### Metadata of the chapter that will be visualized in SpringerLink

<table>
<thead>
<tr>
<th>Book Title</th>
<th>Open Educational Resources in Higher Education</th>
</tr>
</thead>
<tbody>
<tr>
<td>Series Title</td>
<td></td>
</tr>
<tr>
<td>Chapter Title</td>
<td>Reflections on Remixing Open Access Content into Open Educational Resources: A New Paradigm for Sustainable Data-Driven Language Learning Systems Design in Higher Education</td>
</tr>
<tr>
<td>Copyright Year</td>
<td>2023</td>
</tr>
<tr>
<td>Copyright HolderName</td>
<td>The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd.</td>
</tr>
</tbody>
</table>

#### Corresponding Author

<table>
<thead>
<tr>
<th>Family Name</th>
<th>Fitzgerald</th>
</tr>
</thead>
<tbody>
<tr>
<td>Particle</td>
<td></td>
</tr>
<tr>
<td>Given Name</td>
<td>Alannah</td>
</tr>
<tr>
<td>Prefix</td>
<td></td>
</tr>
<tr>
<td>Suffix</td>
<td></td>
</tr>
<tr>
<td>Role</td>
<td></td>
</tr>
<tr>
<td>Division</td>
<td>Department of Computer Science</td>
</tr>
<tr>
<td>Organization</td>
<td>The University of Waikato</td>
</tr>
<tr>
<td>Address</td>
<td>Te Whare Wananga o Waikato, Gate 1, Knighton Road, Hamilton, 3240, New Zealand</td>
</tr>
<tr>
<td>Division</td>
<td></td>
</tr>
<tr>
<td>Organization</td>
<td>Durham University</td>
</tr>
<tr>
<td>Address</td>
<td>Durham, UK</td>
</tr>
<tr>
<td>Email</td>
<td><a href="mailto:alannah.fitzgerald@waikato.ac.nz">alannah.fitzgerald@waikato.ac.nz</a></td>
</tr>
<tr>
<td>ORCID</td>
<td><a href="http://orcid.org/0000-0003-0392-2740">http://orcid.org/0000-0003-0392-2740</a></td>
</tr>
</tbody>
</table>

#### Author

<table>
<thead>
<tr>
<th>Family Name</th>
<th>Wu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Particle</td>
<td></td>
</tr>
<tr>
<td>Given Name</td>
<td>Shaoqun</td>
</tr>
<tr>
<td>Prefix</td>
<td></td>
</tr>
<tr>
<td>Suffix</td>
<td></td>
</tr>
<tr>
<td>Role</td>
<td></td>
</tr>
<tr>
<td>Division</td>
<td>Department of Computer Science</td>
</tr>
<tr>
<td>Organization</td>
<td>The University of Waikato</td>
</tr>
<tr>
<td>Address</td>
<td>Te Whare Wananga o Waikato, Gate 1, Knighton Road, Hamilton, 3240, New Zealand</td>
</tr>
<tr>
<td>Email</td>
<td><a href="mailto:shaoqun.wu@waikato.ac.nz">shaoqun.wu@waikato.ac.nz</a></td>
</tr>
<tr>
<td>ORCID</td>
<td><a href="http://orcid.org/0000-0001-9566-005X">http://orcid.org/0000-0001-9566-005X</a></td>
</tr>
</tbody>
</table>

#### Author

<table>
<thead>
<tr>
<th>Family Name</th>
<th>König</th>
</tr>
</thead>
<tbody>
<tr>
<td>Particle</td>
<td></td>
</tr>
<tr>
<td>Given Name</td>
<td>Jemma</td>
</tr>
<tr>
<td>Prefix</td>
<td></td>
</tr>
<tr>
<td>Suffix</td>
<td></td>
</tr>
<tr>
<td>Role</td>
<td></td>
</tr>
<tr>
<td>Division</td>
<td>Department of Computer Science</td>
</tr>
<tr>
<td>Organization</td>
<td>The University of Waikato</td>
</tr>
</tbody>
</table>
This chapter presents a new paradigm for sustainable data-driven language learning systems design in higher education that draws on qualitative reflections spanning a decade (2012–2022) with stakeholders from an ongoing global research study with the FLAX (Flexible Language Acquisition) and F-Lingo projects at the University of Waikato in Aotearoa New Zealand (Fitzgerald (2019) A new paradigm for open data-driven language learning systems design in higher education; König et al. (2022) Smart CALL).

Design considerations are presented for remixing domain-specific open access content into Open Educational Resources (OER) for academic English language provision across formal and non-formal higher education contexts. Primary stakeholders in the research collaboration include the following three groups: (1) Knowledge organisations that provide open access to academic content—libraries and archives, including the British Library and the Oxford Text Archive, universities in collaboration with MOOC providers and the CORE (COnnecting REpositories) open access aggregation service at the UK Open University; (2) Researchers who mine and remix academic content into corpora and open data-driven language learning systems—converging from the fields of open education, computer science and applied corpus linguistics; (3) Knowledge users who re-use and remix academic content into OER—English for Academic Purposes (EAP) practitioners from university language centres. Automated content analysis was carried out on a corpus of interview and focus discussion data with the three stakeholder groups in this research. We discuss themes arising from the research data that reflect the different stakeholders’ experiences of remixing open access research content that has been produced within the academy for re-use as open educational content for teaching and learning features of academic language within open data-driven language learning systems. These open learning systems have been specifically designed to scale with OER expansion and traction in mind for their sustainable uptake both within and beyond the brick and mortar of the traditional university. The new paradigm presented in this chapter challenges, as the OER movement must, established business models and deeply embedded cultural or institutional norms that present obstacles to OER expansion and traction and the sustainability of the movement. One persistent challenge concerns the lack of open education policy across the higher education sector for full open
access (for use, modification, adaptation) via Creative Commons licensing to content produced within the academy. Thus, while this research has theoretical and practical implications in applied linguistics, computer science, language teaching and learning and open education, more generally, it also has significant cultural, business model and policy implications for higher education.

**Keywords**

(separated by '-') Data-driven learning - Design-based research - English for academic purposes (EAP) - Higher education - Massive open online courses (MOOCs) - Open access - Open educational practices - Open educational resources (OER) - Systems design
Chapter 6

Reflections on Remixing Open Access Content into Open Educational Resources: A New Paradigm for Sustainable Data-Driven Language Learning Systems Design in Higher Education

Alannah Fitzgerald, Shaoqun Wu, Jemma König, Steven Shaw, and Ian H. Witten

Abstract
This chapter presents a new paradigm for sustainable data-driven language learning systems design in higher education that draws on qualitative reflections spanning a decade (2012–2022) with stakeholders from an ongoing global research study with the FLAX (Flexible Language Acquisition) and F-Lingo projects at the University of Waikato in Aotearoa New Zealand (Fitzgerald (2019) A new paradigm for open data-driven language learning systems design in higher education; König et al. (2022) Smart CALL). Design considerations are presented for remixing domain-specific open access content into Open Educational Resources (OER) for academic English language provision across formal and non-formal higher education contexts.

