Handbook of Research on Integrating Technology Into Contemporary Language Learning and Teaching

Bin Zou
Xi'an Jiaotong-Liverpool University, China

Michael Thomas
University of Central Lancashire, UK

A volume in the Advances in Educational Technologies and Instructional Design (AETID) Book Series

IGI Global
DISSEMINATOR OF KNOWLEDGE

Published in the United States of America by
IGI Global
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA, USA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Names: Zou, Bin, 1968- editor. | Thomas, Michael, 1969- editor.
Title: Handbook of reasearch on integrating technology into contemporary language learning and teaching / Bin Zou and Michael Thomas, editors.
Description: Hershey, PA : Information Science Reference, [2018] | Includes bibliographical references.
Identifiers: LCCN 2017035886| ISBN 9781522551409 (h/c) | ISBN 9781522551416 (eISBN)
Subjects: LCSH: Language and languages--Study and teaching--Technological innovations. | Second language acquisition--Computer-assisted instruction.
Classification: LCC P53.855 .H373 2018 | DDC 418.0078/5--dc23 LC record available at https://lccn.loc.gov/2017035886

This book is published in the IGI Global book series Advances in Educational Technologies and Instructional Design (AETID) (ISSN: 2326-8905; eISSN: 2326-8913)

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Chapter 25
Automatically Augmenting Academic Text for Language Learning: PhD Abstract Corpora With the British Library

Shaoqun Wu
University of Waikato, New Zealand

Alannah Fitzgerald
Concordia University, Canada

Ian Witten
University of Waikato, New Zealand

Alex Yu
Waikato Institute of Technology, New Zealand

ABSTRACT

This chapter describes the automated FLAX language system (flax.nzdl.org) that extracts salient linguistic features from academic text and presents them in an interface designed for L2 students who are learning academic writing. Typical lexico-grammatical features of any word or phrase, collocations, and lexical bundles are automatically identified and extracted in a corpus; learners can explore them by searching and browsing, and inspect them along with contextual information. This chapter uses a single running example, the PhD abstracts corpus of 9.8 million words derived from the open access Electronic Theses Online Service (EThOS) at the British Library, but the approach is fully automated and can be applied to any collection of English writing. Implications for reusing open access publications for non-commercial educational and research purposes are presented for discussion. Design considerations for developing teaching and learning applications that focus on the rhetorical and lexico-grammatical patterns found in the abstract genre are also discussed.

DOI: 10.4018/978-1-5225-5140-9.ch025

INTRODUCTION

A growing body of research aims to understand the linguistic features of academic text and their bearing on the problems faced by students learning to write. To support this work, corpora of academic writing have been built as a reference and research base. These include lists of academic words, syntactic patterns characteristic of academic writing, and distinctive linguistic characteristics of multi-word sequences that fulfill discourse functions. The research raises many pedagogical implications, and it should be possible to apply the findings to academic teaching practice. How, then, can we bridge the gap between expert and student writing? Suggestions include helping students understand the importance of learning common collocates or recurrent lexical or grammar patterns in different contexts (Coxhead, 2007), making commonly used lexical bundles more accessible (Hafner & Candlin 2007), and providing students with more realistic writing models (Hyland, 2008a).

There are computer tools that help teachers and learners analyze and study language features. Some allow users to upload text and examine the vocabulary it uses. Concordancers (for example, COCA, Compleat Lexical Tutor, and SKELL), initially designed for linguists, are frequently used to explore corpora with a view to exposing linguistic patterns. Some present short snippets of text, while others present items in paragraph-length units; some limit the items that are retrievable. Teachers and students have been using concordances or alike to obtain, organize, and study real language data derived from corpora. This approach is called data-driven learning (Johns, 1991), and is advocated by many researchers (for instance, Boulton & Thomas, 2012; Boulton & Pérez-Paredes, 2014; Chang, 2014; Cobb & Boulton, 2015; Vyatkina, 2016; Boulton & Cobb, 2017).

We have designed and constructed a language learning system called FLAX that takes academic texts, and automatically extracts linguistic features that have been identified in the research literature. Our design is principled and underpinned by two theories: noticing hypothesis (Robinson, 1995; Schmidt, 2001) and inductive (discovery) learning (Bernardini, 2002). Noticing is facilitated through input enhancement and enrichment that has been proven to be effective in students' recognition and recall of language features, for example, collocations (Sharwood, 1993; Sonbul & Schmitt, 2013; Szudarski & Carter, 2016). FLAX presents important components in academic texts−academic words, key concepts, collocations, and lexical bundles−in a way that draws them to the attention of students. External resources (Wikipedia) are linked to these components to gives students opportunities to encounter them in various authentic contexts, and repeatedly. Simple interfaces are developed so that students can use information discovery techniques (e.g., searching and browsing) that they have become familiarized with through search engines (e.g., Google, Bing) to discover and study the language features of their interests.

The aim of this chapter is not to explain how the FLAX system works behind the scenes: that would be a technical discussion that is relatively uninteresting from the point of view of language education. Instead, we aim to illustrate what it does by describing the result of processing the PhD abstract corpora hosted by the British Library. It works entirely automatically, without any human input, and can be applied to any collection of academic text—for example, samples of writing collected by an individual teacher; an entire textbook; or essays written by students (provided that any texts intended for use are all available electronically).

This chapter is structured as follows. The next section summarizes the research background, including a brief survey of existing corpora and their interfaces, and research on the identification and use of wordlists, collocations and lexical bundles for teaching and learning. The next section describes the PhD abstract corpora that are used as examples throughout the chapter. This section will present a brief

review of the literature on the genre and functions of abstracts in academic text. A discussion will also be presented on the reuse of digital open access content managed by the British Library for non-commercial research and development projects. The PhD Abstract corpora in FLAX provide an example of digital collections reuse with the open access PhD theses managed by EThOS at the British Library following changes in legislation for text and data mining for non-commercial research and education purposes. Following this discussion on abstracts and how we came to work with the British Library in developing the PhD Abstract corpora, we examine the facilities FLAX provides for searching, browsing, and viewing articles in the corpora. These facilities include highlighting automatically identified collocations, calling up auxiliary real-world information mined automatically from Wikipedia, and viewing lexical information obtained from wordlists. Next we look at ways of searching and browsing collocations and lexical bundles, and saving selected collocations for future reference. Then we examine ways of searching and browsing words and word families, and sentences grouped by word pattern. Finally, we draw some conclusions.

Our overall aim is to demonstrate that teachers and students can be exposed to salient lexical, grammatical, and pragmatic features of academic text by a system that automatically augments documents with facilities relevant to academic teaching practice. The work raises the obvious question of whether the system has real educational value, applied—as it is in this paper—to a standard corpus of academic written English, or indeed to any other suitable corpora. A full educational study is far beyond the scope of this paper. Nevertheless, we hope to convince readers that the scheme has great promise for helping language learners produce high-quality academic writing.

USING ACADEMIC TEXTS IN LANGUAGE LEARNING

Academic text is acknowledged to be of considerable value in language learning and teaching, and many pedagogical implications have arisen from studies of academic corpora. This section describes language features that are of particular importance in language acquisition and therefore are the focus of this project.

Corpora

Over the years, corpora have been developed for researchers and teachers to investigate linguistic features that are present in academic genres, and to help them identify problem areas in student academic writing. Some are built from highly graded university assignments, in a range of disciplines and across different genres—essays, reports, critiques, etc. The Michigan Corpus of Upper-Level Student Papers[1] contains 830 A-graded papers (2.6 million words), while the British Academic Written English corpus[2] contains 2860 assignments (6 million words); both collections span the broad disciplinary areas of Arts and Humanities, Social Sciences, Biological and Health Sciences, and Physical Sciences.

Documents written by language learners give insights into their writing development needs. The International Corpus of Learner English[3] contains 6000 EFL (English as a Foreign Language) texts (3.7 million words) written by advanced learners with diverse first languages—Chinese, Japanese, Italian, Spanish, French, German, Polish, etc. Other corpora are under development, such as the Cambridge Academic English Corpus (400 million words of spoken and written English) and the Corpus of Academic Learner English at Universität Bremen.

Students and teachers can interact with these corpora through accessing online user interfaces, using standard concordance tools, or by downloading entire collections. For example, the Michigan corpus provides online facilities for users to browse papers by student level, nativeness, textual feature (abstract, definitions, literature review etc.), discipline, and paper type (essay, proposal, report etc.); or to search for papers that contain a particular word or phrase.

