

A New Paradigm for Open Data-Driven Language Learning Systems Design  
in Higher Education

Alannah Fitzgerald

A Thesis  
In  
The Department  
Of  
Education

Presented in Partial Fulfilment of the Requirements  
For the Degree of  
Doctor of Philosophy (Educational Technology) at  
Concordia University  
Montreal, Quebec, Canada

November 2018

**CONCORDIA UNIVERSITY**  
**SCHOOL OF GRADUATE STUDIES**

This is to certify that the thesis prepared

By: Alannah Teresa Dysart Fitzgerald

Entitled: A new paradigm for data-driven language learning systems design in  
higher education

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Educational Technology)

complies with the regulations of the University and meets the accepted standards with respect to  
originality and quality.

Signed by the final examining committee:

\_\_\_\_\_  
Dr. Pavel Trofimovich Chair

\_\_\_\_\_  
Dr. Pascual Perez-Paredes External Examiner

\_\_\_\_\_  
Dr. Miranda D'Amico External to Program

\_\_\_\_\_  
Dr. Richard Schmid Examiner

\_\_\_\_\_  
Dr. Vivek Venkatesh Examiner

\_\_\_\_\_  
Dr. Steven Shaw Thesis Supervisor

Approved by

Dr. Steven Shaw, Graduate Program Director

January 10, 2019

\_\_\_\_\_  
Dr. André Roy, Dean  
Faculty of Arts and Science

## **ABSTRACT**

### **A new paradigm for open data-driven language learning systems design in higher education**

**Alannah Teresa Dysart Fitzgerald, Ph.D.**

**Concordia University, 2019**

This doctoral thesis presents three studies in collaboration with the open source FLAX project (Flexible Language Acquisition [flax.nzdl.org](http://flax.nzdl.org)). This research makes an original contribution to the fields of language education and educational technology by mobilising knowledge from computer science, corpus linguistics and open education, and proposes a new paradigm for open data-driven language learning systems design in higher education. Furthermore, the research presented in this thesis uncovers and engages with an infrastructure of open educational practices (OEP) that push at the parameters of policy for the reuse of open access research and pedagogic content in the design, development, distribution, adoption and evaluation of data-driven language learning systems.

Study 1 employs automated content analysis to mine the concept of open educational systems and practices from qualitative reflections spanning 2012-2019 with stakeholders from an ongoing multi-site design-based research study with the FLAX project. Design considerations are presented for remixing domain-specific open access content for academic English language provision across formal and non-formal higher education contexts. Primary stakeholders in this ongoing research collaboration include the following: knowledge organisations – libraries and archives including the British Library and the Oxford Text Archive, universities in collaboration with Massive Open Online Course (MOOC) providers; an interdisciplinary team of researchers; and knowledge users in formal higher education – English for Academic Purposes (EAP) practitioners. Themes arising from the qualitative dataset point to affordances as well as barriers with the adoption of open policies and practices for remixing open access content for data-driven language learning applications in higher education against the backdrop of different business models and cultural practices present within participating knowledge organisations.

Study 2 presents a data-driven experiment in non-formal higher education by triangulating user query system log data with learner participant data from surveys (N=174) on the interface designs and usability of an automated open source digital library scheme, FLAX. Text and data mining approaches (TDM) common to natural language processing (NLP) were applied to pedagogical English language corpora, derived from the content of two MOOCs, (Harvard

University with edX, and the University of London with Coursera), and one networked course (Harvard Law School with the Berkman Klein Center for Internet and Society), which were then linked to external open resources (e.g. Wikipedia, the FLAX Learning Collocations system, WordNet), so that learners could employ the information discovery techniques (e.g. searching and browsing) that they have become accustomed to using through search engines (e.g. Google, Bing) for discovering and learning the domain-specific language features of their interests. Findings indicate a positive user experience with interfaces that include advanced affordances for course content browse, search and retrieval that transcend the MOOC platform and Learning Management System (LMS) standard. Further survey questions derived from an open education research bank from the Hewlett Foundation are reused in this study and presented against a larger dataset from the Hewlett Foundation (N=1921) on motivations for the uptake of open educational resources.