Primary stakeholders in the research collaboration include the following three groups:
1. Knowledge organisations that provide open access to academic content—libraries
2. Academic institutions that provide academic content
3. Technology providers that provide technology infrastructure and tools

A. Fitzgerald  ·  S. Wu  ·  J. König  ·  I. H. Witten
Department of Computer Science, The University of Waikato, Te Whare Wananga o Waikato, Gate 1, Knighton Road, Hamilton 3240, New Zealand
e-mail: alannah.fitzgerald@waikato.ac.nz

S. Wu
e-mail: shaoqun.wu@waikato.ac.nz

J. König
e-mail: jemma.konig@waikato.ac.nz

I. H. Witten
e-mail: ihw@waikato.ac.nz

A. Fitzgerald
Durham University, Durham, UK

S. Shaw
Department of Education, Concordia University, Faubourg Ste-Catherine Building, 1610 St. Catherine West, Montreal, QC H3G 1M8, Canada
e-mail: steven.shaw@concordia.ca

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
and archives, including the British Library and the Oxford Text Archive, universities in collaboration with MOOC providers and the CORE (COnnecting REpositories) open access aggregation service at the UK Open University; (2) Researchers who mine and remix academic content into corpora and open data-driven language learning systems—converging from the fields of open education, computer science and applied corpus linguistics; (3) Knowledge users who re-use and remix academic content into OER—English for Academic Purposes (EAP) practitioners from university language centres. Automated content analysis was carried out on a corpus of interview and focus discussion data with the three stakeholder groups in this research. We discuss themes arising from the research data that reflect the different stakeholders’ experiences of remixing open access research content that has been produced within the academy for re-use as open educational content for teaching and learning features of academic language within open data-driven language learning systems. These open learning systems have been specifically designed to scale with OER expansion and traction in mind for their sustainable uptake both within and beyond the brick and mortar of the traditional university. The new paradigm presented in this chapter challenges, as the OER movement must, established business models and deeply embedded cultural or institutional norms that present obstacles to OER expansion and traction and the sustainability of the movement. One persistent challenge concerns the lack of open education policy across the higher education sector for full open access (for use, modification, adaptation) via Creative Commons licensing to content produced within the academy. Thus, while this research has theoretical and practical implications in applied linguistics, computer science, language teaching and learning and open education, more generally, it also has significant cultural, business model and policy implications for higher education.

Keywords Data-driven learning · Design-based research · English for academic purposes (EAP) · Higher education · Massive open online courses (MOOCs) · Open access · Open educational practices · Open educational resources (OER) · Systems design

6.1 Introduction

In this chapter section, we present a new research paradigm for sustainable data-driven language learning systems design in higher education. This research paradigm provides a theoretical and conceptual framework supported by a review of the relevant literature from intersecting fields in this research. As we progress through the chapter, the paradigm will be unpacked in greater detail in the subsequent sections as we drill down into the specifics of the research contexts, materials and methods that have been employed in the research with the three stakeholder groups. Reflections on remixing open access content into OER for teaching and learning features of academic English, along with the affordances and challenges encountered by the stakeholder groups, will be presented in the final discussion section of this chapter.
6.1.1 Research Paradigm

A basic premise underpinning the new research paradigm presented in this chapter is that open data-driven language learning systems design as an approach is learner-centric and operates with the interface to the learner. Whether the learner is operating fully online in non-formal or informal learning mode or in a blended modality that is based both within and beyond the formal university language classroom, this approach requires that the tools and interfaces, and indeed the academic language corpora, be openly accessible and remixable for development or adaptation to meet this specific learner requirement. This method is different from existing Data-Driven Learning (DDL) approaches which assume specialised knowledge or experience with DDL tools, interfaces and strategies, operating on mostly inaccessible corpora in terms of cost or design, or assuming training to, hopefully, compensate for this lack of knowledge and experience (Fitzgerald, 2019; Pérez-Paredes et al., 2018).

The term DDL was coined by corpus linguistics and EAP pioneer, Tim Johns, to refer to a computer-driven language learning support approach with linguistic content that has been automatically analysed, enriched and transformed into a data-mined resource that learners can browse and query (Johns, 1991a). Johns envisioned every language learner as “a Sherlock Holmes” with direct access to the evidence of real-world language data (Johns, 2002, p. 108). In a similar vein to contemporary advocates for using and developing a broad spectrum of data literacies with open data in higher education (Atenas et al., 2015), Johns also envisioned DDL as developing data literacies for understanding and interpreting linguistic data for direct applications in language learning, specifically in the context of higher education (Johns, 2002; Pérez-Paredes et al., 2018).

From a research and development (R&D) standpoint, the paradigm presented here also operates with the interface to knowledge organisations (universities, libraries, archives) and researchers who are engaging with open educational practices to push at the parameters of open policy for the non-commercial re-use and remix of authentic research and pedagogic content that is increasingly abundant in digital open access format for text and data mining (TDM) purposes. This open access content is highly relevant to learning features of specialist varieties of English from across the academy but is otherwise off-limits for development into proprietary learning materials by the commercial education publishing industry (Fitzgerald et al., 2015, 2017; Wu et al., 2018). Indeed, the open corpus development work presented in this chapter would not have been possible had it not been for the campaigners for copyright reform, the Internet activists, the open policymakers, the open-source software developers and the advocates for open access, open data and open education that have made these resources available for re-use and remix.

This paradigm leads down several paths, including research into understanding how users actually perceive, appropriate and use the approach based on the open tools and resources provided. This inquiry informs their design and development in an R&D process that is presented here through the methodological lens of design-based research (Fitzgerald, 2019). This approach will be fundamentally different than if we
assume the user is actually a DDL or linguistics expert or that such an expert will be the learner’s interface to the system by preparing output for the learner to experience and learn from (Johns, 1991b). This approach will necessarily also be different than if we assume the user is always a formally registered student at a university with access to EAP support that may or may not offer DDL or linguistics expertise for learning the language features of specific discourse communities from across the academy. The assumption behind this new paradigm that the right tools and resources can allow the end-learner to drive the processes autonomously is fundamentally revolutionary. This premise goes to the original contribution to the knowledge of this research but also challenges and directs researchers and practitioners in the field to consider and take up this new direction with open data-driven language learning systems design for applications that can be scaled in higher education to meet the increasing numbers of learners who are coming online in increasingly uncertain times (Fitzgerald, 2019; König et al., 2022).

The focus on domain-specific language learning support via data-driven approaches is, of course, also decidedly different from the current English for Academic Purposes (EAP) paradigm, which in mainstream practice has been steadily evolving away from its roots in English for Specific Purposes (ESP), domain specificity and DDL processes towards the generic skills and knowledge programs currently in vogue that are arguably being steered by generic EAP coursebook publications from the commercial education publishing industry (Gillett, 2018). Thus, this is also a new paradigm based on DDL approaches, driving domain-specific language learning support for EAP across formal, non-formal and informal learning modalities in higher education. It will transform, potentially, the focus of DDL systems design developments in language support and learning in general towards the non-specialist end-learner but also hopefully help re-establish the centrality of language specificity to the field of EAP (Anthony, 2018).

This new paradigm is necessarily rooted in greater multi- or trans-disciplinarity (Colpaert, 2004, 2018). Given the goal of facilitating, in particular, the increasing number of learners who are coming online in these uncertain times, and users of large-scale MOOC platforms who are trying to function in domain-specific subject areas that are invariably offered in the English language, the approach requires collaboration and cooperation among platform providers, subject academics and instructors, educational technologists, software developers, educational researchers, EAP practitioners, linguists with expertise in corpus-based and DDL approaches and policymakers in knowledge organisations (libraries, universities, archives). It has to be remarked, also, that the value and significance of this multi-disciplinary work is amplified by our current situation in higher education with the pandemic, which has seen a massive, urgent push to move learning online with an accompanying impetus to identify, adapt and leverage learning content worldwide and to exploit open educational resources, in particular.
6.2 Research Context

The open access movement in research and higher education has bolstered unprecedented access to artefacts of the academy in the form of published research articles, in addition to online platforms and services for accessing unpublished theses and pedagogic materials. One example is open access to transcribed video lectures and course reading content from the world’s leading universities and institutions with an expanding provision in MOOCs. A further example is an open access to a growing corpus of over half a million PhD theses from universities across the UK with the British Library’s Electronic Theses Online Service (EThOS). Both of these examples will feature for discussion in this chapter with respect to the nuanced meanings of openness and the tensions around human and machine re-use of content; the latter of which involves computational processes whereby texts and data are crawled and mined by software to build on and create new knowledge and derivative resources. Specifically, the research presented in this chapter is concerned with stakeholder reflections on a new paradigm for the co-design and co-development of data-driven language learning systems derived from open access content.

A definition for open access appeared for the first time in the declaration of the Budapest Open Access Initiative (BOAI):

By “open access” to [peer-reviewed research literature], we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited (BOAI, 2002).

Colpaert (2016) from the field of Computer Assisted Language Learning (CALL) divides uses for data in language education into two main categories depending on divergent goals for re-use: data as content and data as information. The former data category includes authentic content found on the Web, including open access content that makes up the primary focus of this chapter. In contrast, the latter category includes information about data, otherwise known as metadata, which we also make use of in our research and refer to in this chapter.

