocabulary
19 Genre analysis
21 Intercultural rhetoric
28 PhD defences and vivas

Note

1 A lemma is the uninflected form of a word that can be used as a headword for a dictionary entry. For example, *believe* is the lemma representing *believe, believes, believed*, and *believing*.

References

Ackermann, K., De Jong, J.H.A.L., Kilgarriff, A. & Tugwell, D. (2011) The Pearson International Corpus of Academic English (PICAE). Paper 47. *Proceedings of the Corpus Linguistics Conference*. [Online] Available from: www.birmingham.ac.uk/research/activity/corpus/publications/conference-archives/2011-birmingham.aspx [Accessed: 1st November 2014].

Ädel, A. (2006) *The Use of Metadiscourse in Argumentative Texts by Advanced Learners and Native Speakers of English*. Amsterdam: John Benjamins.

Ädel, A. & Erman, B. (2012) Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes.* 31 (2). 81–92.

Alsop, S. & Nesi, H. (2009) Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora.* 4 (1). 71–83.

Archer, D., Wilson, A. & Rayson, P. (2002) *Introduction to the USAS Category System.* [Online] University Centre for the Computer Corpus Research on Language. Lancaster University, UK. Available from: http://ucrel.lancs.ac.uk/usas/usas%20guide.pdf [Accessed: 1st November 2014].

Baker, P. (2006) *Using Corpora in Discourse Analysis.* London: Continuum.

Biber, D. (1988) *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Biber, D. (2006) *University Language: A Corpus-Based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.

Biber D. & Conrad, S. (1999) Lexical bundles in conversation and academic prose. In Hasselgard, H. & Oksfjell, S. (eds). *Out of Corpora: Studies in Honor of Stig Johansson* (pp. 181–189). Amsterdam: Rodopi.

Biber, D. & Gray, B. (2010) Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes.* 9 (1). 2–20.

Biber, D., Conrad, S. & Cortes, V. (2004a) If you look at … : Lexical bundles in university teaching and textbooks. *Applied Linguistics.* 25 (3). 371–405.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E. & Urzua, A. (2004b). *Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Writing Academic Language Corpus* (ETS TOEFL Monograph Series, MS-25). Princeton, NJ: Educational Testing Service.

Callies, M. & Zaytseva, E. (2011) The *Corpus of Academic Learner English* (CALE): A new resource for the study of lexico-grammatical variation in advanced learner varieties. In Hedeland, H., Schmidt, T. & Wörner, K. (eds). *Multilingual Resources and Multilingual Applications* (Hamburg Working Papers in Multilingualism B 96) *Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology* (GSCL) (pp. 51–56). Hamburg: University of Hamburg.

Carey, R. (2013) On the other side: formulaic organizing chunks in spoken and written academic ELF. *Journal of English as a Lingua Franca*. 2 (2). 207–228.

Carrió-Pastor, M. L. (2013) A contrastive study of the variation of sentence connectors in academic English. *Journal of English for Academic Purposes*. 12 (3). 192–202.

Cava, A. M. (2011) Abstracting science: A corpus-based approach to research article abstracts. *International Journal of Language Studies*. 5 (3). 75–98.

Charles, M. (2003) 'This mystery…': A corpus-based study of the use of nouns to construct stance in theses from two contrasting disciplines. *Journal of English for Academic Purposes*. 2(4). 313–326.

Charles, M. (2006) The construction of stance in reporting clauses: A cross-disciplinary study of theses. *Applied Linguistics*. 27(3). 492–518.

Chen, L. (2010) An investigation of lexical bundles in ESP textbooks and electrical engineering introductory textbooks. In Wood, D. (ed.). *Perspectives on Formulaic Language: Acquisition and Communication* (pp. 107–125). New York/London: Continuum.

Chen, Q. & Ge, C. (2007) A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). *English for Specific Purposes*. 26 (4). 502–514.

Chen, Y-H. & Baker, P. (2010) Lexical Bundles in L1 and L2 Academic Writing. *Language Learning and Technology*. 14 (2). 30–49.

Cobb, T. (n.d.). *The Compleat Lexical Tutor*. Available from: www.lextutor.ca/ [Accessed: 1st November 2014].

Cobb, T. & Horst, M. (2004). Is there room for an AWL in French? In Bogaards, P. & Laufer, B. (eds). *Vocabulary in a Second Language: Selection, Acquisition, and Testing* (pp. 15–38). Amsterdam: John Benjamins.