Study 3 presents a data-driven experiment in formal higher education from the legal English field to measure quantitatively the usefulness and effectiveness of employing the open Law Collections in FLAX in the teaching of legal English at the University of Murcia in Spain. Informants were divided into an experimental and a control group and were asked to write an essay on a given set of legal English topics, defined by the subject instructor as part of their final assessment. The experimental group only consulted the FLAX English Common Law MOOC collection as the single source of information to draft their essays, and the control group used any information source available from the Internet to draft their essays. Findings from an analysis of the two learner corpora of essays indicate that members of the experimental group appear to have acquired the specialized terminology of the area better than those in the control group, as attested by the higher term average obtained by the texts in the FLAX-based corpus (56.5) as opposed to the non-FLAX-based text collection, at 13.73 points below.

## **Acknowledgements**

This doctoral thesis research was made possible due to a number of important people, projects, and opportunities that I encountered on what has turned out to be a most humbling, rich and rewarding PhD journey.

With special thanks to my doctoral supervisor, Steven Shaw, for encouraging me to take a global focus with my work and for supporting my collaboration with the open source data-driven Flexible Language Acquisition project ([flax.nzdl.org](http://flax.nzdl.org)) at Waikato in Aotearoa/New Zealand, in addition to supporting numerous successful applications for funding and for fellowships at leading UK higher education institutions (the Open University, Oxford, Durham). A good supervisor must be an advocate for the postgraduate student through the increasingly bureaucratised doctoral candidature – I sincerely thank Steven for his sensitivity and his support in navigating my way through unforeseen delays with ethics clearance and the extended knock-on effects with having to set up new research sites to collect data. More importantly, I thank Steven for fighting for extra time for me to care for my mother in the final years of her life before submitting this thesis for examination. I also thank my doctoral committee members, Richard Schmid and Vivek Venkatesh, and Graduate Program Coordinator, Nadine Wright, for supporting Steven throughout his supervision of my doctoral candidature in ensuring that I complete my PhD. Also, with grateful appreciation to the Fonds de recherche du Québec – Société et culture (FRQSC) for partly funding my doctoral research.

I thank the serendipitous moment in time and space that led me and Shaoqun Wu of the FLAX project to present at the same eLearning conference event in Villach, Austria, leading to my on-going collaboration with her at the Greenstone Digital Library Lab at Waikato's Computer Science Department. My collaboration with Shaoqun has proven to be the longest and most fruitful professional working relationship of my career, and I thank her and her family for their generosity in hosting me at their home during my research visits. I also thank my current postdoctoral supervisor and FLAX project lead, Ian Witten, for his support in encouraging me to scope out open content for the co-design and co-creation of digital collections for learning academic English with the FLAX project, and to carry out educational research with various stakeholders around the world working in formal and non-formal higher education. With admiration for Ian's successful vision and leadership in developing and disseminating the award-

winning Greenstone digital library open-source software, upon which the FLAX system is built, for accelerating the use of information technology for social and economic benefit in the global south and the global north where value is placed on facility with technology and the English language. I thank Li Liang for her support and her willing participation to reflect on her own doctoral work with the FLAX project at Waikato under the supervision of Margaret Franken and Shaoqun Wu in parallel to my own doctoral work. With appreciation and admiration for Jemma König's PhD work, also at Waikato's Department of Computer Science, in developing the F-Lingo software and taking the vision of data-driven domain-specific terminology learning support directly to the MOOC platform space. It is a wonderful opportunity to be able to collaborate with Jemma on research into the uptake of F-Lingo in MOOCs through my current postdoctoral fellowship work with Waikato.

With thanks to my open-minded and supportive former English for Academic Purposes (EAP) colleagues at Durham University – Steve Kirk, Louise Greener, Terri Edwards, Jeff Davidson, Clare Carr and Leslie Kendall – for getting behind and participating in my first OER academic practice fellowship managed by the Open University (OU) at Durham University English Language Centre. Thanks also to the numerous inspiring individuals and project teams whom I met and worked with at the OU for my fellowships with the Support Centre in Open Resources for Education (SCORE), the OER Research Hub, and the Global OER Graduate Network (formerly with the OU Netherlands).