Corpus linguistics researchers have demonstrated the importance of viewing language as data (Anthony, 2014; Hunston, 2002; McEnery et al., 2006; Sinclair, 2004). By way of extension, DDL can be viewed as a means for language teachers and learners to obtain, organise and study authentic language data derived from corpora in language education (Boulton & Cobb, 2017; Boulton & Pérez-Paredes, 2014; Boulton & Thomas, 2012; Chang, 2014; Cobb & Boulton, 2015; Vyatkina, 2016). There remains a persistent lack of exposure to and use of corpus-based systems and Natural Language Processing (NLP) tools by language practitioners in mainstream language education, however:
Many of the 15 million English teachers in the world today, according to the British Council Annual Report (2010), have never heard of corpora, while many who are familiar with their use by lexicographers and grammarians are not aware that they can use them themselves, as could their students. (Thomas, 2017, p. 17)

For this R&D project, we identified a range of open, authentic domain-specific text and data sources that are of perceived value to the EAP community yet are off-limits for commercial re-use and development by the English language content publishing industry. In this chapter, we will share reflections on our work with representatives from knowledge organisations that manage and curate digital open access content, such as the British Library, who are working at the cutting edge of reforms in UK copyright law to create open access policy within their Research and Reuse Committee. In line with the Fair Use Doctrine, which is a limitation to US copyright law, an important exception and limitation to UK copyright law for TDM was introduced in 2014, whereby permissions were established for the non-commercial re-use of digital research content following an independent government report (Hargreaves, 2011).

One of the aims of this research has been to bring corpus linguistics researchers and EAP practitioners to the interface of data-driven language learning systems design for higher education through open initiatives in software development, research, education and publishing that support the co-design, co-creation and distribution of such systems. A further aim of this research has been to explore the potential of working with open, authentic academic texts that afford language specificity (Hyland, 2002; Strevens, 1988) in the development of teaching and learning resources for EAP that reflect the specific language and discourse features from target academic communities.

We will discuss the perceived value that EAP researchers, teachers and managers place on the efficacy of utilising authentic open access academic texts and corpora in data-driven approaches for blended learning. These perceived educational values will be weighed against the perceived risks held by knowledge organisations and the individuals working therein, such as curators, subject academics and educational technologists, regarding the remix and re-use of digital open access content and collections for non-commercial research and education purposes. For the scope of this chapter, we will explore the following research questions:

1. To what extent can open access content foster open educational practices among academic English language stakeholders for designing, developing and evaluating data-driven language learning resources?

2. What impact do the underlying business models and cultural practices of institutions and organisations have on open educational practices for remixing open access content in the design, development, implementation and dissemination of resources for EAP in higher education?
6.2.1 Research Materials

With this research, we have placed particular emphasis on co-designing and co-creating language learning systems for pedagogic purposes rather than for corpus linguistics research purposes. Drawing on the concept of knowledge mobilisation (Levin, 2011), our goal is to engage relevant stakeholders in moving available knowledge from research in corpus linguistics, open education and computer science (NLP and TDM) towards knowledge users, namely EAP practitioners and learners. The goal is for knowledge users to not only benefit from the research but to collaborate directly in an iterative design-based research process. Intermediaries working in knowledge organisations have acted as brokers and open education champions in this research by creating access to knowledge artefacts that are valued for re-use in EAP via initiatives in open access policy and reforms in copyright law.

Although the findings from this research are tied to issues with designing and developing open access content into data-driven learning systems, wider issues vis-à-vis the re-use and remix of open access content in language materials development practices will also be discussed as they apply to both classroom teaching and online learning. The Appendix at the end of this chapter provides an overview of our work to date. It identifies the knowledge organisations, researchers and knowledge users who have collaborated on the design and development of open data-driven systems for learning aspects of academic English in formal and non-formal higher education contexts with the FLAX\(^1\) and F-Lingo\(^2\) projects.

6.3 Research Methods and Results

Methods for collecting data from different participant groups in different locations over a period of years included: focus discussions, face-2-face and Skype interviews and email exchanges stemming from project meetings on observations and evaluations shared in this situated research. Three knowledge organisations have participated in the research (The British Library, The Oxford Text Archive and the Connected Repositories research group at the UK Open University). Eight researchers working in the area of corpus and computational linguistics and open education have participated in the research from higher education institutions in Aotearoa, New Zealand, Spain, Canada and the United Kingdom. Seven knowledge users working in EAP teaching and management from two UK universities have also participated in the research. Automated content analysis (ACA) was carried out on the complete corpus employing the Leximancer software version 4.5, and then on sub-corpora corresponding to data from the three stakeholder groups engaged in this research—knowledge organisations, researchers and knowledge users. Results from the ACA in this study were checked and then triangulated with participants.

---
\(^1\) http://flax.nzdli.org/greenstone3/flax.
\(^2\) https://chrome.google.com/webstore/search/flingo.
in this qualitative research to create opportunities for participants to comment on transcripts and emerging findings. Thematic and conceptual findings in the datasets were then confirmed with participants as they pertain to reflections on the iterative design processes for designing open data-driven systems for academic English. The complete corpus and ACA visualisation maps of key themes and concepts from this R&D project are available for viewing on the Open Science Foundation\footnote{https://osf.io/gbkzp/} data platform.

6.3.1 Design-Based Research

Action research is a widely employed methodology in English language education research and teacher training programmes (Burns, 2009) and shares many of the same principles as design-based research (DBR). Pragmatism is central to both approaches, often employing mixed methods of inquiry to arrive at tangible solutions to educational problems. Within action research cycles, individual teaching practitioners carry out classroom teaching interventions to observe, record and reflect on the impact of these interventions over time to inform and improve their classroom and online teaching practice (Reason & Bradbury, 2007). In design-based research, another layer exists that requires educational practitioners to collaborate with research and design teams (Anderson & Shattuck, 2012).

Although DBR has sustained great interest from researchers and practitioners within the instructional design and educational technology milieu, it is nevertheless a long-term and very resource-intensive exploratory research method with goals and outcomes that are difficult to define. The literature on DBR attests to “a series of approaches, with the intent of producing new theories, artefacts, and practices” (Barab & Squire, 2004, p. 2). More specifically, these approaches have been defined as multiple research cycles that include numerous iterations of analysis, design, development, evaluation and revision (Burkhardt, 2006; Walker, 2006; Amiel & Reeves, 2008; Hakkarainen, 2009; McKenney & Reeves, 2012). Data are collected over a minimum of several weeks but, in most cases, are collected over several months or years (Herrington et al., 2007) as has been the case with our research, which has been ongoing for over a decade now (Fitzgerald, 2019).

6.3.2 Automated Content Analysis

Automated Content Analysis (ACA) is situated within the framework of computational social sciences. It refers to a range of algorithms that employ probabilistic models, namely topic models and concept mapping models (Blei, 2012a, b), that iteratively infer the themes and concepts present within a corpus. ACA can be traced back
to the theoretical underpinnings of Latent Semantic Indexing (LSI; Papadimitriou et al., 1998), leading to the three-level Bayesian model of Latent Dirichlet Allocation (LDA; Blei et al., 2003). The current state of the art with ACA models involves the identification and analysis of higher levels of complexity found within thematic structures (Blei, 2012b). Current ACA systems include features for analysing “syntax, concept hierarchies, document networks and temporal trends in themes, furthering our ability to visualize and explore the literature” (Nunez-Mir et al., 2016). ACA is primarily used to automatically analyse text in digital format but also, increasingly, media content, e.g., images (Boumans & Trilling, 2016).

In this section, we look through the analytical lens offered by ACA at the different themes and concepts from each of the three participant groups in this study: knowledge organisations, researchers and knowledge users. Due to the limited scope of this publication, we will only be looking at the results of the top four themes in each sub-dataset for the three participant groups in this research. Where we present a summary and discussion of results from all three sub-datasets, themes and concepts from the data will be italicised.