Abstracts

Abstracts play a number of important roles in academic text. Identified primarily as a sub-genre (Swales & Feak, 2009) they have been characterized as the "gatekeepers" (Swales, 1990) of academic fields, and as "self-promotional tools" (Hyland, 2000) for authors to market and legitimize their writing within academic and professional communities. In addition to summarizing and distilling the content of the larger associated texts they point to, abstracts also enable efficient "scanning-reading strategies" (Lock, 1988) for readers who would otherwise be overburdened by having to keep up with *the hyper-production of knowledge in their fields* (Hyland, 2000, p. 64). Even though widely held as a sub-genre they possess "stand-alone mini-text" qualities (Hakin, 2001) with the growing consensus among academics that they may often be the only part of a paper read via abstracts databases. With reference to research articles, Hyland underscores the capacity of well-constructed abstracts to lend professional credibility to both the research and the writer of the research:

After the title the abstract is generally the readers' first encounter with a text, and is often the point at which they decide whether to continue reading and give the accompanying article further attention, or to ignore it. The research and the writer are therefore under close scrutiny in abstracts and, because of this, writers have carefully, and increasingly, tended to foreground their main claims and present themselves as competent community members. (Hyland, 2000, p. 63)

Abstracts also function as metadata (along with titles and keywords) for the improved searchability and ranking of a paper, thesis, and etcetera via search engines. More pointedly, the abstract is often the only part of a paper that is accessible within subscription-based publications (Bordet, 2014; 2015). This point of abstracts functioning as metadata, and therefore increasing their accessibility, is central to the development of the PhD abstract corpora in FLAX. Metadata, which currently includes the abstracts of 450,000 doctoral theses from UK universities, was harvested from EThOS[4] (managed by the British Library), which will be discussed in more detail in the following section of this chapter.

Words and Wordlists

Much effort has been spent on identifying language features in academic text. Coxhead (2000) developed a list of 570 academic words from a 3.5M word corpus of academic writing, which has become a widely used resource for teachers and students. A number of competing vocabulary lists have been created—for example the University Word List (Xue & Nation, 1984), the Academic Words List (Coxhead, 2000), the Academic Keyword List (Paquot, 2012) and the Academic Vocabulary List (Gardner & Davies, 2014). Computer tools such as the Vocabprofiler available at the Compleat Lexical Tutor website[5] help teachers and students analyze the vocabulary in a text with reference to this and other word lists.

Apart from academic words, the research also has focused on other pervasive words in academic writing for example, so-called shell nouns (e.g., *fact, approach, problem, result*), reporting verbs (e.g. *argue, suggest, define, claim*), and pronouns (e.g. *we*, *I*, *it*). Shell nouns are also known by various names: general nouns (Halliday & Hasan, 1976) anaphoric nouns (Francis, 1986), carrier nouns (Ivanič, 1991), enumerative nouns (Hinkel, 2004), signaling nouns (Flowerdew J., 2003) and stance nouns (Jiang & Hyland, 2015). They carry little or no meaning in themselves: instead, readers infer their meaning from the context. They glue text together by fulfilling functions like characterization (e.g. *the* ***problem*** *with this technique*), temporary concept-formation (e.g. *the same* ***result***), and linking (e.g. *this* ***fact***) (Aktas & Cortes, 2008). Reporting verbs are used when citing other people's work. Thompson and Ye (1991) categorized them in terms of denotation and evaluation, and further distinguished the perspective of a paper's author from that of someone who cites their work. For example, *state*, *point out*, *accept* and *challenge* are predominantly used to refer to an author's stance, while *anticipate*, *acknowledge*, *remark*, *utilize* generally refer to the present writer's interpretation. Hyland (1999) refined this categorization by dividing denotation into research acts (e.g. *observe, analyze, explore*), cognitive acts (e.g. *believe, conceptualize, suspect*), and discourse acts (e.g. *discuss, hypothesize, state*); he also grouped evaluation into factive (e.g. *acknowledge, point out, establish*), non-factive (e.g. *argue, address, suggest*), and counter-factive evaluation (e.g. *overlook, exaggerate, ignore*). Thomas and Hawes (1994) divided reporting verbs into finding verbs (e.g. *find, observe, show*), procedure verbs (e.g. *categorize, evaluate, examine*), tentativity verbs (e.g. *postulate, propose, suggest*), and certainty verbs (e.g. *state, report, document*). We also looked at the pronouns *we*, *I*, and *it*. The importance of first-person pronouns in academic prose and their underuse in student writing has been recognized by many researchers (Ädel, 2006; Harwood, 2005; Hyland, 2002). The inclusion of *it* offers students an alternative way of dealing with the writer's identity. Hyland and Tse (2005) note that the anticipatory *it*-clause fragment pattern (e.g. *It is interesting to note that*) highlights the writer's stance towards an argument but at the same time conceals his or her identity and reduces their responsibility for the claim.

Of course, learning vocabulary involves far more than simply memorizing words in lists or looking them up in dictionaries. Nation (2013) discusses many aspects of knowing a word, in terms of both receptive and productive capabilities. Academic writing requires specialized knowledge of language in different genres (Hyland & Tse, 2007), and teachers need to guide students through a range of contexts to examine how a word is used in a particular disciplinary area and prioritize words that are significant in that area. Table 1 shows the top 50 of Coxhead's academic words that appear in the Arts and Humanities and the Physical Sciences sections of the British Academic Written English (BAWE) corpus wherein only 18 words are found to be in common between the two sub corpora (bolded in the Table), for example, *theory*, *process*, *factor* (Wu & Witten, 2016). Whereas, when we examine the specificity of words, and their usage, in the sub corpora of the BAWE, we can see that, for example, *data*, *code, equation*, *error*, *output*, *ratio* are particularly important for students studying the physical sciences.

Collocations

The importance of collocation knowledge in academic writing has been widely recognized. Hill (1999) observes that students with good ideas often lose marks because they do not know the four or five most important collocations of a keyword that is central to what they are writing about. L2 student text tends to be cumbersome and error prone because of insufficient collocation knowledge. For example, in an

Table 1. Top 50 academic words in the BAWE Arts and Humanities and Physical Sciences

Arts and Humanities	Physical Sciences
theory, evidence, **create**, role, text, individual, **structure**, period, **process**, **area**, approach, context, economic, feature, culture, site, focus, concept, **image**, **analysis**, **factor**, conclusion, **method**, **function**, identify, **similar**, despite, aspect, issue, **occur**, contrast, revolution, identity, community, gender, cultural, **source**, establish, **involve**, **indicate**, interpretation, status, **element**, **require**, **affect**, link, notion, significant, style, physical	data, design, **process**, **method**, energy, error, equation, project, **function**, **require**, **analysis**, output, **structure**, ratio, code, **area**, section, **theory**, display, **factor**, obtain, **source**, **create**, computer, **element**, **occur**, define, constant, component, input, stress, range, **image**, team, phase, reaction, **involve**, variable, maximum, achieve, **similar**, **indicate**, **affect**, technology, potential, parameter, layer, technique, strategy, financial

essay on "Smoking" one student wrote *the smokers who rely on cigarettes and have to smoking everyday* instead of using phrases such as *heavy smoker* or *addicted to smoking*.

Topic-specific corpora are valuable resources that help students build up collocation knowledge within the areas that concern them. When investigating semantic associations in a corpus of business English, Nelson (2006) discovered that the lexical environment surrounding business words is rich, diverse and semantically connected. He noted that collocates of certain common words are more formulaic than in general English. It is also true that the same word in different disciplines exhibits a different range of collocations. For example, Wu and Witten (2016) compared the usage of the word *image* (and inflected forms) in the Arts and Humanities and the Physical Sciences areas in the BAWE corpus that contain similar numbers of student texts and similar numbers of occurrences of the word *image*. Table 2 demonstrates variation in sense, grammatical patterns, and associated collocates. The dominant sense of *image* is a mental representation in the Arts and Humanities and a physical representation in the Physical Sciences. The word is commonly used in different grammatical patterns and associated with different sets of verb and adjective collocates, which reflect these two shades of meaning.

Lexical Bundles

To become proficient writers, students need to develop a repertoire of discipline-specific phrases. Recently, Biber and colleagues developed the notion of "lexical bundles," which are multi-word sequences with distinctive syntactic patterns and discourse functions that are commonly used in academic prose (Biber & Barbieri, 2007; Biber, Conrad & Cortes, 2003, 2004; Biber & Conrad, 1999; Biber & Johans-

Table 2. The word image in Arts and Humanities, and the Physical Sciences in BAWE

	Arts and Humanities	Physical Sciences
sense	mental representation of a thing	physical representation of a thing
occurrences	694	668
grammatical pattern	verb + *image* + *of*, *the image of*	verb + *image*, *The image is* + verb
verb collocates	*restore, contain, offer, challenge, create, reflect, abstract, have, conjure, mystify, leave, perpetuate, construct, conflict, give*	*assemble, convey, produce, combine, display, choose, convert, analyze, rotate, flip*
adjectival collocates	*final, prevailing, subjective, weak, strong, public, intense, mirror, typical, horrific, poetic, brutal, peculiar, strange, negative, mental, unique, romantic, bleak, eternal*	*particular, original, similar, larger, clear, physical, virtual, certain, digital, continuous, large, intellectual*

son etc., 1999). Typical patterns include noun phrase + *of* (*the end of the*, *the base of the*), prepositional phrase + *of* (*as a result of, as a part of*), *it* + verb/adjective phrase (*it is possible to, it is necessary to*), *be* + noun/adjective phrase (*is one of the*, *is due to the*), and verb phrase + *that* (*can be see that, studies have shown that*). Such phrases fulfill discourse functions such as referential expression (framing, quantifying and place/time/text-deictic), stance indicators (epistemic, directive, ability) and discourse organization (topic introduction/elaboration, inferential and identification). Hyland's (2008b) follow-up study compared the most frequent 50 four-word bundles in texts on biology, electrical engineering, applied linguistics and business studies, and discovered substantial variation between the areas. These findings point to the need for students to understand relevant discourse features in their subject domains.