Cooke, R. & Birch-Becaas, S. (2008) Help on the spot: Online assistance for writing scientific English. *LSP and Professional Communication*. 8 (2). 96–111.
Cortes, V. (2004) Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes.* 23 (4). 397–423.
Cotos, E. (2011) Potential of automated writing evaluation feedback. *CALICO Journal*. *28* (2). 420–459.
Coxhead, A. (2000) A new Academic Word List. *TESOL Quarterly*. 34 (2). 213–239.
Coxhead, A. (2011) The Academic Word List 10 years on: Research and teaching implications. *TESOL Quarterly*. 45 (2). 355–362.
Coxhead, A. & Hirsh, D. (2007) A pilot science word list for EAP. *Revue Française de Linguistique Appliqueé.* XII (2). 65–78.
Cutting, J. (2012) Vague language in conference abstracts. *Journal of English for Academic Purposes*. 11 (4). 283–293.
Dang, T. N. Y. & Webb, S. (2014) The lexical profile of academic spoken English. *English for Specific Purposes. 33*. 66–76.
Davies, M. (2011) The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*. 25 (4). 447–464.
Davies, M. (2014) Google Scholar and COCA-Academic: Two very different approaches to examining academic English. *Journal of English for Academic Purposes.* 12 (3). 155–165.
Deroey, K. L. B. (2011) What they highlight is…: The discourse functions of basic wh-clefts in lectures. *Journal of English for Academic Purposes.* 11 (2). 112–124.
Deroey, K. L. B. & Taverniers, M. (2012) 'Just remember this': Lexicogrammatical relevance markers in lectures. *English for Specific Purposes*. 31 (4). 221–233.
Durrant, P. (2009) Investigating the viability of a collocation list for students of English for academic purposes. *Journal of English for Specific Purposes.* 28 (3). 157–179.
Durrant, P. (2014) Discipline and level specificity in university students' written vocabulary. *Applied Linguistics*. 35 (3). 328–356.
Evans, S. & Morrison B. (2011) The first term at university: Implications for EAP. *ELT Journal.* 65 (4). 387–397.
Evison, J. (2012) A corpus linguistic analysis of turn-openings in spoken academic discourse: Understanding discursive specialisation. [Online] *English Profile Journal*. 3. e4. Available from: http://journals.cambridge.org/action/displayIssue?jid=EPJ&volumeId=3&seriesId=0&issueId=-1 [Accessed:1st November 2014].
Farr, F. (2003) Engaged listenership in spoken academic discourse: The case of student–tutor meetings. *Journal of English for Academic Purposes.* 2 (1). 67–85.
Gardezi, S. A. & Nesi, H. (2009) Variation in the writing of economics students in Britain and Pakistan: the case of conjunctive ties. In Charles, M., Hunston, S. & Pecorari, D. (eds). *Academic Writing: At the Interface of Corpus and Discourse* (pp. 236–250). London: Continuum.
Gardner, D. & Davies, M. (2014) A new Academic Vocabulary List. *Applied Linguistics.* 35 (3). 305–327.
Garside, R. & Smith, N. (1997) A hybrid grammatical tagger: CLAWS4. In Garside, R., Leech, G. & Mcenery, A. (eds). *Corpus Annotation:Linguistic Information from Computer Text Corpora* (pp. 102–121). London: Longman.
Gilquin, G. & Paquot, M. (2008) *Too Chatty: Learner Academic Writing and Register Variation*. *English Text Construction*. 1 (1). 41–61.
Gilquin, G., Granger, S. & Paquot, M. (2007) Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes.* 6 (4). 319–335.
Gray, B. (2013) More than discipline: Uncovering multi-dimensional patterns of variation in academic research articles. *Corpora.* 8 (2). 153–182.
Gries, S. T. (2008) Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*. 13 (4). 403–37.
Hardy, J. & Römer, U. (2013) Revealing disciplinary variation in student writing: A multi-dimensional analysis of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora.* 8 (2). 183–208
Haywood, S. (n.d.) *Academic Vocabulary*. [Online] Available from: www.nottingham.ac.uk/alzsh3/acvocab/index.htm [Accessed: 1st November 2014].
Heatley, A. & Nation, P. (1996) *Range.* [Online] Wellington, New Zealand: Victoria University of Wellington. Available from: www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx [Accessed: 1st November 2014].
Hiltunen, T. & Mäkinen, M. (2014) Formulaic language in economics papers: Comparing novice and published writing. In Gotti, M. & Giannoni, D. S. (eds). *Corpus Analysis for Descriptive and Pedagogic Purposes: English Specialised Discourse* (pp. 347–368). Bern: Peter Lang.