Our reasons for employing the Leximancer ACA software to analyse the qualitative datasets were two-fold: to increase validity and to determine the lexical co-occurrence of information extracted from natural language into semantic or conceptual patterns using automated methods. Leximancer has been designed to mitigate subjectivity and researcher bias in the traditional content analysis processes of manual text analysis, coding and intercoder reliability testing (Weber, 1990). Through powerful automated methods, Leximancer is devised to make the human analyst aware of “the global context and significance of concepts and to help avoid fixation on particular anecdotal evidence” (Smith & Humphreys, 2006, p. 262). Leximancer performs two types of analysis on a ranked list of lexical terms found in a unified body of text or corpus: conceptual analysis and relational analysis. The conceptual analysis measures the presence and frequency of concepts in a document set by extracting words, phrases, or collections of words that represent a concept. The relational analysis is concerned with measuring the co-occurrence of concepts within a document set, extracting these co-occurring concepts to show their relationship.

The design principles that underpin the Leximancer software are founded on observations from the fields of corpus linguistics, computational linguistics and psycholinguistics, resulting in the development of the semantic and relational Leximancer algorithms that are employed in both stages of the software’s co-occurrence information extraction technique (see Smith, 2000a, b, 2003). Leximancer was employed to mine the total qualitative dataset and sub-datasets for each participant group, resulting in a thesaurus of words identified within each corpus analysed along with their related meanings and surrounding words or collocates. The complete corpus and ACA visualisation maps of key themes and concepts from this R&D project are available for viewing on the Open Science Foundation data platform.

As shown in Fig. 6.1, closely related words from the complete qualitative dataset in this study are identified by the ACA software as concepts and are represented as dots
within thematic circles of interrelated concepts on a concept map. The key below the map indicates how many times the central themes occurred in the corpus. Important themes are mapped with warm colours. For example, research and FLAX appear in red and brown on the concept map (Angus et al., 2013). These two dominant themes are represented as tightly packed circles containing concept dots in close proximity to one another. The spatial alignment of these dots indicates how closely related concepts are within each key theme (Smith & Humphreys, 2006). For instance, research, corpus, able, EAP, teaching and learning are closely related concepts within the dominant research theme. Thematic circles are sometimes shown as overlapping with one another when concepts occur close to or across neighbouring themes, such as the concepts for corpus and learning within the open and research themes, which are central to this ongoing design-based research with the FLAX project and will provide a basis for the discussion section of this chapter.

6.3.2.1 Knowledge Organisations

The Leximancer analysis of data from the knowledge organisations group reveals text as the major theme. The concepts within this key theme of text emphasise experimentation with corpora and stuff, with one frequent example in the dataset being the EThOS (Electronic Thesis Online Service) PhD thesis content at the British Library, in addition to the terms around re-use, and what you are able to do when using texts with text and data mining. The second most prominent theme is work with concepts reflecting the importance of doing work in the open as central to this design-based research with knowledge organisations. In close orbit to the text theme are the overlapping themes of trying and example, representing the third and fourth most frequent themes in the dataset, coming in closely behind the work theme. Of note in the trying theme are the connected concepts of people trying to do things. Re-use is the concept shared between the overlapping text and example themes. Also apparent in the theme are the key interlinked concepts of example, collections and metadata for what can probably be looked at with respect to research and development that focus on the re-use of text and their metadata from digital collections. In the discussion section, we will explore these themes and concepts further with reference to the terms and conditions around open access content re-use in this research with knowledge organisations.

6.3.2.2 Researchers

We now turn to interview data with education researchers who have worked with the FLAX project. The first researcher interviewed was Researcher 4, a legal English corpus researcher at the University of Murcia in Spain who developed the British Law Reports Corpus (BLaRC) with judicial hearings from around the world that subscribe to the English common law system. The corpus was made available with an open access government licence from the British and Irish Legal Institute (BAILI).
Fig. 6.1 Concept map and key derived from automated content analysis of the complete qualitative dataset
Researcher 6 also conducted doctoral research into lexical bundles with the FLAX project, focusing on the Chinese and New Zealand EAP contexts.

When we look at the Leximancer conceptual analysis for the researchers’ group, of note are four prominent and overlapping themes: FLAX, students, teachers and time. Moreover, the concepts of access, different, research, online, language and learning appear in the overlapping foci areas of these top four central themes. In this section, we will summarise the findings from these concepts, which will form the basis for the discussion section of this researcher participant group later in the chapter. The access concept in particular, which appears in the overlap between the FLAX and students themes in the ACA, is expressed in the data as issues related to conducting research that provides students with access to and use of different corpora, data and systems in FLAX that can support their online language learning with formal language courses and non-formal MOOCs. Of interest, the access concept is also expressed in the data in relation to the issue of gaining access to students through working with language teachers to conduct research into the use of the FLAX system. This last point on access is further extended into the sixth most frequent theme in the dataset, study, with concepts expressing the need for user studies on the uptake of FLAX. In addition, the issue of access is further expressed by how teachers may be interested in working with the FLAX project but are limited in terms of the fourth most frequent theme, time, due to the heavy emphasis placed on teaching and learning and not on conducting research at their institutions.

6.3.2.3 Knowledge Users

Of the seven EAP practitioners who participated in the research, only one (Knowledge User 1) from Queen Mary University of London (hereafter referred to as QMUL), had extensive experience with using corpus tools in his classroom teaching, namely the Sketch Engine\(^5\) suite of tools for querying and sketching corpora. The three other participants at QMUL (Knowledge User 2, Knowledge User 3 and Knowledge User 4) all had a background in CALL for developing online EAP resources, most notably Academic English Online.\(^6\) The three EAP teachers at Durham University who are former EAP teaching colleagues of Researcher 1 (Knowledge User 5, Knowledge User 6 and Knowledge User 7), were early adopters and advocates for using open-source software and OERs in their classroom teaching as a means of ensuring that their students had access to high-quality free and open online teaching and learning resources during and after their courses had finished. The EAP practitioners in this study expressed that the motivation to adopt open educational practices as they apply to academic practice in higher education was a motivating factor for participating in the research with the FLAX project.

The dominant themes arising from the Leximancer analysis of interviews and focus discussions from project meetings with knowledge users—EAP teachers and

---

\(^5\) [https://www.sketchengine.eu/](https://www.sketchengine.eu/).

\(^6\) [http://aeo.sllf.qmul.ac.uk/](http://aeo.sllf.qmul.ac.uk/).
course managers—are EAP followed closely by students, things and people. In summary, results from the ACA of this sub-dataset point to issues concerned with the concepts of EAP and the teaching of academic English language from the largest theme, EAP. The second-largest theme in the data, students, reveals issues around materials for teaching students that teachers are developing themselves or those materials that have been developed by commercial publishers and reflections on what does and does not work in practice. The third most frequent theme in the dataset, things, is representative of concepts related to what needs to be done with research using things and materials. In the fourth most frequent theme, people, an interesting interplay of concepts are revealed in reference to people as being those EAP teachers working in universities who do or do not create access to open resources for education, and also in reference to people outside of the university who can and cannot access open resources for education. The themes and concepts outlined here in this section will be explored in more depth in the corresponding discussion section of this paper on knowledge users.

The work at Durham University took the form of an OER cascade training project with the participating EAP practitioners and their students that introduced them to four online data-driven text analysis language learning systems: Lextutor,7 AntConc,8 Word and Phrase9 and FLAX. This OER cascade training work led to collaborative evaluations and further development iterations of the Learning Collocations collection in FLAX with the addition of the open access British Academic Written English (BAWE) corpus managed by the Oxford Text Archive for a specific focus on academic English collocations. It was written up as a case study for the UK Higher Education Academy (Fitzgerald, 2013). This work at Durham also resulted in the development of the full-text BAWE collections in FLAX that focused on novel ways to search and browse augmented academic texts that represented different genre types from across the disciplines of the arts and humanities, the social sciences, the physical sciences and the life sciences (Wu & Witten, 2016).

The work at QMUL focused on design collaborations with open access PhD thesis abstract content managed by the British Library for the development of domain-specific micro-corpora and interactive games with Android mobile apps for uptake on QMUL’s pre-sessional EAP programmes (Fitzgerald et al., 2014). The work with QMUL led to a further design iteration with the development of the much larger PhD Abstract collections in FLAX of 9.8 million words (Wu et al., 2018).