Some studies compare student and expert writing in terms of the lexical bundles they exhibit. Cortes (2004) identified a set of target bundles from history and biology publications and compared them with ones extracted from student writing at different university levels. He concluded that students rarely used the target bundles correctly. Hyland (2008a) discovered different structural and functional cluster patterns in three genres—research articles, master's theses, and doctoral dissertations. Expert writing contains fewer clusters, mostly taking the form noun phrase + *of* used to organize text, in contrast to the plethora of different clusters used for structuring activities and experiences as exhibited in student writing. When comparing lexical bundles in L1 and L2 academic writing Chen and Baker (2010) found a similar trend: more verb phrase based bundles and discourse organizers occur in student writing, whereas bundles in expert writing are mainly noun phrase based referential markers.

THE PHD ABSTRACT CORPORA

Throughout this paper we use the PhD abstract corpora as an example of an extended body of academic text. These abstracts come from the Electronic Theses Online Service (EThOS) managed by the British Library. The entire corpus was downloaded through OAI-PMH[6] harvesting protocol. Each PhD abstract file contains the text of the abstract, and the metadata such as the title, creator, date, subjects, and publisher. We used the subject field containing Dewey Decimal Classification code to divide abstracts into the areas of Arts and Humanities, Social Sciences, Life Sciences and Physical Sciences, and for each area, further into disciplines and subjects. Table 3 shows the number of abstracts, the number of running words, the average length of an abstract, and disciplines in each area. We built a digital library collection (a set of electronic documents) for each of the four areas.[7] Our work is an extension of the Greenstone digital library system,[8] which is widely used open source software that enables end users to build large collections of documents and metadata that are searchable and browseable, and to serve them on the Web (Witten, I.H., Bainbridge, D. & Nichols, D.M., 2010).

To provide a brief description of the harvesting process, the digital PhD abstracts were automatically analyzed, enriched, and transformed into a resource that second-language and novice research writers can browse and query. This automated text and data mining work using the EThOS toolkit[9] was made possible due to changes in legislation introduced in 2014 for the text and data mining of open access publications for non-commercial purposes (Fitzgerald & Wu, forthcoming). It is anticipated that the practical contribution of the FLAX tools and the PhD abstract corpora will benefit second-language and novice research writers in understanding the rhetorical moves and language used to achieve the persuasive and promotional aspects of the written research abstract genre. It is also anticipated that users of the collections will be able to develop their arguments more fluently and precisely through the practice

of research abstract writing to project a persuasive voice as is required in specific research disciplines (Hyland & Tse, 2005a; Hyland & Tse, 2005b; Bondi & Lores Sanz, 2014).

Some useful research has been conducted into the writing of abstracts with particular emphasis on rhetorical moves (Bhatia, 1993; Hyland, 2000; Bordet, 2015), and how features of lexico-grammar support the different rhetorical moves present in abstracts. For example, Bordet's 4-move rhetorical classification system [Context, Research statement, Method, Results] is combined with identifiable features of lexico-grammar to guide readers by way of "lexical paving" through the argumentation of a text:

…a succession of lexical patterns' variations around reiterated pivot keywords within a text forms a sort of "lexical paving" whose integration with the rhetorical moves contributes to the coherence of the argumentation in a text, as expected by a specified discourse community. (Bordet, 2015, p. 45)

Where abstract corpora have been deployed in research, however, they have been limited to only a few disciplines and are often inaccessible for querying purposes by researchers as well as EAP teachers and learners due to copyright restrictions. Concerning the use of corpus-based language teaching materials in language instruction, as mentioned earlier in the introduction section of this chapter, Tim Johns is often regarded as the pioneer in the field, coining the term Data-Driven Learning (DDL) to refer to the method of inferring the rules of language by directly observing them in corpora using text analysis tools. Johns envisioned every language learner as "a Sherlock Holmes" with direct access to the evidence of real-world language data (Johns, 2002, p. 108). And, like contemporary advocates for using and developing data literacies with open data in higher education (Atenas, Havemann, & Priego, 2015), Johns also envisioned DDL as developing data literacies for understanding and interpreting linguistic data for direct applications in language learning (Johns, 2002). In response to the lack of accessible abstract corpora, which reflect variation and change in academic discourse from across the disciplines, a guiding research question leads our project development work with the PhD abstract corpora:

To what extent can openly accessible PhD abstract corpora, which are inclusive of all subject domains, enrich data-driven learning for second-language and novice research writers?

Table 3. Number of abstracts and disciplines in each area

Area	Abstracts	Running Words	Ave. Words per Abstract	Discipline
Physical Sciences	7825	2,695,500	345	Architecture, Astronomy, Chemistry, Computer science, Earth sciences and geology, Engineering, Manufacturing, Mathematics, etc.
Social Sciences	8769	3,117,800	356	Commerce, communications, and transportation, Economics, Education, Law, Library and information sciences, Management and public relations, Political science, sociology and anthropology, etc.
Life Sciences	6251	2,233,400	357	Agriculture, Animals (Zoology), Biology, Fossils and prehistoric life, Medicine and health, Plants, etc.
Arts and Humanities	5525	1,827,170	331	Arts, History, Linguistics, Language, Music, Philosophy, Psychology, Religion etc.

One of the driving principles of the open education and open access movements in higher education and research is to ensure *enhanced availability of discoverable, reusable and repurposable academic open content.* (JISC, 2011). EThOS is the United Kingdom's national PhD thesis service, which aims to maximize the findability and accessibility of doctoral research theses. EThOS supports the UK Government's open access mandate that publications resulting from publicly funded research are made freely available for all researchers, providing opportunities for further research. Following suit, most universities now require PhD graduates to deposit their thesis in a local institutional repository. Services such as EThOS further facilitate access to this material by way of providing an umbrella service for managing as well as archiving PhD theses from UK universities, including the oldest British thesis on record dated 1738 by Thomas Charles Hope from the University of Edinburgh (who was, incidentally, a teacher of the young Charles Darwin).

In addition to EThOS, the Digital Research Team at the British Library has developed British Library Labs[10], an Andrew Mellon Foundation funded initiative, which encourages and supports scholars and inspires the reuse of the British Library's digital collections and data in exciting and innovative ways. In 2016, FLAX was awarded a British Library Labs award in the Teaching and Learning category for the reuse of digital collections in language education. Prior to this in 2015, by way of a pilot study, we explored EThOS at the British Library with language teachers at Queen Mary University of London for creating micro-corpora of approximately 30 abstracts each that corresponded with subject areas for the summer pre-sessional EAP programs at Queen Mary Language Centre. It was discovered that working with abstracts saved time for busy teachers in building micro collections with the software directly onto the FLAX website. Abstracts also proved to be ideal in terms of size for developing web-based and mobile language learning activities using the suite of mobile applications for Android from FLAX[11] and thus avoiding issues with large text scrolling on mobile devices. For actual examples, please refer to the micro PhD abstract corpora in the areas of Law[12], and Water Politics and Tourism Studies[13], for the types of web-based and mobile activities developed by EAP teachers at Queen Mary University of London.

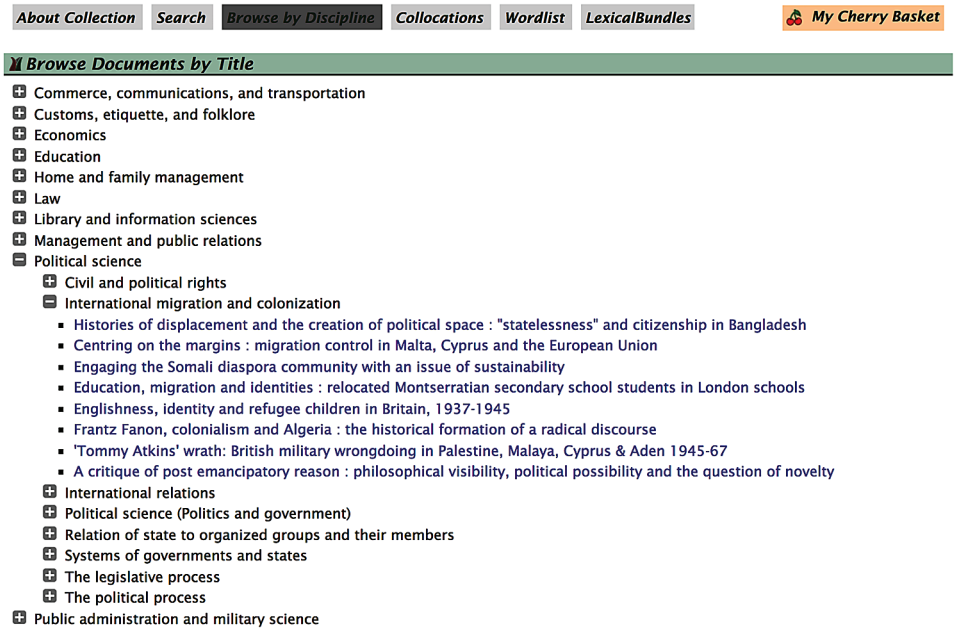
The following sections of this chapter will describe specific functions in the PhD abstract corpora for their uptake in EAP.