I am grateful to my fellow SCORE fellow, Melissa Highton, for her continued interest in my work with the FLAX project. For her and Ylva Berglund-Prytz's recognition of the reuse value of the British National Corpus in the FLAX system for making the corpus more accessible for learning collocations, and for Melissa's invitation to open up more English language corpora managed by the Oxford Text Archive at the University of Oxford's IT Services (formerly Oxford Computing Services) where she worked as Director of Academic IT. This invitation led to the development of the British Academic Written English (BAWE) collections in FLAX, and a further UK OER International fellowship with Oxford and the UK Higher Education Academy for promoting open educational practices internationally with Oxford-managed corpora and Oxford-created OERs. My UK OER international fellowship was a real highlight in my career for better understanding the social and economic implications of open educational practices in seven different countries from across three continents.

With thanks to EAP teachers and managers at Queen Mary University of London - Chris Mansfield, Martin Barge, William Tweddle and Saima Sherazi - for their enthusiastic engagement with developing data-driven collections for EAP teaching and learning with digital open access content from the British Library that resulted in the PhD Abstracts collections in FLAX. Thanks also to Sara Gould of the Electronic Thesis Online Services (EThOS) at the British Library, and Mahendra Mahey of British Library Labs for their championship in the area of open access, for their encouragement in creating learning and teaching derivatives from digital collections curated by the British Library for reuse by third parties for not-for-profit purposes. With thanks also for their recommendation to carry out further text and data mining work with one of the British Library's research partners, CORE (COConnected REpositories), at the OU's Knowledge Media Institute for the reuse of aggregated open access data, metadata and APIs in the development of the new Academic Collocations in English (ACE) collections in FLAX.

Many thanks to the two wonderful Marias at Universidad de Murcia, Maria Jose Marín and María Angeles Orts, for their passion and insightful expertise in legal English and corpus linguistics, and for their willingness to collaborate on research at a distance with their students in legal English language education and translation studies. Thanks also for the addition of Maria Jose Marín's British Law Reports Corpus (BLaRC) in FLAX with a focus on making the corpus accessible for learning and teaching purposes with this special variety of English.

With special thanks also to Kathi Bailey and Ryan Damerow of The International Research Foundation (TIRF) for English Language Education – for their support of doctoral dissertation grant holders beyond the call of duty and their dedication to making our research more meaningful and accessible to practitioners working in English language education the world over. Thanks also to TESOL International for their support in funding the dissemination of my graduate research through the Ruth Crymes fellowship.

Thanks to the OER champions, Pat Lockley behind the English Common Law MOOC with the University of London, and Professor Fisher behind CopyrightX at Harvard Law School, for their commitment to OER and for their encouragement with the reuse of their course content in developing language support derivatives with the FLAX project. I learned a great deal about copyright law as it applies to reuse in education from the CopyrightX course, which has been of direct use in this thesis.

Many thanks to the international ESOL placements through my part-time work as an examiner with Trinity College London, which has indirectly helped to fund this doctoral research, and which has maintained my connection with English language assessment, curricula, teaching and learning in schools and educational institutions throughout the world.

Special thanks to the loving support of my step-father, Michael Dawson, and my sisters, Joan Fitzgerald and Catherine Fitzgerald, throughout this adventurous PhD journey. Thanks to Catherine who was also doing a PhD at the same time as me and who put me onto the amazing doctoral studies guru, Tara Brabazon, and her invaluable digital scholarship for sharing insights into the PhD experience.

With appreciation for the loving memory of my late father, Kevin Fitzgerald, who first introduced me to the UK Open University when he took me to see the film *Educating Rita* in 1985. It took two years to reach our picture house in provincial-town Aotearoa/New Zealand, and I was just at that age – twelve going on thirteen – to appreciate this Pygmalion story of a woman breaking through the class barriers with an emancipatory distance education from the OU. My Dad also took me canvassing with him for the New Zealand Labour Party in those formative years, showing me first-hand that life for those in state-housing areas was very different from life in homes belonging to those who had been to university.