6.4 Discussion

This section discusses prominent themes and interrelated concepts from the ACA of the qualitative datasets. We drill further down into the data that captured reflections

7 https://www.lextutor.ca/.
8 http://www.laurenceanthony.net/software.html.
9 https://www.wordandphrase.info/.
from participants in the research to present relevant themes and concepts identified in transcriptions. Where we refer to actual data for discussion, themes and concepts will be italicised.

6.4.1 The Four Pillars of Re-Use in Knowledge Organisations

Our research with knowledge organisations in developing open corpora for EAP shows that it often comes down to those individuals working on the inside who are reasonably au fait with copyright law as it pertains to open access and open educational practices and who are willing to champion the re-use of resources and encourage the development of open policies within their organisations. We have seen this type of open access policy championship with the EThOS service team manager and the British Library Labs project manager. The progress with policy development for open access and re-use that enable TDM approaches with digital collections at public knowledge organisations such as the British Library is contrasted with the absence of open education policy in higher education institutions, where there has been less progress made with the re-use of educational content. Open access, in most cases, to read-only research publications and, in lesser cases, to pedagogic content has become the default re-use position of most universities and mainstream MOOC providers.

Once again, those individuals who are already open education practitioners who have openly licensed their educational resources with Creative Commons licenses have enabled the FLAX team to develop derivative language learning collections. Open licensing supports their wider practices in open digital scholarship (Weller, 2011)—via blogs, public lectures, MOOCs, networked courses etcetera—to widely promote the subjects they are passionate about. Notably, Professor Fisher of the CopyrightX micro-networked course has deliberately applied his expertise in understanding the ins and outs of copyright law by licensing his teaching and learning content as CC-BY with Creative Commons “to maximise the number and variety of educational projects and derivative works that can be built (directly or indirectly) on our foundation—and thus the set of students who might benefit from our efforts”. (Fisher, 2014, p. 17).

The participating knowledge organisations in this research differ regarding policies and practices around re-use. British Library Labs (BL Labs) is an Andrew Mellon Foundation-funded initiative that supports the remix and re-use of the British Library’s digital collections and data for research and educational purposes. In an interview with the project manager of BL Labs (Knowledge Organisation Representative 1 in the transcript corpus), we discussed the FLAX project research with the EThOS PhD theses dataset for the development of the PhD Abstract collections, wherein he identified four pillars for the re-use of this dataset that can be broadly applied to the re-use of other digital collections at the British Library:
(1) “Do we have an expert with curatorial knowledge of a particular collection who is on board with re-use? Some curators are not concerned about that at all. All they care about is the preservation and not about who uses it.

(2) Do we know where it, the collection, is? A description of something is one thing but who actually has the digital files? Can they be accessed?

(3) Is there any metadata? That obviously helps enormously because it means that you can then release the metadata, normally. But even metadata has licenses as well… so, who owns that metadata?

(4) Is the collection close to being copyright-cleared? And what I mean by that, I actually mean, is it, could it potentially, easily, be available under an open licence?”

(Interview with BL Labs Manager, British Library, London, UK)

With the harvested PhD theses in EThOS at the British Library, the provenance is very mixed, whereby there is no one set of terms and conditions for re-use of the open access content found therein. This phenomenon is largely a reflection of the different universities where the research was carried out and is dependent on whether or not there were industry investments in the research, for example, which would result in copyright stakes. Due to this mixed provenance, the British Library has undertaken measures to balance any possible research instances of re-use with any identifiable potential risks such as mass copying, misrepresenting and misquoting of the EThOS dataset. As with the Oxford Text Archive, a cautious approach has been adopted at the British Library with respect to TDM, whereby collections are only available for non-commercial re-use purposes on a request-only basis. The BL Labs manager does, however, acknowledge the iterative nature of research and encourages the practice of “dogfooding” at the British Library, whereby collections management teams, such as the EThOS team, engage in internal research on the re-use and remix of collections to anticipate affordances and hindrances with conducting research:

Knowledge Organisation Rep 1: First, to work with a collection it’s important to ensure that there’s a human being who can tell you the story of that collection because you don’t know what may be lurking in there and it may not be about legal issues. It could be political. It could be financial. But that information isn’t always documented.

Researcher 1: Sorry to interrupt you there, but were there any issues around EThOS?

Knowledge Organisation Rep 1: Well, I think there are still issues really because the problem of doing this work is because the intellectual property is going to be dependent on the institution and their relationship with their students. It seems that that is not straightforward with all the different institutions. So, if you do a PhD at an institution, you’re under the IPR for that work, and I think that different universities have different views and policies.

Researcher 1: So, it’s not always automatically the student’s work? I thought it was.

Knowledge Organisation Rep 1: All I know is that some work, some PhD work, is embargoed because it has commercial sensitivities in there. So, for example, somebody might…

Researcher 1: Because they’ve been funded by…?
Knowledge Organisation Rep 1: Yeah, because they’ve been funded by Panasonic, for example.

Researcher 1: Yeah, I get that.

Knowledge Organisation Rep 1: There could be, depending on the PhD and the funding stream, so it could not only be the university, it could be the funder, the funder might have certain requirements. It could be commercial. It could be a funding council. What you’re getting is a harvested bunch of stuff in EThOS where the provenance is very mixed, and I think the team have decided to take a very cautious approach in terms of being able to do things like text and text and data mining, so, you know, it’s on a request only basis. Because, especially, you know, about the possibility that there could be commercial re-use.

Researcher 1: Yes, I think that’s getting back to your original point about the library wanting to know what your research questions were before doing the work.

Knowledge Organisation Rep 1: Exactly.

Researcher 1: And that’s when somebody puts in a request, for example. We want to re-use these texts for these purposes, and this is what the end result will look like kind of.

Knowledge Organisation Rep 1: Yeah, but the problem with that is, in our experience, is that research doesn’t work like that. With research you don’t know what you’re going to get. You might know your research questions, but the whole point and nature of research is that it’s iterative. You know, you experiment.

Researcher 1: I’m glad to hear you say that because, you know, that was our experience with the Oxford Text Archive when we requested the BAWE corpus. Because we didn’t know in advance that we’d be Wikifying whole texts but then we had the technology to do it. In particular, I mean, all the prior work we had done with Wikipedia mining at the Digital Library Lab at Waikato. And, we thought, well, Wikification may well be useful for language learning so let’s add this functionality for learners. So, the BAWE collections became our first Wikified collections, and you can see this feature in our subsequent collections, including the PhD Abstract collections with EThOS metadata. But this work with Wikification wasn’t in our initial request to the OTA, which was instead very general in terms of what we were proposing to do.

Knowledge Organisation Rep 1: Yeah, I think in general, I understand why there needs to be this clarity but unfortunately, it’s a complete misunderstanding of the whole scholarly process. The scholarly process is actually incredibly creative, and you know, you don’t know by the very nature of research, that you don’t know what you’re going to find. And, you know, it’s surprising what comes along the way. Ideas will come along the way, and that’s just the nature of research. So, we have found that really challenging. And, what we’ve decided to do, I think, is to be working on research questions where they can be sort of dealt with on a case-by-case basis, and also to agree on what the outcomes are going to be. So that, like, if people want to publish work, what actually can be published, and what can’t be published because of the sensitivities at the moment. We’re also having quite a lot of requests to do text and data mining work with our non-print legal deposit stuff.

(Interview excerpt with BL Labs Manager, British Library, London, UK)
6.4.2 Issues of Access in DDL Research

The automated content analysis of the entire qualitative dataset reveals a direct link between the knowledge organisations and researchers’ sub-groups with overlapping themes of access. Put simply, access to digital collections that can be re-used by researchers, in this case, corpus linguistics and open education researchers, is due in no small part to the open access and open education policies adopted by knowledge organisations and the gatekeepers working within those organisations who implement these policies to promote open access and re-use.

We turn first to a discussion with Researcher 4 in this study on the perceived affordances of re-using and remixing open access publications for open data-driven learning in DDL research with reference to the BLaRC of 8.85 million words (Marín et al., 2014), which is derived from open access judicial hearings licensed with a government license and available from the BAILII online service. Marín developed the BLaRC due to the lack of relevant, authentic materials for teaching the specific area of legal English in EAP. We invited her to include her corpus on the FLAX website so that it would be openly accessible for data-driven language learning in addition to corpus linguistics research. Researcher 4 was interviewed about the making of the BLaRC, which highlights the affordance of the access concept as a prominent concept in the interview data with applied corpus linguistics researchers, and how this had enabled the development of legal English resources from open access content in comparison with proprietary legal content services that require licence subscriptions:

Researcher 1: You know, my next question: Could you even have built the BLaRC without those open government licenses on all of those documents, those judicial hearings in the BAILII (British and Irish Legal Information Institute)?