NAVIGATING THE PHD ABSTRACT CORPORA

Figure 1 shows the main page of the Social Science collection. The buttons at the top of the page indicate the following:

- *About Collection* introduces the content of the collection;
- *Search* enables you to search the collection at the level of abstracts, paragraphs, sentences or collocations;
- *Browse by Discipline* lets you browse the documents by title;
- *Collocations* allows you to study the collocations present in the abstract corpora, including the *Top 100* collocations and their lexico-grammatical patterns;
- *Wordlist* presents the words in the collection, sorted by how often they occur;
- *Lexical Bundles* lists four-word sequences with distinctive syntactic patterns and discourse functions;
- *My Cherry Basket* shows you any collocations ("cherries") that you have collected.

Figure 1. Main page of the Social Science collection



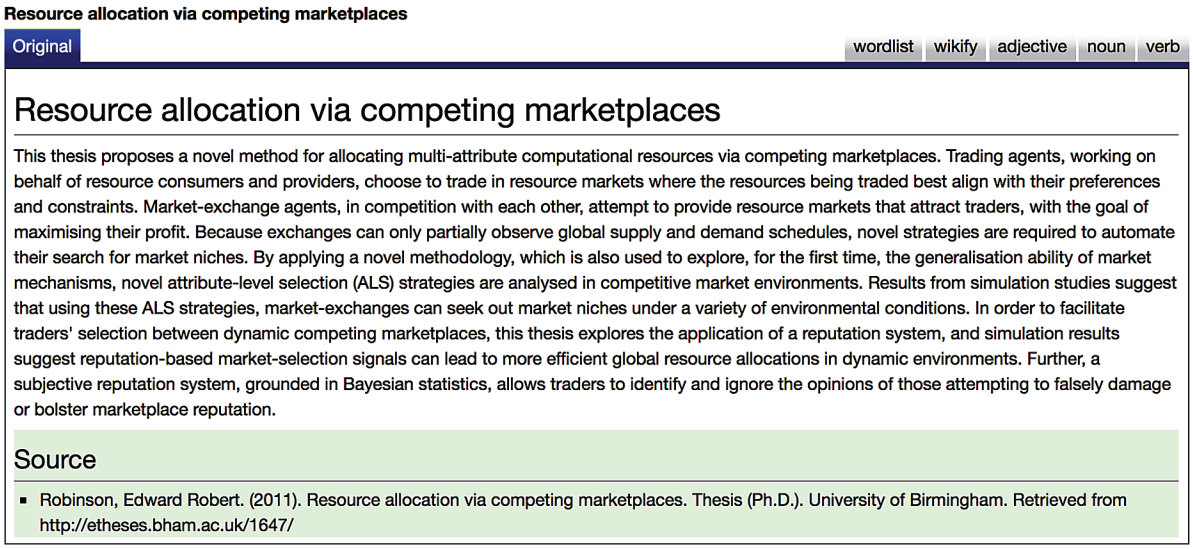
In Figure 1, the *Browse by Discipline*, *Political Science, International Migration and Colonization* buttons have been clicked in that order, which brings up the titles of all articles classified as *International Migration and Colonization*. Users can select a different discipline (the *Browse by Discipline* button), or search for articles containing a particular word (the *Search* button). These are standard Greenstone facilities.

EXPLORING DOCUMENTS

Clicking on an abstract title brings up the full text of the abstract itself, in a view like that shown in Figure 2. This particular abstract is titled, "Resource allocation via competing marketplaces" and can be accessed by selecting *Browse by Discipline* and then *Commerce, communications, and transportation*, and *Commerce (Trade).* As well as viewing the original document, other views can be obtained by clicking the buttons along the top of Figure 2:

1. Wordlist view, which allows users to study the vocabulary used in the document by clicking on the *wordlist* button;
2. Wikipedia view, which helps users grasp the meaning of concepts that are mentioned in the document that have been mined with the Wikipedia toolkit by clicking on the *wikify* button;
3. Collocation view, which allows users to examine lexical compounds that occur in the document, divided into collocations that involve adjectives, nouns, and verbs by clicking on the corresponding *adjective*, *noun* and *verb* buttons.

Figure 2. An abstract in the Social Science collection



These functions are described in more detail in the following subsections of this chapter.

Wordlist View

The Wordlist view, which is shown in Figure 3, allows users to analyze the range of vocabulary used in the article. The drop-down box above the text (currently showing "academic words") provides five options: the most frequent 1000 and 2000 words, taken from wordlists used in language teaching (West, 1953); academic words included in the list by Coxhead (2000); other off-list words, which are often specific to the content of the document; and keywords.

Figure 3 shows "academic words", and their distribution in the document is indicated beside the dropdown box (21%). Clicking a highlighted word leads to a page that shows all the sentences in the collection containing that word.

Figure 4 shows the keyword view, in which the words *trading, trader, market, resources, marketplaces,* and so on, are highlighted in blue. The bar beside the dropdown box is a slider that the user can manipulate by dragging right or left to reveal more words: moving it to the right makes the system less selective, highlighting more words; conversely, moving it to the left reduces the number of highlighted words. At the very left end only one keyword, *resource,* is given, while at the right end of the slider all content words are displayed.

Keywords are calculated by a heuristic method commonly deployed in information retrieval (known as Term Frequency–Inverse Document Frequency or TF-IDF, and described by, for example, Witten et al., 1999). First, documents are parsed, and the nouns, adjectives, verbs, and adverbs are designated as content words. For each such word, a score is calculated that reflects how important the word is to the document, based on the number of times it occurs in the document (which increases the score) and the number of times it occurs in the collection as a whole (which decreases it).

Figure 3. Academic words highlighted in the Wordlist view

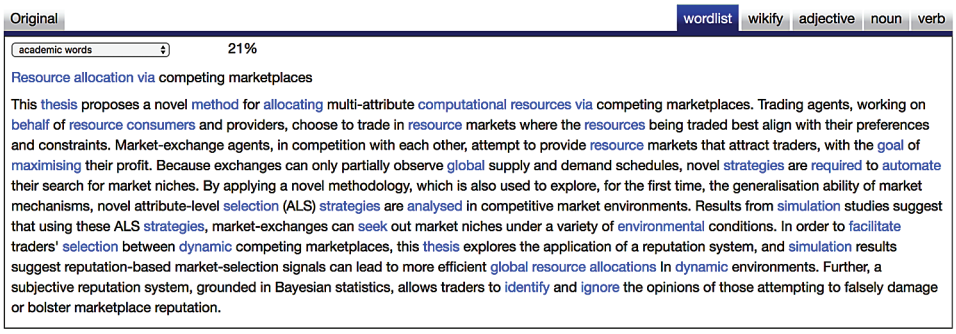
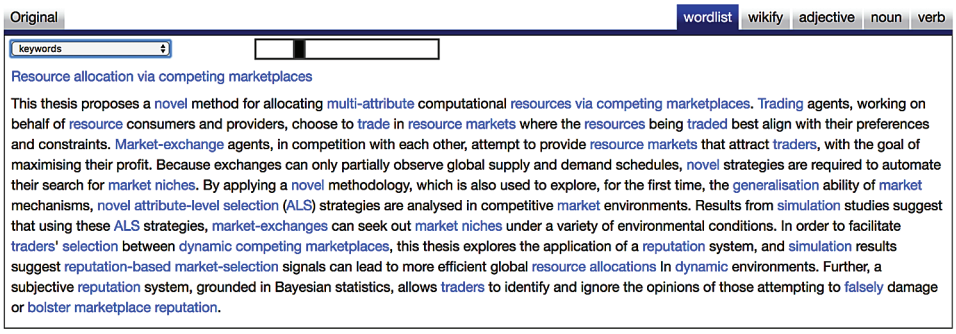


Figure 4. Keywords highlighted in the Keywords view

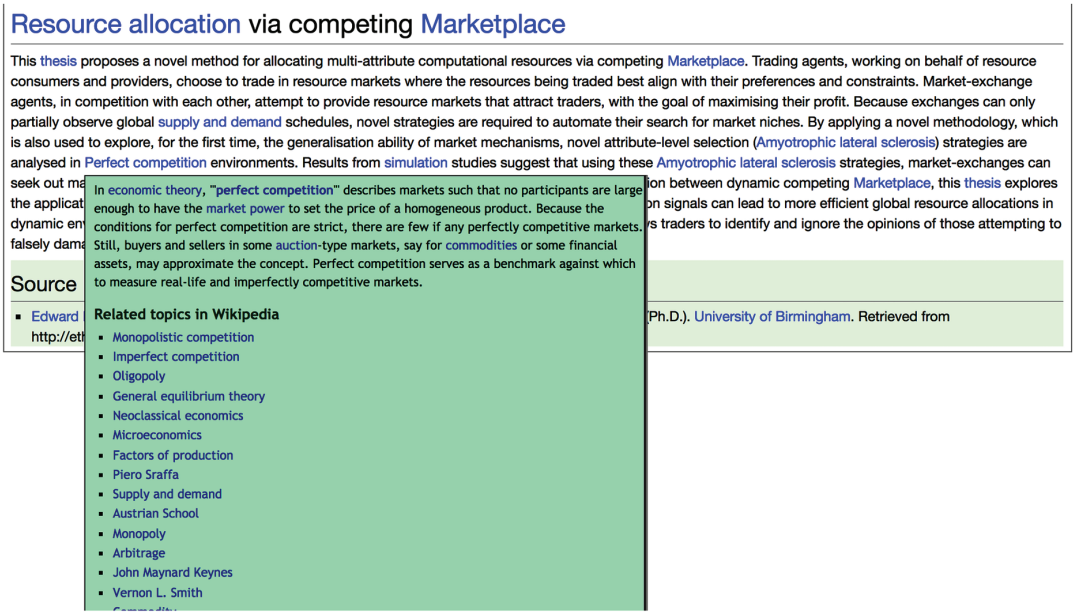


Wikipedia View

The Wikipedia view, illustrated in Figure 5, is obtained by clicking the button labeled *wikify*. It links the terminology used in the abstract to Wikipedia, highlighting concepts that are defined therein. In Figure 5, the phrases *resource allocation, marketplace, supply and demand, perfect competition*, *Amyotrophic lateral sclerosis*, *reputation system*, etc., have been identified as pertinent concepts. These could be Wikipedia article titles, or "redirects" (i.e., synonyms) defined in Wikipedia itself. Clicking on any highlighted phrase in the document brings up information extracted from Wikipedia in a popup window: the definition, hyperlinked to the Wikipedia article; followed by a list of related topics in Wikipedia that can also be clicked and further explored. In Figure 5, *perfect competition* has been selected, and in this case the list of related articles gives the names of prominent economists in this area. Their full Wikipedia entries are only one click away.