With special gratitude to my dear friend, Katherine McAlpine, who went well beyond the call of duty to support me in person while caring for my mother and helping me through my mother's memorial service process. With fond gratitude to friends and extended family who generously hosted me and supported me on my numerous research travels between Canada, New Zealand, the UK, India and South Korea, making this PhD possible: Ann Clark, Sarah Clark, Andy Hamilton, David Butler, Douglas MacDougall, Damian Rumph, Biliana (Bibi) Velkova, Kathleen Lehan, Tara Rousseau, Yazmin Shroff, Katherine McAlpine, Mark Zeman, Iain Sands, Teresa Cowie, Robert Weinkove, Fraser Barrett, Benoit Aboumrad, Marie Aboumrad, Marc-André Dupras, Marie-Soleil Thellen, Anne Goldenberg, Simone Geday, Nathalie Rathle, Richard Hinch, Aidan Hinch, Guilaine Royer, Francis Brisebois, Ioana Contu, Aurélie Wales-Gaudreau, Danny Gaudreau, Jean-Sébastien Senécal (Tats), Patrice D'Amours, Laura Noël, Michèle Martin, Bastien Partensky, Manuelle Langlois (Manue), Kirsten Cameron, Alexandre Bustros, Shakib Ahsan, Vinay Suján, Christine Aitken and Jongsam Lee. Many thanks also to my Master's thesis supervisor, Richard Fay, at the University of Manchester, and for the support of yet more old



friends from my English language teaching days in Korea, who have shared discussions with me about my doctoral work over the years: Elana Wright, Christina Olivieri, Kimura Byol (Mihee Nathalie Cho), Michelle Wouters, Cheri-Ann Nakamaru, Sean Duke, Heather Evans and Nancy Hayne. I thank Dave Laroche-Gagnon for his help with formatting my survey data. A special thanks also to Karen Garry for providing a loving home and life for our jet-setting Korean ginger cat, Orangee, whom I left in Vancouver at the beginning of my PhD journey to the UK, and who is now 22 years old and obviously living on to see me through the completion of my doctorate!

With deep appreciation and gratitude to Karine Rathle for her inspiration as a dance educator and dance scientist, and for her ability to live the artist's way by staying true to herself and her dreams. This PhD has been with us the whole time we have known each other. Thanks to her love, her grace, her support, and her understanding in encouraging each of us to pursue careers that inspire us.

This thesis is dedicated to my mother, Mary Cowie, whose empathic heart, love of life, ability to positively transcend the seemingly impossible, and whose strength of conviction that it is important to create a better more joyful world, have been the guiding inspiration for my life and for my work in education. To my mother's connection to the past and her stories of the teachers in our family, including my late father, Kevin Fitzgerald's, six sisters who became school teachers. To my mother's maternal grandfather, Kieran Hogan, who became a teacher later in life. Whose first language was Gaelic and who as a younger man became fluent in Te reo Māori in the 1860s and acted as a translator for local iwi to ensure they were not cheated in their transactions with English colonial authorities on the gold fields of Ross during the West Coast Gold Rush. To my mother's mother, Mary Hogan, who along with her two sisters lived and taught in rural schools populated with local Māori and Pakeha from the 1920s to the 1960s, who excelled in supporting all of her students through their reading, writing and arithmetic proficiency and school certificate exams, and who developed a reputation for helping students with special needs long before the terms dyslexia and dyscalculia were coined. To my mother's connection to the future – she will remain my twinkling north star as I continue on my lifelong educational journey.

## Contribution of Authors

The different chapters of this thesis are based on the following peer-reviewed publications, three of which are central to the studies described and discussed in chapters 3, 4, and 5, with supporting publications that show work leading up to and beyond the central texts in this manuscript-based thesis.

All of the central manuscripts included in this thesis for Studies 1, 2, and 3 have been co-authored with my colleagues at the University of Waikato in Aotearoa/New Zealand, Dr. Shaoqun Wu and Emeritus Professor Ian H. Witten of the FLAX project based at the Greenstone Digital Library Lab. For Study 2, collaboration has also included co-authorship with my PhD supervisor, Professor Steven Shaw, of Concordia University in Canada, and with doctoral candidate, Jemma König, of the F-Lingo project, which is also based at the Greenstone Digital Library Lab at the University of Waikato. In the case of Study 3, Dr. Maria Jose Marín from the University of Murcia in Spain is second author, and she has expanded the study in a further supporting publication with herself as first author, and Dr. María Angeles Orts, also of the University of Murcia, and myself as co-authors. Where the use of the first-person singular pronoun, *I*, is used in the manuscript-based chapters of the thesis this refers to my direct engagement with research contexts. Where the first-person plural pronoun, *we*, is used this refers to both mine and my co-authors' direct collaboration in developing, evaluating and discussing FLAX collections from the research. For the bridging material that introduces and expands on the manuscripts as they bear relevance to the overall coherence of this thesis, I resume use of the first-person singular pronoun, *I*.