Researcher 4: No, that’s the thing, that’s the thing. The amazing discovery was the BAILII [...] I was thinking about buying a licence for LexisNexis, I think it’s called. There are a couple of them, which cost a fortune, a fortune. I’m not sure but I think law firms, they pay, I don’t know, four or five thousand pounds a year for having that kind of thing, which is amazing [...]

Researcher 4: Actually, the University of Murcia doesn’t have access to that database because one of my colleagues was in Madrid, she was a visiting researcher there, and she downloaded like a hundred thousand texts from LexisNexis because she didn’t know that the BAILII existed. So, when she came here, and we were talking, and I said, look there’s this site [the BAILII] and they have added a lot of overseas legal documents, including United States documents. They have the whole planet in there. It’s amazing how much stuff you can find. So, to me it was a huge, huge discovery. That was the best thing that could have happened to me. That’s why I started my research on legal corpora. I mean that was one of the reasons.

Researcher 1: Access is so key, isn’t it? And I’m sure that’s a big part of why the BAILII exists as well because they knew people couldn’t access LexisNexis.

(Interview excerpt with Researcher 4, via Skype from Murcia, Spain)

The experience of Researcher 6 in this study regarding attempts to carry out DDL research with language teachers and learners in China highlights another aspect of
the access concept as it intercepts with the dominant themes for FLAX, students and teachers within the qualitative dataset. Her greatest challenges were with securing access to research sites with students and teachers in China to test out the efficacy of the Learning Collocations collection in the FLAX system. She and Researcher 1, both of whom come from the field of education, discussed the role of use or user studies—prevalent concepts within the data—with tools and projects like FLAX that stem from computer science as they are applied to the students theme for educational researchers:

Researcher 1: They talk a lot about user studies in computer science, don’t they?
Researcher 6: Yeah, but those user studies are only to prove that the tool works.
Researcher 1: Right, the focus is not to prove that learning has occurred with use of the tool.
Researcher 6: No, the purpose of such user studies in computer science is not to promote the application of the tool. So, for them the end of their project is that the tool has been developed successfully but for English teachers with English language learning tools, that is the beginning. But between the end of computer scientists completing the development of a learning tool and the beginning of English language teachers adopting a learning tool in their teaching there is a gap.

(Interview excerpt with Researcher 6, University of Waikato NZ)

The importance of user studies in this design-based research leads into our final section of analysis on the data collected with knowledge users, EAP teachers and managers at two UK universities, Durham and Queen Mary.

6.4.3 Barriers to Remixing Texts in Data-Driven EAP Materials Development

Collaborative work with Durham and Queen Mary revealed that data-driven approaches are not embedded within materials development and classroom teaching practices at these two UK universities. However, online corpus-based resources have a valued place as supplementary EAP materials. Most DDL tools and corpus-based systems were viewed by the majority of participants at Durham and QMUL as stand-alone web-based reference resources for students to explore outside of classroom teaching time.

Issues stemming from the design-based research carried out with Durham and QMUL include the limited amount of time EAP teachers have in the classroom with students to focus on discrete language items and the infeasibility of shepherding large groups of students in developing and mining personalised domain-specific corpora for focused help with dissertation and thesis writing, for example. This is despite some promising findings from research into DDL approaches with smaller, more tailored EAP classes for building Do-It-Yourself digital corpora with students to help with PhD thesis writing (Charles, 2012, 2015).
The focus-group discussions with managers at QMUL on the increased availability of open access content point to what EAP practitioners are now able to do with academic things, resources and materials for use/using with students as they emerge in this sub-dataset for the top four themes related to knowledge users: EAP, students, things and people. Knowledge User 2, manager of multimedia language support at QMUL, describes the approach of developing transferable skills in EAP materials development with revising and repurposing open access research publications as being one that is closer to traditional approaches with the re-use of authentic language content for classroom teaching purposes:

Knowledge User 2: You know, I think the thing about open educational resources, the question here, or part of the question here, which we discovered in this project, for example, is if you take a text, a raw text, which is not adapted for teaching like an article, it has EAP potential because it’s an authentic academic article. Then the ability to use that and to put it into materials, or adapt it, modify it, or change it under the Creative Commons thing is the revelation. Because we’ve all been doing it for years anyway, from copying it from a book or something when we’ve not supposed to have been adapting it, changing it, or whatever.

(Knowledge User 2, focus-group discussion excerpt, Queen Mary University of London UK)

From the same focus discussion, the pre-sessional course director at QMUL, Knowledge User 3, talks about the barriers to people working in universities from openly sharing EAP materials across institutions and how they are tied to each university’s business model with the aim of promoting their particular brand of EAP courses and materials as a unique selling feature. He also discusses the rise in influence of commercially produced EAP publications and the re-use of third-party materials from these publications as seeping into university EAP course materials development practices, which in turn creates a further barrier to sharing due to copyright infringement:

Knowledge User 3: There is a certain degree of openness but there is also this desire for everything to be branded, and a certain amount of clutching to your chest, especially about pre-sessional materials. [...] This is Queen Mary material, this is Southampton material, this is Durham material. But I think when you get back to the institutional level, those are where the real barriers lie because people are, and that comes down to the cut n paste culture that means a lot of third-party materials end up in our materials and are branded as being in-house but a lot of them are not really. You know, the ideas come from published materials and they’re probably not properly acknowledged anyway because they’re only being used internally. And part of that barrier to sharing more openly is raising an awareness of our existing practices and this means they don’t want to share between institutions because they’re worried that people will see just how much cut n paste is going into those materials. And I think the loser is the student, you know, because if people were really producing and sharing the best that they could amongst institutions to then create the best EAP pre-sessionals then the students would obviously benefit.

(Knowledge User 3, focus-group discussion excerpt, Queen Mary University of London UK)

From a meeting with QMUL EAP teacher, Knowledge User 1, the concepts of open and access, which congregate in the people theme, relate to frequent references in the data of how people outside the university can also benefit from education and resources that are openly accessible via the Internet:
Knowledge User 1: This open-source software and open access approach to data-driven learning resources does threaten current business models in EAP provision, doesn’t it? This idea of yours to re-use the artefacts of the academy. This really bucks some people in academia.

Researcher 1: Tell me more about that because that’s what I think is important to be doing in higher education, but I realise that this isn’t everyone’s priority.

Knowledge User 1: That’s what I think is important as well. It’s the ivory tower, isn’t it? It’s the secret garden behind the firewall of the ivory tower.

Knowledge User 1: Now, yes, I need people within this higher education environment [Queen Mary] to re-use these academic texts but I also need people to come into this FLAX environment, people who need to interface with this environment for whatever academic English need they have, and that’s what FLAX does for them in a manageable way. It makes it accessible not only to people who are using it in situ within the privileged brick-n-mortar of the academy but for people who, like I say, need to interface with that in some way outside of the academy, and, oh, that matters. The resource is not just locked inside our intranet-based VLE [Virtual Learning Environment] where I have developed learning resources with links out to FLAX on the web, which is really a Mickey Mouse version of FLAX in here.

(Meeting excerpt with Knowledge User 1, Cutty Sark pub in Greenwich, London, UK)

6.4.4 A Crisis in EAP Identity

An emerging tension in formal EAP is the issue of EAP practitioner identity in the neoliberal university (Ding & Bruce, 2017; Hadley, 2015; Hyland, 2002). Where are EAP service units placed in universities, and more importantly, how are they received and perceived by the wider academy? At its best, EAP is viewed as drawing on and contributing to a rich knowledge base from research in systemic functional linguistics, genre theory, corpus linguistics, academic literacies and critical EAP (Ding & Bruce, 2017). At its worst, EAP has been conceived as having “accepted the role as an economic and intellectual short-cut… [with] maximum throughput of students with minimum attainment levels in the language in the shortest possible time”. (Turner, 2004, pp. 96–97).