The process of relating words and phrases with running text to Wikipedia articles is called "wikification," and Milne and Witten (2013) describe the method we use. It has three steps. First, sequences of words in the text that may correspond with Wikipedia articles are identified using the names of the articles, as well as their redirects and every referring anchor text used anywhere in Wikipedia. Second, situations where multiple articles correspond to a single word or phrase are disambiguated. For example, the word *kiwi* may refer to a bird, a fruit, a person from New Zealand, or the New Zealand national rugby league team, all of which have distinct Wikipedia entries. A machine learning classifier is used to make

Figure 5. Wikipedia articles are highlighted in the Wikipedia view



the appropriate choice, taking into account the prior probability of the mapping, semantic relatedness to other concepts in the same document, and some contextual information. The third step selects the most salient linked (and disambiguated) concepts to include in the output. Again, a machine learning approach is used that combines prior probability, relatedness to context, disambiguation confidence, generality, location and spread. This process is described in greater detail by Milne and Witten (2013), including the machine learning methods and models used.

Collocation View

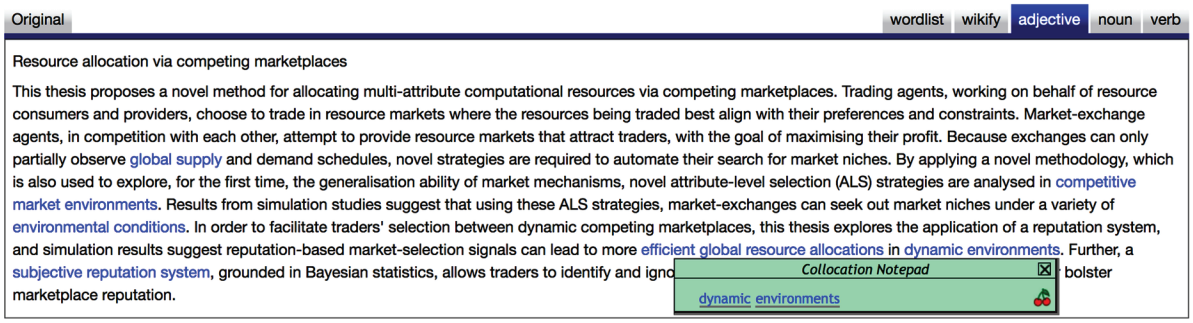
Benson et al. (1986, p.ix) define collocations as follows:

In any language, certain words combine with certain other words or grammatical constructions. These recurrent, semi-fixed combinations, or collocations, can be divided into two groups: grammatical collocations and lexical collocations.

We focus on lexical collocations, which have structures verb + noun, adjective + noun, noun + noun, noun + *of* + noun, and preposition + noun.

Once an article has been selected, the collocation view is accessed by clicking one of the *adjective*, *noun* or *verb* tabs as shown in Figures 2–4. Each tab shows collocations starting with that word type; for example, the *adjective* panel hosts collocations starting with an adjective. Collocations are highlighted in the text, to help students notice them and study their context. In the example shown in Figure 6, collocations related to the subject of the article, marketplaces—*global supply*, *competitive market environment, environmental conditions, subjective reputation system*—stand out from the surrounding text, attracting the student's attention. The collocation *dynamic environments* has been clicked to reveal the *Collocation*

Figure 6. Collocations, highlighted in the collocation view



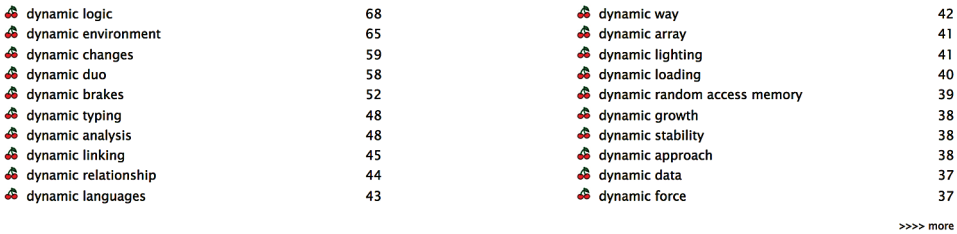
Notepad pop-up window in green. The underlined words, *dynamic* and *environments*, are hyperlinked to entries that contextualize the same words in an external collocations database,[14] built from all of the written text in Wikipedia (three million Wikipedia articles comprising three billion words).

The FLAX system makes it easy for users to study further collocations related to these two words through the much larger collocations database in FLAX . For example, clicking *dynamic* in Figure 6 generates a further popup, shown in Figure 7, that lists *dynamic logic, dynamic environment, dynamic changes, dynamic duo, dynamic brakes* etc., along with their frequency in that corpus; likewise, clicking on *environments* in Figure 6 reveals more collocations in the database such as *hostile environments, certain environments, various environments, arid environments, social environment, integrated development environment*, and so on. Furthermore, users can see samples of these collocations in context by clicking on the links to reveal relevant extracts from the Wikipedia articles mined in the Learning Collocations database. Wu et al. (2016) describe the full functionality of the large Wikipedia collocations database in FLAX.

EXPLORING COLLOCATIONS AND LEXICAL BUNDLES

We have described the various ways a user can view individual articles in the Social Sciences collection, including showing the collocations that it contains. This section describes other facilities for exploring the language used in the collection, outside the context of a particular abstract in the sub corpus. Users can investigate collocations by searching or browsing. They can save and organize their favorite collocations for future use when writing. Finally, they can study the lexical bundles in the collection.

Figure 7. Related collocations for the word dynamic



As well as the standard search function of locating articles, paragraphs, or sentences that contain a particular word or words, the "Search" button at the top of the user interface allows users to search collocations in the collection. Figure 8 shows results for the search term; *impact*, which has returned 1763 collocations of which the first four can be seen here, along with their contexts. They are grouped under tabs that reflect the syntactic roles of the associated word or words, Noun + of (shown), Verb, Adjective, and Noun. The dominant pattern for our chosen word *is noun + of + impact* or *impact of + noun* (half the total of 1763), as in *analysis of the impact*, *understanding of the impact*, *impact of globalization*. The next most dominant pattern is verb + *impact* + *of* (nearly a fourth of the total), for example: *examines the impact of*, *assess the impact of, explore the impact of* etc.

Collocations can be browsed as well as searched, using the *Collocations* button. Then, an alphabetic selector leads to the word in question. Clicking the letter *i*, followed by the word *impact*, obtains the collocations shown in Figure 8.

Cherry Picking

Many educators encourage language learners to collect their favorite collocations for possible use in later writing. In FLAX, this can be done using the "cherry-picking" interface shown in Figure 8. We use the metaphor of cherries because, like collocations, they are tasty fruit that come in small clusters. Clicking on the "cherries" icon that follows collocations as shown in Figures 6, 7 and 8, launches the interface. In Figure 9, the collocation *global supply* has been selected to add to the user's personal cherry basket. If desired, it can be assigned to a category or categories, or a new category can be created for it.

Figure 10 shows a student's cherry basket that displays collocations that have been picked and placed in two categories: *commerce* and *economics*. The usual options are provided for organizing the basket: collocations can be deleted or moved into different categories, new categories can be created and old ones deleted, and the contents of the basket can be printed or emailed to oneself or to learning peers (see the *Print friendly* button in Figure 10).