As the lead author of the three publications presented in this thesis, I was responsible for scoping academic English language content with knowledge organisations (libraries and archives, open access data aggregation services, and universities in collaboration with MOOC providers) co-designing corpora with Dr. Wu and participating language researchers and teachers, designing research interventions and instruments, collecting participant data that has been reinforced with log data from the FLAX system, conducting analyses, and co-authoring manuscripts with my colleagues based in New Zealand and Spain.

## **Thesis manuscripts and supporting publications**

As per the directives of the editors working for the various publishers in preparing these manuscripts, the two journal articles and one book chapter central to chapters 3-5 of this thesis are very concise documents, one of which has been published and two of which have been submitted for publication. There are, however, obvious limitations to the manuscript-based thesis format. With particular reference to Study 3 a good deal of data and insights have not been reported because they did not fit the requirement for the publication in terms of focus and format. Nonetheless, reference to a more detailed publication which builds on Study 3 and reflects further research questions, methods, observations, conclusions and findings with my collaborators in Spain is discussed in the front matter preceding Chapter 5 of this thesis. In order to ensure that this thesis reads as a coherent whole, additional bridging material has been included herein with references to supporting research with the FLAX project that foreground each study and show the relationship between the central studies in this thesis. The main thrust of this thesis demonstrates the iterative nature of all three inter-related research studies and their implications for current as well as planned future research, in addition to implications for policy and pedagogy.

## **Study 1**

The central manuscript for Study 1 in Chapter 3 of this thesis presents collaborative work from 2012 to the present time on the various corpora that have been developed with key stakeholders as part of this doctoral research: knowledge organisations, researchers and knowledge users.

### ***Central manuscript for Study 1 in Chapter 3:***

Fitzgerald, A., Wu, S. & Witten, I.H. (Submitted). Reflections on remixing open access content for data-driven language learning systems design in higher education.

### ***Supporting papers for Study 1 in Chapter 3:***

Wu, S., Fitzgerald, A., Yu, A. & Witten, I.H. (2019). Developing and evaluating a learner-friendly collocation system with user query data. *International Journal of Computer-Assisted Language Learning and Teaching*, 9(2), 53–78.

Wu, S., Fitzgerald, A., Witten, I.H. & Yu, A. (2018). Automatically augmenting academic text for language learning: PhD abstract corpora with the British Library. In B. Zou, M. Thomas (Eds.), *Integrating Technology into Contemporary Language Learning and Teaching*, (pp. 512-537). IGI Global.

Fitzgerald, A., Wu, S., & Barge, M. (2014). Investigating an open methodology for designing domain-specific language collections. In S. Jager, L. Bradley, E. J. Meima, & S. Thouësny (Eds), *CALL Design: Principles and Practice*. In *Proceedings of the 2014 EUROCALL Conference* Groningen, The Netherlands, (pp. 88–95). Dublin: Research-publishing.net. doi:10.14705/rpnet.2014.000200.

Fitzgerald, A. (2013). *TOETOE International: FLAX Weaving with Oxford Open Educational Resources*. Open educational resources international case study with the University of Oxford. Commissioned by the Higher Education Academy (HEA) and the Joint Information Systems Committee (JISC), United Kingdom.

Fitzgerald, A. (2013). *Openness in English for Academic Purposes*. Open educational resources case study with Durham University: Pedagogical development from OER practice. Commissioned by the Higher Education Academy (HEA) and the Joint Information Systems Committee (JISC), United Kingdom.

## **Study 2**

The central manuscript for Study 2 in Chapter 4 of the thesis presents work with mining MOOC pedagogical content for designing, developing and evaluating data-driven domain-specific terminology learning support for non-formal learners enrolled on MOOCs with Coursera and edX, and with CopyrightX at Harvard Law School.