There has been an upswing in commercially produced EAP publications with a notable shift in focus towards generic academic skills and processes. The increasing prominence of generic EAP publications can be seen to exacerbate the growing fissure in EAP practitioner identity with the emergence of two opposing camps: English for General Academic Purposes (EGAP) versus English for Specific Academic Purposes (ESAP). The received definitions and understandings from the literature indicate that EAP is a subset of English for Specific Purposes (ESP) (see ETIC, 1975; Widdowson, 1983; Swales, 1985; Flowerdew & Peacock, 2001; Howatt, 2004; Belcher, 2010; Charles & Pecorari, 2016; Anthony, 2018). However, this understanding of EAP as being concerned with the teaching and learning of domain-specific language appears to have become conflated and confused as the popularity of generic skills-based
EAP textbooks, subscription-based supplementary online resources and programmes continues to rise (Gillett, 2018).

The absence of data-driven approaches in the design of EAP classroom teaching and online materials is a recurring theme in the sub-dataset from knowledge users. In a focus-group discussion with former teaching colleagues at Durham (Knowledge User 5 and Knowledge User 6), reflections turned towards collaborative work that involved trialling corpora and data-driven approaches for EAP (Fitzgerald, 2013). The discussion drew comparisons between the explicit focus on the teaching and learning of domain-specific language against a growing perception that the culture and practice of EAP are moving away from a focus on language towards generic skills, and the implications that this shift in focus might have for teachers and students:

Knowledge User 6: I think one major issue with EAP is that it has become so un-language focused. It’s moved so far away from teaching language. And, students, of course, can’t understand this because that’s what they think they’re paying for. They think we’re there to teach them the English. I think I’m there to teach them the English but the powers that be think that we’re there to teach them EAP.

Researcher 1: I mean we didn’t do any, there was no time in the timetables for language, right?

Knowledge User 6: No, for language, nothing. It’s all just skills.

Knowledge User 5: I couldn’t believe it when I started teaching EAP.

Knowledge User 6: Skills and process. And this is so deeply concerning when they don’t have the language to express their ideas.

Knowledge User 5: I think that’s why when they started this redundancy thing, oh well, I didn’t fight it because I’m not teaching language in EAP and I enjoy teaching language.

(Focus-group discussion excerpt with Knowledge User 5 & Knowledge User 6, Café Nero, Durham UK)

Corpora provide teachers and learners with access to linguistic data that show how language is used across a variety of real-world communication contexts. There have been many successful commercial language coursebook publications that are informed by corpora. However, many more coursebook publications appear to fly in the face of evidence-based approaches to materials writing for meeting the demands of an English language education content industry that seems to be driven, first and foremost, by market research rather than research into whether or not materials have positively influenced teaching, learning and language acquisition. A meeting with EAP teacher, Knowledge User 1, highlights some of the issues with EAP materials writing with commercial publishers. Despite materials not always drawing on evidence of how language actually works, they are still widely marketed for sales distribution and consumption:

Knowledge User 1: What I saw with him [EAP materials writer with Oxford University Press] was, with his presentation at IA TEFL [International Association for Teaching English as a Foreign Language] was, that it was no more or less like really saying that THESE materials he is selling are THE exponents that we need to teach students. And it was still
very much along the lines of we need to teach them yet more fixed phrases. And I was like sitting there and thinking some yes, some no, but prove it. I can—Can you? And he was putting up his examples, and I had my tablet open using FLAX, and I was going that example of his works, and that works, that doesn’t work, that works, that doesn’t work. But he’s just basing it on his own judgement. And I’m just sitting there testing. Just right in front of him, testing his materials.

Researcher 1: And, you would have thought that he would have tested his examples with a corpus-informed approach before presenting them at IATEFL let alone publishing them with OUP. You have to wonder where the quality control lies if at all.

[...]

Knowledge User 1: The vast majority of my colleagues at Queen Mary have been pretty open-minded, and they’ve been looking at FLAX and they can see that it’s real academic language data. It’s the authenticity of it.

Researcher 1: Yes, that always wins out, doesn’t it?

Knowledge User 1: Of course, it does but first of all they need to know that these non-commercial data-driven systems exist and that’s where the commercial publishers have the upper hand.

( Meeting excerpt with Knowledge User 1, Cutty Sark pub in Greenwich, London UK)

6.5 Conclusion

With initiatives in open access and the changes to copyright legislation that have brought about TDM limitations and exceptions, we have seen the greatest distance travelled with this design-based research, resulting in the co-creation of the following language learning collections that remix open access content for learning features of academic English: the largest English language collocations collections used by learners online (Wu et al., 2021), the full-text BAWE collections in collaboration with EAP teachers at Durham University, the EThOS PhD abstract corpora with participating EAP practitioners from Queen Mary University of London, the legal English BLaRC collection by Dr Maria Jose Marín from the University of Murcia, and the Academic Collocations in English (ACE) corpora with the COnnecting REpositories (CORE) aggregation and Application Programming Interface (API) services at the UK Open University. There is a growing sense that knowledge organisations such as the British Library and the Oxford Text Archive and aggregation and API services such as CORE are interested in non-commercial educational re-use applications of open access content that are aligned with the Budapest Open Access Initiative. Indeed, by far the biggest impact of openness in the higher education sector has been with open access, showing the importance of knowledge organisations in promoting accessible and reusable research (Finch Group, 2012).

The research presented on remixing MOOC content with TDM approaches provides proof of concept for the importance of licensing MOOC content openly for much-needed data-driven support with domain-specific language in non-formal
education that has re-use value in formal EAP education (Fitzgerald et al., 2017). This increased value from open language learning online is echoed by the upswing in the enrolment in language MOOCs that have emerged during the pandemic (Martín-Monje & Borthwick, 2021). However, findings from our research point to a current problem with the scalability of developing derivative OERs from MOOC content, with the example presented here of providing data-driven language support in the MOOC context. This problem is apparent in current mainstream MOOC provision where current business models do not anticipate a need for the open licensing of course content, and where open educational practices are mostly limited to those subject academics and learning technologists who were already open digital scholars before engaging in MOOC and networked learning pedagogy. Rather, current MOOC business models appear to focus on charging learners for increased access to learning content. This phenomenon has been presented here as an issue that open education policy makers, in collaboration with Creative Commons, are actively lobbying to address. As a work-around solution for embedding the functions and open corpora of FLAX directly into a MOOC platform interface, research is currently being carried out by Dr Jemma König at the University of Waikato with the development of F-Lingo, a Chrome extension. F-Lingo works on top of the FutureLearn platform to support content-based learning of domain-specific terminology and concepts for academic and professional English. Nonetheless, this work with F-Lingo would still require higher education institutions to allow the traversing and re-use of All Rights Reserved course content for the R&D of automated language learning support in the MOOC context (Fitzgerald et al., 2019; König et al., 2022).

The observed absence of data-driven approaches to support EAP provision at two UK universities, and the apparent shift away from language teaching, as noted in focus-group discussions with teachers and managers, give pause for understanding current practices with EAP materials development for classroom and online learning in a time of increased uptake of generic EAP course books from commercial publishers. By drawing attention to the underlying business models and cultural practices that higher education institutions and organisations adopt, we also arrive at a closer understanding of the values placed on research, or lack thereof, with online and classroom materials development and teaching in the field of EAP.

The new paradigm for open data-driven language learning systems design presented through this research has also argued for greater access to and re-use of the artefacts of the academy and professional domains such as law, for example, that are taught and studied at higher education institutions. In this chapter, we have demonstrated the perceived value that corpus linguistics researchers and knowledge users working within EAP place on pedagogic, professional and research texts that can be mined for aspects of domain-specific language with data-driven learning systems. In addition to the value placed on open educational practices that can be fostered to re-use, remix and redistribute EAP resources for uptake across formal and non-formal higher education in increasingly uncertain times.