Figure 8. Exploring collocations associated with the word impact

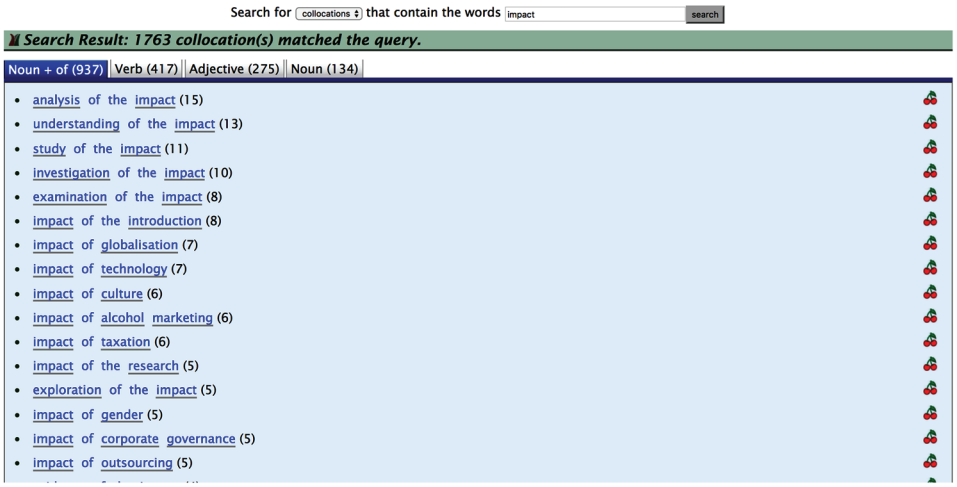


Figure 9. Collocation "cherry picking" interface

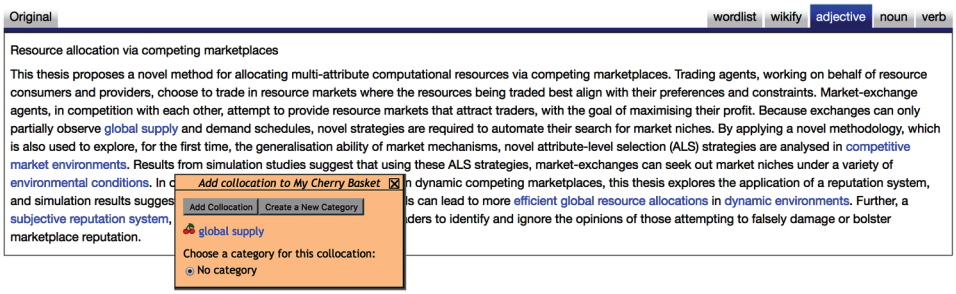


Figure 10. A personalized cherry basket

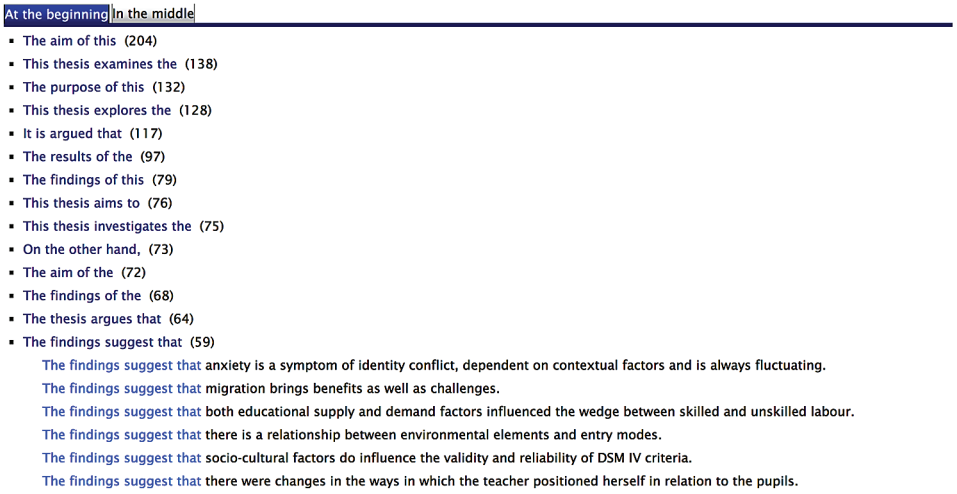
Browsing Lexical Bundles

"Lexical bundles," namely multi-word sequences with distinctive syntactic patterns and discourse functions that are commonly used in academic prose, have recently become prominent in research on academic language learning. To help users explore them, our software extracts all four-word phrases that appear in the collection and sorts them by frequency. We chose four-word sequences because it seems that most discourse function bundles are four-word combinations—at any rate, this is the length that appears most often in the literature.

Bundles at the beginning (we call them "head bundles") and in the middle ("middle bundles") of sentences are treated separately. Researchers generally discard infrequent bundles: we decided to use document counts instead of occurrence counts, and (after experimenting with various thresholds) discarded bundles that appear in fewer than three documents. This yields about 800 head bundles and 3500 middle bundles for each of the PhD abstract sub corpora.

Clicking the *Lexical Bundles* button as shown in Figure 11 displays the lexical bundles extracted from the collection. Head bundles and middle bundles appear under separate tabs, sorted by frequency. Figure 11 shows the top head bundles—*The aim of this*, *This thesis examines the*, *The Purpose of this*, *This thesis explores*, *It is argued that*, *The results of this*—along with their frequencies. Clicking a bundle—here, *The findings suggests that*—expands it to show the corresponding sentences.

Figure 11. Lexical bundles at the beginning of sentences

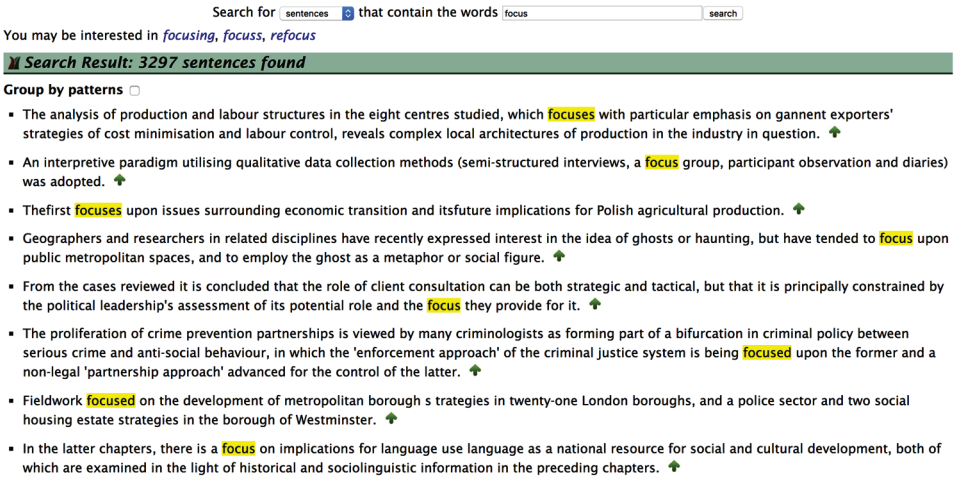


EXPLORING WORDS

Searching documents for particular words and phrases is one of the standard functions of digital libraries. However, library users generally seek information about the content of articles, whereas language learners want to know how words are used. The same search mechanism can be applied, but the results should be displayed in different ways.

Users can search documents, paragraphs, sentences and collocations containing a particular word, along with its variants. Here we focus on sentence and paragraph searching. Figure 12 shows the first 8 of 3297 sentences that contain the word *focus*; sentences containing the inflected forms *focuses* and *focused* are also returned by this search. To recognize inflected forms of a query word, a lemma list

Figure 12. Result of searching for sentences containing the word focus



containing about 15,000 entries is consulted.[15] Clicking the "arrow" icon at the end of a sentence pops up the paragraph that contains the sentence, to show its context.

Other terms or alternatives derived from the query word, in this case the word *focusing*, *focus*, *refocus* are given at the top of Figure 12 as further possible searches. Although *focus* has only three derived terms, there are often several: for example, the query word *analysis* yields the derivatives *analyze, analytical, analyze, analytic, analyst*, and *analytically*, which are presented in descending order of frequency in the collection.

Search queries can contain more than one word, in which case sentences are returned that contain all the query terms. For phrase searching, a query can be enclosed by quotation marks; for example, *"focus group"* returns sentences containing this phrase, while *focus group* returns sentences that contain these two words.

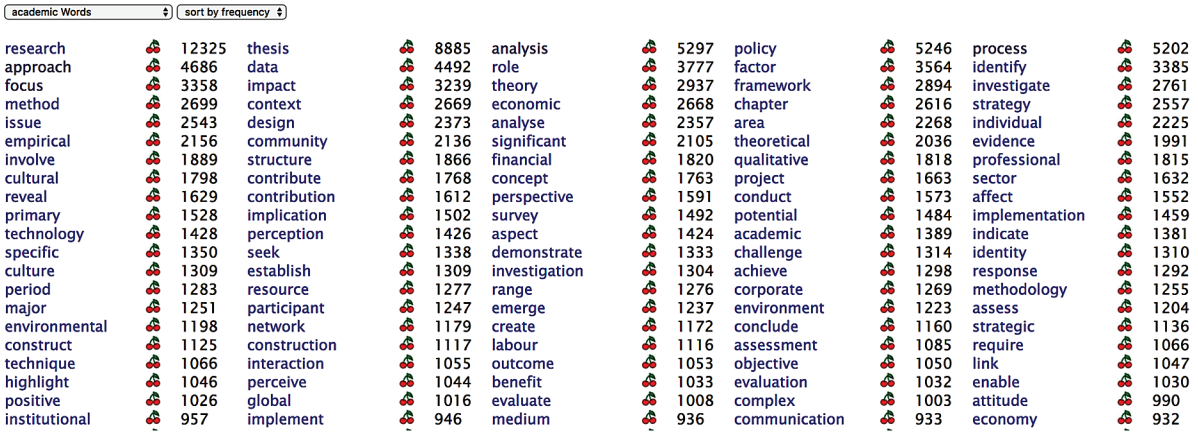
Wordlists

Users can explore academic words in the collection by clicking the *Wordlist* button at the top of the user interface, yielding the screen shown in Figure 13. The words can be sorted alphabetically or by frequency, as seen in Figure 13; in either case, frequency in the collection is shown alongside the word. Clicking the word itself retrieves sentences containing it, on a page like that of Figure 12. The "cherries" icon links to the collocations associated with the word, yielding the same sort of display as seen previously in Figure 8. Note that in each of these cases all inflected forms of the word are also included in the search.