Hsu, W. (2013) Bridging the vocabulary gap for EFL medical undergraduates: The establishment of a medical word list. *Language Teaching Research.* 17 (4). 454–484.
Hyland, K. (2004) Disciplinary interactions: Metadiscourse in L2 postgraduate writing. *Journal of Second Language Writing*. 13 (2). 133–151.
Hyland, K. (2005) *Metadiscourse.* London: Continuum.
Hyland, K. (2008) As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*. 27 (1). 4–21.
Hyland, K. (2009) *Academic Discourse*. London: Continuum.
Hyland, K. (2012) *Disciplinary Identities: Individuality and Community in Academic Discourse.* Cambridge: Cambridge University Press.
Hyland, K. & Tse, P. (2007) Is there an 'academic vocabulary'? *TESOL Quarterly*. 41 (2). 235–253.
Issitt, S. (2011) How an L2 learner corpus can identify areas of quantifiable improvement in students' written discourse. *Proceedings of the CL2011 conference*, Birmingham 20–22 July 2011. [Online] Available from: www.birmingham.ac.uk/research/activity/corpus/publications/conference-archives/2011-birmingham.aspx [Accessed: 1st November 2014].
Johns, T. (1991) Should you be persuaded – two examples of data-driven learning materials. In Johns, T. & King, P. (eds). *English Language Research Journal.* 4. 1–16.
Lee, D. & Chen, X. (2009) Making a bigger deal of smaller words: Function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing.* 18 (4). 281–296.
Lee, D. Y. W. & Swales, J. (2006) A corpus-based EAP course for NNS doctoral students: Moving from available specialised corpora to self-compiled corpora. *English for Specific Purposes.* 25 (1). 56–75.
Leedham, M. (2014) *Chinese Students' Writing in English: Implications from a Corpus-Driven Study*. London: Routledge.
Li, Y., & Qian, D. D. (2010) Profiling the academic word list (AWL) in a financial corpus. *System*. 38 (3). 402–411.
Lin, C-Y (2012) Modifiers in BASE and MICASE: A matter of academic cultures or lecturing styles? *English for Specific Purposes.* 31 (2). 117–126.
Liu, D. (2012) The most frequently-used multi-word constructions in academic written English: A multi-corpus study. *English for Specific Purposes.* 31 (1). 25–35.
Liu, J. & Nesi, H. (1999) Are we teaching the right words? A study of students' receptive knowledge of two types of vocabulary: Subtechnical and technical. In Bool, H. & Luford, P. (eds). *Academic Standards and Expectations: The Role of EAP* (pp. 141–147). Nottingham: Nottingham University Press.
Martinez, I. A., Beck, S. C. & Panza, C. B. (2009) Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes.* 28 (3). 183–198.
Mauranen, A. (2003) The corpus of English as lingua franca in academic settings. *TESOL Quarterly.* 37 (3). 513–527.
Mauranen, A. (2012) *Exploring ELF: Academic English Shaped by Non-Native Speakers*. Cambridge: Cambridge University Press.
Milton, J. (1998). Exploiting LI and interlanguage corpora in the design of an electronic language learning and production environment. In Granger, S. (ed.) *Learner English on Computer* (pp. 186–198). London: Longman.
Mudraya, O. (2006) Engineering English: A lexical frequency instructional model. *English for Specific Purposes*. 25 (2). 235–256.
Nesi, H. (2014) Corpus query techniques for investigating citation in student assignments. In Gotti, M. & Giannoni, D. S. (eds). *Corpus Analysis for Descriptive and Pedagogic Purposes: English Specialised Discourse* (pp. 85–106). Bern: Peter Lang.
Nesi, H. & Basturkmen, H. (2009) Lexical bundles and discourse signalling in academic lectures. In Mahlberg, M. & Flowerdew, J. (eds). *Lexical Cohesion and Corpus Linguistics* (pp. 23–44). Amsterdam: John Benjamins.
Nesi, H. & Gardner, S. (2012) *Genres across the Disciplines: Student Writing In Higher Education*. Cambridge: Cambridge University Press.
Paquot, M. (2010) *Academic Vocabulary in Learner Writing: From Extraction to Analysis*. London: Bloomsbury.
Paquot, M., Hasselgård, H. & Ebeling, S. O. (2013) Writer/reader visibility in learner writing across genres. A comparison of the French and Norwegian components of the ICLE and VESPA learner corpora. In Granger, S., Gilquin, G. & Meunier, F. (eds). *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead: Proceedings of the First Learner Corpus Research Conference* (LCR 2011) (pp. 377–387). Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.