### ***Central manuscript for Study 2 in Chapter 4:***

Fitzgerald, A., Wu, S., König, J., Witten, I.H., & Shaw, S. (Submitted). Designing and evaluating an automated open data-driven language learning support system for MOOCs.

***Supporting paper for Study 2 in Chapter 4:***

Wu, S., Fitzgerald, A., & Witten, I. H. (2014). Second language learning in the context of MOOCs. In S. Zvacek, M. T. Restivo, J. Uhomoibhi, & M. Helfert (Eds.), *Proceedings of the 6th International Conference on Computer Supported Education* (pp. 354–359). Barcelona: Scitepress.

**Study 3**

The central manuscript for Study 3 in Chapter 5 of the thesis presents a corpus-based study on the reuse of MOOC content derived from course lectures and readings, and legal documents from the public domain for uptake in formal legal English language education and translation studies at the University of Murcia in Spain.

***Central manuscript for Study 3 in Chapter 5:***

Fitzgerald, A., Marín, M.J., Wu, S., & Witten, I. H. (2017). Evaluating the efficacy of the digital commons for scaling data-driven learning. In M. Carrier, R. Damerow, K. Bailey (Eds.), *Digital language learning and teaching: Research, theory and practice*. Global Research on Teaching and Learning English Series (pp. 38–51). New York: Routledge & TIRF.

***Supporting papers for Study 3 in Chapter 5:***

Marin, M.J., Ortis Llopaz, M. & Fitzgerald, A. (2017). A Data-driven learning experiment in the legal English classroom using the FLAX platform. *Miscelánea: A Journal of English and American Studies*, 55, 34–67.

Fitzgerald, A., Wu, S. & Marin, M.J. (2015). FLAX - Flexible and open corpus-based language collections development. In K. Borthwick, E. Corradini, A. Dickens (Eds.), *10 years of the Languages, Linguistics & Area Studies (LLAS) eLearning symposium: case studies in good practice* (pp. 215–227). Dublin: Research-publishing.net. doi:10.14705/rpnet.2015.000281.

## TABLE OF CONTENTS

LIST OF TABLES.....	XIX
LIST OF FIGURES .....	XXI
GLOSSARY.....	XXIV
ACRONYMS .....	XXIX
<b>CHAPTER 1: GENERAL INTRODUCTION .....</b>	<b>1</b>
A NEW PARADIGM FOR OPEN DATA-DRIVEN LANGUAGE LEARNING SYSTEMS DESIGN IN HIGHER EDUCATION .....	1
RESEARCH VISTAS.....	3
<i>Open Educational Resources (OER) research fellowships</i> .....	3
<i>The open source Flexible Language Acquisition (FLAX) language project</i> .....	5
TEXT AND DATA MINING (TDM) .....	7
<i>Opening the way for text and data mining</i> .....	8
<i>Mining modern language corpora for pedagogical purposes</i> .....	9
OVERARCHING REVIEW OF THE LITERATURE .....	10
<i>Developments with Data-Driven Learning (DDL) in language education</i> .....	10
<i>Developments with reusing artefacts of the academy in corpus development</i> .....	15
<i>Developments with openness in higher education</i> .....	19
<b>CHAPTER 2: RESEARCH METHODS .....</b>	<b>27</b>
KNOWLEDGE MOBILISATION .....	27
DESIGN ETHNOGRAPHY (DE).....	32
<i>A framework for design ethnography</i> .....	33
<i>Ethnographic toolkit</i> .....	34
<i>Engaging context</i> .....	36
<i>Moving in</i> .....	38
<i>Data gathering and analysing</i> .....	39
<i>Ethnography-4-Design (E4D)</i> .....	40
<i>Frameworking</i> .....	41
<i>Generating design concepts</i> .....	41
<i>Ethnography-2-Design (E2D)</i> .....	41