FLAX also provides a separate interface for students to access a list of shell nouns, reporting verbs, and pronouns by way of organizing useful words for academic writing. The list comprises:

- 160 reporting verbs, e.g. *question, argue, accept*;
- 36 shell nouns, e.g. *form, fact, result*;
- 3 pronouns: *We, I, It*;

Figure 13. The academic words in the collection



- 240 adjectives, e.g. *positive, visual, important*;
- 123 adverbs, e.g. *particularly, initially, significantly*.

Figure 14 shows how these words are presented. Reporting verbs are grouped by function, such as agreement (e.g. *admits, accepts, supports*), argument and persuasion (e.g. *assures, justifies, emphasizes*), evaluation, and examination (e.g. *analyses, compares, investigates*). The shell nouns are from Aktas and Cortes's list (2008). We added 240 adjectives and 123 adverbs that occur frequently in the PhD abstract corpora.

All words in the list are sorted by frequency as they are distributed across the PhD abstract corpora, which are then divided into four sub corpora. Frequencies from the Social Sciences corpora are shown by default, but students can switch to another sub corpus (e.g. Arts and Humanities) by selecting it from the drop-down list at the top. Again, clicking a word displays usage patterns in the Social Sciences sub corpus in the same form as shown previously in Figure 12.

Grouping Words by Pattern

The "Group by pattern" option, shown near the top of Figure 12 but turned off by default, allows users to study word usage by showing salient lexico-grammatical patterns. We group patterns by word position—near the beginning or in the middle—because these provide different views of a word's usage patterns. Figure 15 shows (in the bar at the top) that 1296 patterns are found for the word *focus*. They are separated into two tabs, *At the beginning* (758 patterns) and *In the middle* (538 patterns). The word *focus* can be used as a verb or a noun, and the most common pattern has the form *The study focuses on*

Figure 14. Useful words for academic writing

(Figure 15), which occurs—with different subjects—in many sentences (418). Clicking on the pattern brings up examples from the corpus; they include *This thesis focuses on*, *This research focuses on*, and *The analysis focuses mainly on* and all exhibit the same pattern: noun + *focuses* + *on*-phrase. The next most common patterns in Figure 15, *The focus of*, *We focus on, Most research has focused on, The focus is on* and so on, demonstrate that the most dominant usage of the verb *focus* near the beginning of sentences is subject + *focus* + *on*-phrase, and the noun *focus* are *focus* + *of*-phrase + be and *focus* + *be* + *on* respectively.

How *focus* is used in the middle of sentences shows the same trend whereby *focus* is used as a verb or a noun. The *focus* + *on*-phrase is the most common pattern for the verb *focus*, while the noun form of *focus* is followed by various prepositional phrases such as *focus* + *of*-phrase (*the focus of this work*, *the focus of my investigation*), *focus* + *on*-phrase *(a strong focus on the routine, a dual focus on the formative experiences),* and *focus* + *for*-phrase (*a focus for the development, a suitable focus for review*).

The *Group by pattern* interface also makes it easy for users to identify collocations. For example, *The factor that/which*, and noun + verb + *factor* + *that*, and verb + *factor* + *that* and verb + *factor* + *for/in/of* are the top two usage patterns of the word *factor* at the beginning and in the middle of sentences correspondingly. Expanding the verb + *factor* + *that/which* pattern generates these verb and adjective collocates of *factor,* for example:

be the key/specific/additional factors which
factors which build/ shape/determine/determine
identify/investigate/examine/outline/assess/operationalize the factors that
various/social/contextual/critical/underpinning factors that

It is also possible to retrieve patterns for function words like pronoun *it*, preposition *in*, adverb *however*. Table 4 gives the top ten patterns associated with *it* at the beginning of a sentence.

Wu and Witten (2016) have discussed the pattern generation procedure and evaluation in a separate paper. Here a summary is given. Texts stored in FLAX are split into sentences, part-of-speech tagged and chunked into syntactic phrases (noun, verb, preposition, etc.). Users present a query term and FLAX retrieves chunked sentences containing the query term and extracts the phrases surrounding it, arranges them by syntactic pattern, and sorts them by frequency. The procedure was evaluated statistically with

Figure 15. Sentences containing focus at the beginning position, grouped by pattern

Table 4. The ten most common patterns of it at the beginning of sentences

Pattern	Sample
It + be + verb + that	It is argued that / It was found that / It is suggested that
It + verb + noun + of	It provides an understanding of / It considers the impact of
It + verb + that	It argues that / It shows that / It concludes that
It + verb + to + verb	It seeks to understand / It aims to contribute / It attempts to identify
It + verb + that	It seems that / It appears that
It + verb + how	It examines how / It investigates how / It reveals how
It + be + adjective + to + verb	It is possible to identify / It is important to consider
It + be + adjective + that	It was evident that / It is clear that / It is vital that
It + verb + noun + which/that	It adopts an approach which / It identifies the channels that
It + verb + noun + in	It fills a gap in / It suggests ways in / It has a role in

respect to Coxhead's Academic Word List and West's (1953) General Service List. Assuming that the idea of syntactic grouping makes intuitive sense, the procedure's success can be assessed by measuring the ratio of sentences covered by a pattern to the total number of sentences containing that term. For example, if *provide* + noun + *of* covers 30% of sentences containing the term *provide*, it is fair to conclude that this is one of the typical usages for *provide*. The results show that for any given word, the two most frequent beginning and middle patterns cover 41% and 22% respectively of sentences containing that word, increasing to 59% and 36% for the top five patterns. This indicates that language learners can benefit from focusing on these patterns when studying the lexico-grammatical patterns of a particular word (Wu & Witten, 2016).

CONCLUSION

We have shown how academic text can be augmented to facilitate language learning. The process embodied in the FLAX system is guided by findings recorded in the research literature. It extracts useful language learning material from the input documents, including academic words, key words and concepts; collocations; typical word usage patterns; and lexical bundles. All these are made easily accessible through a unified searching and browsing interface.

Document text is presented in different views, each focusing on a particular linguistic aspect, with the aim of drawing users' attention to different aspects of language, and increasing their awareness of how ideas are expressed. To further enrich and expand student knowledge, external resources—Wikipedia—are automatically linked into the collection to provide additional context, both pragmatic and linguistic. Users can get a better feeling for the facilities the FLAX system provides by exploring its use online.

It should be emphasized again that although our description has focused on a particular corpus, the PhD abstract corpora, for illustrative purposes, the process is entirely automatic. Using the open-source FLAX software, anyone can assemble a collection of documents and have it built into a digital library with all the facilities described here. For example, other standard corpora can be used, or a set of documents

chosen by a teacher, perhaps in a particular discipline, or even samples of student writing. However, if the documents are not academic ones the facilities—particularly those involving lexical bundles—may not be as useful as illustrated here.

Can this system actually improve language learning in a practical setting? The answer will depend on how it is used and how it is linked in with other aspects of formal language teaching, and in self-directed informal language learning. While we hope that the examples given here make a compelling case for its potential utility, user studies will certainly be required to prove the point—and to suggest further directions for development. The idea of automatically enhancing text for the purposes of language learning is just beginning to be explored.

REFERENCES

Ädel, A. (2006). *Metadiscourse in L1 and L2 English*. Amsterdam: John Benjamins Publishing Company. doi:10.1075/scl.24

Aktas, R. N., & Cortes, V. (2008). Shell nouns as cohesive devices in published and ESL student writing. *Journal of English for Academic Purposes*, *7*(1), 3–14. doi:10.1016/j.jeap.2008.02.002

Atenas, J., Havemann, L., & Priego, E. (2015). Open data as open educational resources: Towards transversal skills and global citizenship. *Open Praxis*, *7*(4), 377–389. doi:10.5944/openpraxis.7.4.233

Benson, M., Benson, E., & Ilsen, R. F. (1986). *The BBI combinatory dictionary of English: A guide to word combinations*. Amsterdam: John Benjamins. doi:10.1075/z.bbi1(1st)

Bernardini, S. (2002). Exploring new directions for discovery learning. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis* (pp. 165–182). Amsterdam: Rodopi. doi:10.1163/9789004334236_015

Bhatia, V. K. (1993). *Analysing genre: Language use in professional settings*. London: Longman.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, *26*(3), 263–286. doi:10.1016/j.esp.2006.08.003

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

Biber, D., Conrad, S., & Cortes, V. (2003). Lexical bundles in speech and writing: an initial taxonomy. In A. Wilson, P. Rayson, & T. McEnery (Eds.), Corpus linguistics by the lune: A festschrift for Geoffrey Leech (pp. 71–92). Frankfurt/Main: Peter Lang.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at . . .: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405. doi:10.1093/applin/25.3.371

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman grammar of spoken and written English*. London: Longman.