Pérez-Llantada, C. (2014) Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage. *Journal of English for Academic Purposes.* 14 A1–A2. 84–94.

Poos, D. & Simpson, R. (2002) Cross-disciplinary comparisons of hedging: Some findings from the Michigan Corpus of Academic Spoken English. In Reppen, R., Fitzmaurice, S. & Biber, D. (eds). *Using Corpora to Explore Linguistic Variation* (pp. 3–21). Philadephia, PA: John Benjamins.

Römer, U. & Brook O'Donnell, M. (2011) From student hard drive to web corpus (part 1): The design, compilation and genre classification of the *Michigan Corpus of Upper-level Student Papers* (MICUSP). *Corpora.* 6 (2). 159–177.

Römer, U. & Swales, J. (2010) The Michigan Corpus of Upper-level Student Papers (MICUSP). *Journal of English for Academic Purposes*. 9 (3). 249.

Seidlhofer, B. (2000) Operationalising intertextuality: Using learner corpora for learning. In Burnard, L. & Mcenery, T. (eds). *Rethinking Language Pedagogy from a Corpus Perspective* (pp. 207–224). Frankfurt am Main: Peter Lang.

Seidlhofer, B. (2001) Closing a conceptual gap: The case for a description of English as a lingua franca. *International Journal of Applied Linguistics.* 11 (2). 133–158.

Seidlhofer, B. (2012) Corpora and ELF. In Hyland, K., Chau, M. H. & Handford, M. (eds). *Corpus Applications in Applied Linguistics* (pp. 135–149). London: Continuum.

Simpson, R. C., Briggs, S. L., Ovens, J. & Swales, J. M. (2002) *The Michigan Corpus of Academic Spoken English.* Ann Arbor, MI: The Regents of the University of Michigan.

Simpson-Vlach, R. & Ellis, N. (2010) An academic formulas list: New methods in phraseology research. *Applied Linguistics*. 31 (4). 487–512.

Sinclair, J. M. (ed.) (1987) *Collins COBUILD English Language Dictionary.* London: Collins.

Sinclair, J. M. (2004) *Trust the Text*: *Language, Corpus and Discourse*. London: Routledge.

Swales, J. M. (2004) *Research Genres: Exploration and Applications*. Cambridge: Cambridge University Press.

Swales, J. M. & Burke, A. (2003) 'It's really fascinating work': Differences in evaluative adjectives across academic registers. In Leistyna, P. & Meyer, C. F. (eds). *Corpus Analysis: Language Structure and Language Use* (pp. 1–18). New York: Rodopi.

Thompson, P. (2000) Citation practices in PhD theses. In Burnard, L. & McEnery, T. (eds). *Rethinking Language Pedagogy from a Corpus Perspective* (pp. 91–102). Frankfurt: Peter Lang.

Thompson, P. & Nesi, H. (2001) The British Academic Spoken English (BASE) Corpus project. *Language Teaching Research.* 5 (3). 263–264.

Tribble C. (1997) Improving corpora for ELT: Quick and dirty ways of developing corpora for language teaching. In Lewandowska-Tomaszczyk, B. & Melia, P. (eds). *Practical Applications in Language Corpora – Proceedings of PALC '97* (pp. 107–117). Lodz: Lodz University Press.

Valipouri, L. & Nassaji, H. (2013) A corpus-based study of academic vocabulary in chemistry research articles. *Journal of English for Academic Purposes.* 12 (4). 248–263.

Walsh, S. (2013) Corpus linguistics and conversation analysis at the interface: Theoretical perspectives, practical outcomes. In Romero-Trillo, J. (ed.). *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies* (pp. 37–52). New York: Springer.

Ward, J. (2009) A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purpose.* 28 (3). 170–182.

West, M. (1953) *A General Service List of English Words*. London: Longman.

Wood, D. & Appel, R. (2014) Multiword constructions in first year business and engineering university textbooks and EAP textbooks. *Journal of English for Academic Purposes.* 15 A1–A2. 1–13.

Wulff, S., Swales, J. & Keller, K. (2009) 'We have about seven minutes for questions': The discussion sessions from a specialized conference. *English for Specific Purposes.* 28 (2). 79–92.

Xue, G. & Nation, I. S. P. (1984) A university word list. *Language Learning and Communication.* 3 (1). 215–29.