<i>Moving along</i> .....	42
<i>Prototyping</i> .....	43
<i>Artefacts</i> .....	43
<i>Moving out</i> .....	44
DESIGN-BASED RESEARCH (DBR).....	45
<i>Models for understanding and conducting DBR</i> .....	46
<i>Macro DBR cycle 1: Augmented full-text FLAX corpus design</i> .....	49
<i>Open gratis vs open libre</i> .....	55
<i>Macro DBR cycle 2: FLAX Learning Collocations system design</i> .....	57
OVERVIEW OF RESEARCH SITES AND ORIGINAL CONTRIBUTIONS TO KNOWLEDGE .....	63
INTRODUCTION TO STUDY 1 .....	66
<i>Open access as the content reuse default in higher education</i> .....	69
<b>CHAPTER 3: STUDY 1</b> .....	<b>70</b>
<b>REFLECTIONS ON REMIXING OPEN ACCESS CONTENT FOR DATA-DRIVEN LANGUAGE LEARNING SYSTEMS DESIGN IN HIGHER EDUCATION</b> .....	<b>70</b>
ABSTRACT .....	70
KEYWORDS .....	71
INTRODUCTION.....	71
RESEARCH MATERIALS .....	73
<i>Open data in Computer Assisted Language Learning (CALL)</i> .....	74
RESEARCH METHODS.....	74
<i>Design-Based Research in the context of Design Ethnography</i> .....	74
RESULTS AND ANALYSIS .....	75
<i>Automated Content Analysis (ACA)</i> .....	75
<i>Knowledge organisations</i> .....	77
<i>Researchers</i> .....	79
<i>Knowledge users</i> .....	81
DISCUSSION.....	84
<i>Knowledge organisations</i> .....	85
<i>Researchers</i> .....	90

<i>Knowledge users</i> .....	93
CONCLUSION .....	99
CONNECTING STUDY 1 TO STUDY 2.....	102
INTRODUCTION TO STUDY 2 .....	103
<i>The datafication of higher education</i> .....	103
<b>CHAPTER 4: STUDY 2</b> .....	<b>107</b>
<b>DESIGNING AND EVALUATING AN AUTOMATED OPEN DATA-DRIVEN LANGUAGE LEARNING SUPPORT SYSTEM FOR MOOCS</b> .....	<b>107</b>
ABSTRACT .....	107
KEYWORDS .....	108
INTRODUCTION.....	108
<i>The problem with learning support and the business model behind MOOCs</i> .....	108
<i>The problem with MOOC language barriers: The case for domain-specific terminology learning support</i> .....	110
<i>The problem with MOOC content browsability and searchability: Design principles for an augmented learning platform experience</i> .....	111
<i>Research questions</i> .....	114
<i>Research hypothesis</i> .....	114
METHODS .....	115
<i>Materials</i> .....	115
<i>Procedures</i> .....	117
RESULTS AND ANALYSES .....	118
<i>Demographic data statistics</i> .....	119
<i>Learning support resources used and techniques adopted by non-formal learners</i> .....	121
<i>Language resources used by non-formal learners</i> .....	121
<i>Motivations for using learning support</i> .....	123
<i>Motivations for using FLAX as learning support</i> .....	123
<i>User query pathway analysis</i> .....	124
<i>Automated Content Analysis of quantitative variables for searching and linking</i> .....	132
<i>FLAX user experience evaluation statistics</i> .....	134



<i>Automated Content Analysis of open-ended survey comments on FLAX user experience ..</i>	137
DISCUSSION.....	139
<i>Diverse motivations for adopting learning support in the MOOC space .....</i>	139
<i>Designing and evaluating augmented learning support systems for domain-specific terminology.....</i>	140
LIMITATIONS .....	142
CONCLUSION AND FUTURE RESEARCH.....	143
NOTES .....	145
CONNECTING STUDY 1 AND STUDY 2 TO STUDY 3 .....	146
INTRODUCTION TO STUDY 3 .....	147
<i>The reuse value of domain-specific OERs in higher education teaching and learning.....</i>	147
<b>CHAPTER 5: STUDY 3 .....</b>	<b>149</b>
<b>EVALUATING THE EFFICACY OF THE DIGITAL COMMONS FOR SCALING DATA-DRIVEN LEARNING.....</b>	<b>149</b>
ABSTRACT.....	149
KEYWORDS .....	150
INTRODUCTION.....	150
<i>The growing digital commons and open educational resources.....</i>	150
<i>Open data-driven learning systems in specialised language education .....</i>	152
<i>Research questions .....</i>	152
TOOLS IN THIS STUDY .....	153
<i>Transcending concordance: Augmenting academic text with FLAX.....</i>	153
<i>Research on academic text.....</i>	155
METHODS.....	160
<i>Participants .....</i>	160
<i>Procedure.....</i>	160
RESULTS.....	160
<i>Corpora description and methods of analysis.....</i>	161
ANALYSIS AND DISCUSSION .....	161
<i>Specialised term usage .....</i>	161