Acknowledgements We would like to thank the many contributors and collaborators of this ongoing R&D project over the years, including the Fonds de recherche du Québec—Société et
Appendix

Open Collections in FLAX: Content and Collaborators

**Learning collocations system in FLAX (2009–2022)**

**Content**
- Wikipedia corpus of contemporary English derived from three million Wikipedia articles comprising three billion words (Wu & Witten, 2016; Wu et al., 2021)
- British National Corpus (BNC) of 100 million words (BNC Consortium, 2007)
- British Academic Written English (BAWE) corpus of 2500 pieces of assessed university student writing from across the disciplines
- Academic Collocations in English (ACE) corpora of harvested open access content and metadata from 135 million articles residing in open journals and open repositories

**Knowledge organisations**
- Wikimedia Foundation (Wikipedia corpus); Oxford Text Archive and the UK Higher Education Academy OER International Programme with the University of Oxford (BNC and BAWE corpora); CORE (COnnecting REpositories)\(^\text{10}\) team, UK Open University (ACE corpora)

**Researchers**
- FLAX team

**Knowledge users**
- Waikato University computer science students; Durham University EAP teachers and students; University of Oxford OER International stakeholders

**British Academic Written English (BAWE) collections in FLAX (2012)**

**Content**
- Full texts of the BAWE corpus divided into four sub-collections: Arts & Humanities, Social Sciences, Life Sciences, Physical Sciences

**Knowledge organisations**
- The Oxford Text Archive; UK Higher Education Academy

**Researchers**
- FLAX team

**Knowledge users**
- Durham University EAP teachers and students; University of Oxford OER International stakeholders


(continued)

\(^\text{10}\) [https://core.ac.uk/about#mission](https://core.ac.uk/about#mission).
### MOOC/micro-networked course collections in FLAX (2014–2016)

**Content**
- 8.85 million-word corpus of full-text judicial hearings derived from free legal sources at the British and Irish Legal Information Institute (BAILII) aggregation website

**Knowledge organisations**
- BAILII

**Researchers**
- Universidad Murcia; FLAX team

**Knowledge users**
- Law MOOC learners

### PhD micro-abstract corpora with FLAX mobile activities (2014–2015)

**Content**
- MOOC / Micro-Networked Course lecture transcripts and videos (streamed via YouTube or Vimeo) and case law that reside in the public domain

**Knowledge organisations**
- MOOC host institutions (Harvard University; University of London; Columbia University) with edX and Coursera MOOC providers

**Researchers**
- FLAX team; Universidad Murcia

**Knowledge users**
- MOOC learners and MOOC subject matter experts; legal English translation studies teachers, and students at the University of Murcia

### PhD abstract corpora in FLAX (2015–2016)

**Content**
- 9.8 million-word corpus derived from the metadata, including the abstracts, of over 500,000 PhD theses awarded by UK universities and managed by the Electronic Thesis Online Service (EThOS) at the British Library

**Knowledge organisations**
- British Library Labs and EThOS at the British Library

**Researchers**
- FLAX team

**Knowledge users**
- EAP teachers and learners at Queen Mary University of London

### Academic Collocations in English (ACE) collections in FLAX (2018–2022)

**Content**
- Harvested open access content from open journals and open repositories divided into four sub-collections: Arts & Humanities, Social Sciences, Life Sciences, Physical Sciences

---


12 [https://www.bl.uk/projects/british-library-labs](https://www.bl.uk/projects/british-library-labs).

13 [http://ethos.bl.uk/Home.do](http://ethos.bl.uk/Home.do).
(continued)

<table>
<thead>
<tr>
<th>Knowledge organisations</th>
<th>CORE (COnnecting REpositories) team, UK Open University</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researchers</td>
<td>FLAX and F-Lingo teams</td>
</tr>
</tbody>
</table>
| Knowledge users          | • User query data analysis research with the FLAX LC system learners worldwide  
                           | • Research with MOOC learners via F-Lingo Chrome extension and FutureLearn platform |

References


COnnectedREpositories (CORE): Aggregating the world’s open access papers. (n.d.). Retrieved from https://core.ac.uk/


Fitzgerald, A. (2013). Openness in English for academic purposes. Open Educational Resources Case Study: Pedagogical development from OER practice. Commissioned by the Higher Education Academy (HEA) and the Joint Information Systems Committee (JISC), United Kingdom, 20 pages.


Alannah Fitzgerald is a postdoctoral research fellow with the Computer Science Department at the University of Waikato in Aotearoa, New Zealand, and an honorary fellow with the School of Education at Durham University in the UK. Alannah is responsible for designing open education applications with the F-Lingo and FLAX language projects. With Dr. Wu, Professor Witten and Chris Mansfield, she was awarded a prize in the British Library Labs Competition Teaching and Learning category for re-using digital collections in language education with funding from the Andrew W. Mellon Foundation. Her research interests include open educational resources and practices for designing and developing digital domain-specific language collections (corpora) and for devising and delivering online English language learning interventions that can be scaled and assessed across both formal (classroom-based) and non-formal (MOOC space) higher education contexts.

Shaoqun Wu is Senior Lecturer in Computer Science at the University of Waikato in Aotearoa, New Zealand and is the main developer of the FLAX language project. With Dr. Fitzgerald and Professor Witten, Dr. Wu was awarded first prize in the LinkedUp Vici Challenge for mature open data-driven applications for education by Open Knowledge International with funding from the European Commission. Her research interests include computer-assisted language learning, mobile language learning, supporting language learning in MOOCs, digital libraries, natural language processing and computer science education.
**Jemma König** is Postdoctoral Fellow in the Department of Computer Science at the University of Waikato in Aotearoa, New Zealand and is responsible for developing the F-Lingo Chrome extension for FutureLearn MOOCs. Jemma’s PhD research explored a computational approach to vocabulary testing, language tools and text enrichment. More specifically, focusing on corpus analysis, pseudoword generation, automated vocabulary testing and tracking learners’ interaction with online written language. With Dr. Fitzgerald and Professor Witten, Dr. König was awarded best paper for her work with F-Lingo at the Learning with MOOCs conference in Milwaukee, USA, by the Institute of Electrical and Electronics Engineers (IEEE) Education Society in 2019.

**Steven Shaw** is a professor in the Department of Education, Concordia University, in Montreal. His research and professional work focus on the design, development and implementation of technology to support learning and knowledge sharing, particularly at the enterprise scale in large public and private sector organisations. He co-founded the corporation that developed the first “learning content management system”. He served as the CLO of Eedo Knowledgeware, which for over a decade furnished the leading-edge technology for learning content management, employed by Fortune 500 organisations such as Xerox, Dell, Eli Lilly, Boeing and the largest public sector organisations in the US and UK, including US Treasury and Department of Energy and Foreign and Commonwealth Office and Department of Work and Pensions in the UK. His areas of expertise include software development, systems implementation, content management, taxonomy development and the design and evaluation of training programs and curricula in professional education.

**Ian H. Witten** is Emeritus Professor of Computer Science at the University of Waikato in Aotearoa, New Zealand, with a research career that spans over 40 years. His best-known publication is the book, *Data mining: Practical machine learning tools and techniques*, now in its fourth edition (2016). Professor Witten is also well known for his award-winning open-source software, sharing his advances with thousands of students, teachers and users around the world. These include Greenstone, a digital library platform on which the FLAX system operates. Another successful open-source software is Weka (Waikato Environment for Knowledge Analysis), a data-mining tool. Weka is probably the world’s most widely used machine learning workbench, and in 2017 Professor Witten led three popular Massive Open Online Courses (MOOCs) with Future-Learn: *Data mining with Weka*, *More data mining*, and *Advanced data mining*. In 2017, Professor Witten was awarded an Honorary PhD (Doctor of the University) from the Open University in the United Kingdom for his lifetime contribution to furthering the advancement of research and education.
## Author Queries

**Chapter 6**

<table>
<thead>
<tr>
<th>Query Refs.</th>
<th>Details Required</th>
<th>Author’s response</th>
</tr>
</thead>
<tbody>
<tr>
<td>AQ1</td>
<td>Please note that the footnote has been set in the following sentence “It identifies the knowledge organisations,…”, as footnotes are not allowed in abstracts.</td>
<td></td>
</tr>
<tr>
<td>AQ2</td>
<td>References “Burkhardt (2006), Weber (1990), Fisher (2014), BNC Consortium (2007)” are cited in the text but not provided in the reference list. Please provide the respective references in the list or delete these citations.</td>
<td></td>
</tr>
</tbody>
</table>