Bondi, M., & Lores Sanz, R. (2014). Introduction. In M. Bondi & R. Lores Sanz (Eds.), *Abstracts in Academic Discourse: Variation and Change* (pp. 9–20). Bern: Peter Lang. doi:10.3726/978-3-0351-0701-2

Bordet, G. (2014). Influence of collocational variations on making the PhD abstract an effective "would-be insider" self-promotional tool. In M. Bondi & R. Lores Sanz (Eds.), *Abstracts in Academic Discourse: Variation and Change* (pp. 131–160). Bern: Peter Lang.

Bordet, G. (2015). The role of "Lexical Paving" in building a text according to the requirements of a target genre. In English for Academic Purposes: Approaches and Implications (pp. 43-66). Cambridge Scholars Publishing.

Boulton, A., & Cobb, T. (2017). Corpus Use in Language Learning: A Meta-Analysis. *Language Learning*, *67*(2), 348–393. doi:10.1111/lang.12224

Boulton, A., & Pérez-Paredes, P. (Eds.). (2014). Researching new uses of corpora for language teaching and learning. *ReCALL*, *26*(2).

Boulton, A., & Thomas, J. (2012) Corpus language input, corpus processes in learning, learner corpus product. In Input, Process and Product: Developments in Teaching and Language Corpora. Brno: Masaryk University Press.

Chang, J.-Y. (2014). The use of general and specialized corpora as reference sources for academic English writing: A case study. *ReCALL: the Journal of EUROCALL*, *26*(2), 243–259. doi:10.1017/S0958344014000056

Chen, Y. H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, *14*(2), 30–49.

Cobb, T., & Boulton, A. (2015). Classroom applications of corpus analysis. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139764377.027

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, *23*(4), 397–423. doi:10.1016/j.esp.2003.12.001

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. doi:10.2307/3587951

Coxhead, A., & Byrd, P. (2007). Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Second Language Writing*, *16*(3), 129–147. doi:10.1016/j.jslw.2007.07.002

Flowerdew, J. (2003). Signaling nouns in discourse. *English for Specific Purposes*, *22*(4), 329–346. doi:10.1016/S0889-4906(02)00017-0

Francis, G. (1986). *Anaphoric nouns*. Birmingham, UK: English Language Research, University of Birmingham.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, *35*(3), 305–327. doi:10.1093/applin/amt015

Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors? *System*, *32*(3), 301–319. doi:10.1016/j.system.2004.04.001

Hackin, T. (2001). Abstracting from abstracts. In M. Hewings (Ed.), *Academic writing in context: Implications and applications* (pp. 93–103). Birmingham, UK: Birmingham University Press.

Hafner, C. A., & Candlin, C. N. (2007). Corpus tools as an affordance to learning in professional legal education. *English for Academic Purposes*, *6*(4), 303–318. doi:10.1016/j.jeap.2007.09.005

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Harwood, N. (2005). '*Nowhere has anyone attempted . . . In this article I aim to do just that*': A corpus-based study of self-promotional *I* and *we* in academic writing across four disciplines. *Journal of Pragmatics*, *37*(8), 1207–1231. doi:10.1016/j.pragma.2005.01.012

Hill, J. (1999). Collocational competence. *ETP*, 11.

Hyland, K. (1999). Academic attribution: Citation and the construction of disciplinary knowledge. *Applied Linguistics*, *20*(3), 341–367. doi:10.1093/applin/20.3.341

Hyland, K. (2000). *Disciplinary discourses: Social interactions in academic writing*. London: Longman.

Hyland, K. (2002). Authority and invisibility: Authorial identity in academic writing. *Journal of Pragmatics*, *34*(8), 1091–1112. doi:10.1016/S0378-2166(02)00035-8

Hyland, K. (2008a). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, *18*(1), 41–62. doi:10.1111/j.1473-4192.2008.00178.x

Hyland, K. (2008b). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, *27*(1), 4–21. doi:10.1016/j.esp.2007.06.001

Hyland, K., & Tse, P. (2005a). Evaluative that constructions: Signalling stance in research abstracts. *Functions of Language*, *12*(1), 39–63. doi:10.1075/fol.12.1.03hyl

Hyland, K., & Tse, P. (2005b). Hooking the reader: a corpus study of evaluative that in abstracts. *English for Specific Purposes, 24*(2), 123-139.

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly*, *41*(2), 235–253. doi:10.1002/j.1545-7249.2007.tb00058.x

Ivanič, R. (1991). Nouns in search of a context: A study of nouns with both open- and closed-system characteristics. *International Journal of Applied Linguistics*, *29*, 93–114. doi:10.1515/iral.1991.29.2.93

Jiang, F., & Hyland, K. (2015). 'The fact that': Stance nouns in disciplinary writing. *Discourse Studies*, 1–22. doi:10.1177/1461445615590719

Johns, T. (1991). Should you be persuaded: Two examples of data-driven learning. *English Language Research Journal*, *4*, 1–16.

Johns, T. (2002). Data-driven learning: the perpetual challenge. In B. Kettemann & G. Marko (Eds.), *Teaching and Learning by Doing Corpus Analysis. Proceedings of the Fourth International Conference on Teaching and Language Corpora* (pp. 107-117). Amsterdam: Rodopi.

Joint Information Systems Committee. (2011). *JISC Grant funding 18/11: OER rapid innovation*. Author.

Leńko-Szymańska, A., & Boulton, A. (2015). *Multiple affordances of language corpora for data-driven learning*. Amsterdam: John Benjamins Publishing Company. doi:10.1075/scl.69

Lock, S. (1988). Structured abstracts. *BMJ: British Medical Journal, 297*(6642).

Milne, D., & Witten, I. H. (2013). An open-source toolkit for mining Wikipedia. *Artificial Intelligence*, *194*, 222–239. doi:10.1016/j.artint.2012.06.007

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge, UK: Cambridge University Press.

Nelson, M. (2006). Semantic association in Business English: A corpus-based analysis. *English for Specific Purposes*, *25*(2), 217–234. doi:10.1016/j.esp.2005.02.008

Paquot, M. (2012). *Academic vocabulary in learner writing: From extraction to analysis*. London: Continuum.

Robinson, P. (1995). Attention, memory, and the "noticing" hypothesis. *Language Learning*, *45*(2), 283–331. doi:10.1111/j.1467-1770.1995.tb00441.x

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139524780.003

Sonbul, S., & Schmitt, N. (2013). Explicit and implicit lexical knowledge: Acquisition of collocations under different input conditions. *Language Learning*, *63*(1), 121–159. doi:10.1111/j.1467-9922.2012.00730.x

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge, UK: Cambridge University Press.

Swales, J., & Feak, C. (2009). *Abstracts and the writing of abstracts. The Michigan Series in English for Academic and Professional Purposes*. Ann Arbor, MI: University of Michigan Press. doi:10.3998/mpub.309332

Szudarski, P., & Carter, R. (2016). The role of input flood and input enhancement in EFL learners' acquisition of collocations. *International Journal of Applied Linguistics*, *26*(2), 245–265. doi:10.1111/ijal.12092

Thomas, S., & Hawes, T. P. (1994). Reporting verbs in medical journal articles. *English for Specific Purposes*, *13*(2), 129–148. doi:10.1016/0889-4906(94)90012-4

Thompson, G., & Ye, Y. (1991). Evaluation in the reporting verbs used in academic papers. *Applied Linguistics*, *12*(4), 365–382. doi:10.1093/applin/12.4.365

Vyatkina, N. (2016). Data-driven learning of collocations: Learning performance, proficiency, and perceptions. *Language Learning & Technology*, *20*(3), 159–179.

West, M. (1953). A general service list of English words. Longman, Green & Co.

Witten, I. H., Bainbridge, D., & Nichols, D. M. (2010). *How to Build a Digital Library* (2nd ed.). Burlington, MA: Morgan Kaufmann.

Wu, S., Li, L., Witten, I. H., & Yu, A. (2016). Constructing a collocation learning system from the Wikipedia corpus. *International Journal of Computer-Assisted Language Learning and Teaching*, *6*(3), 18–35. doi:10.4018/IJCALLT.2016070102

Wu, S., & Witten, I. H. (2016). Transcending concordance: Augmenting academic text for L2 writing. *International Journal of Computer-Assisted Language Learning and Teaching*, *6*(2), 1–18. doi:10.4018/IJCALLT.2016040101

Xue, G., & Nation, P. (1984). A university word list. *Language Learning and Communication*, *3*(2), 215–229.

ENDNOTES

1 http://micusp.elicorpora.info/
2 http://www2.warwick.ac.uk/fac/soc/al/research/collect/bawe/
3 http://www.uclouvain.be/en-cecl-icle.html
4 http://ethos.bl.uk/Home.do
5 http://www.lextutor.ca/
6 The Open Archives Initiative Protocol for Metadata Harvesting: http://www.openarchives.org/OAI/openarchivesprotocol.html
7 These collections can be viewed at http://flax.nzdl.org
8 Greenstone is available from http://www.greenstone.org
9 EThOS toolkit is available from: http://ethostoolkit.cranfield.ac.uk/tiki-index.php
10 https://www.bl.uk/projects/british-library-labs
11 https://play.google.com/store/apps/developer?id=FLAX%20TEAM&hl=en
12 http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=flaxc383&if=flax
13 http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=flaxc404&if=
14 The database is available at http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=collocations
15 http://www.lexically.net/downloads/e_lemma.zip