POLICY IMPLICATIONS.....	163
FURTHER RESEARCH.....	164
<b>CHAPTER 6: CONCLUSION.....</b>	<b>166</b>
INTRODUCTION.....	166
OVERVIEW OF KEY FINDINGS.....	166
CONCLUSIONS FROM THE THREE STUDIES.....	167
IMPLICATIONS AND LIMITATIONS .....	171
CURRENT AND FUTURE RESEARCH .....	172
<i>F-Lingo: Scaling automated domain-specific terminology learning support in MOOC platforms.....</i>	<i>172</i>
<i>FLAX Learning Collocations system: Analysing user query data .....</i>	<i>176</i>
<i>FLAX PhD abstract collections: Developing OERs for learning features of lexical paving .....</i>	<i>179</i>
CONCLUDING REMARKS .....	180
<b>REFERENCES.....</b>	<b>182</b>
<b>APPENDICES .....</b>	<b>213</b>
<i>Appendix A. ....</i>	<i>213</i>
<i>Major historical milestones in the progress of Open Access publishing .....</i>	<i>213</i>
<i>Appendix B .....</i>	<i>217</i>
<i>Non-formal Learning Support (Type A) .....</i>	<i>217</i>
<i>Appendix C .....</i>	<i>219</i>
<i>Non-formal Learning Support (Type B) .....</i>	<i>219</i>
<i>Appendix D.....</i>	<i>220</i>
<i>CopyrightX collection in FLAX log data.....</i>	<i>220</i>
<i>Appendix E .....</i>	<i>224</i>
<i>English Common Law MOOC collection in FLAX log data .....</i>	<i>224</i>
<i>Appendix F .....</i>	<i>229</i>
<i>ContractsX MOOC collection in FLAX log data .....</i>	<i>229</i>
<i>Appendix G .....</i>	<i>233</i>
<i>Essay topic list for legal English translation studies .....</i>	<i>233</i>

## ***List of Tables***

Table 1. <i>Open corpora in FLAX: Content and collaborators</i> .....	28
Table 2. <i>Meso-cycle 1. FLAX BAWE Collections</i> .....	50
Table 3. <i>Meso-cycle 2. FLAX Law Collections</i> .....	52
Table 4. <i>Meso-cycle 3. FLAX PhD Abstract Collections</i> .....	53
Table 5. <i>Meso-cycle 4. FutureLearn MOOC Collections via the F-Lingo Chrome extension</i> .....	56
Table 6. <i>Meso-cycle 1. Wikipedia in the FLAX LC system</i> .....	60
Table 7. <i>Meso-cycle 2. The BAWE corpus in the FLAX LC system</i> .....	61
Table 8. <i>Meso-cycle 3. The ACE corpora in the FLAX LC system</i> .....	62
Table 9. <i>Number of abstracts and disciplines in each area of the PhD Abstract Corpora.</i> <i>Reprinted from: Wu, S., Fitzgerald, A., Witten, I.H. &amp; Yu, A. (2018). Automatically augmenting</i> <i>academic text for language learning: PhD abstract corpora with the British Library. In B. Zou, M.</i> <i>Thomas (Eds.), Integrating Technology into Contemporary Language Learning and Teaching,</i> <i>pp. 512-537. IGI Global.</i> .....	84
Table 10. <i>FLAX Law Collections</i> .....	115
Table 11. <i>Survey question: “When you want to find out how to express something in English what</i> <i>resource(s) do you use? You can select more than one.”</i> .....	122
Table 12. <i>Survey question: “What are your/your learners’ main reasons for using the FLAX</i> <i>resources? (Select all that apply)”</i> .....	123
Table 13. <i>Statistics of example query pathway</i> .....	125
Table 14. <i>Survey question: “Evaluate the following statements about your use of FLAX”</i> .....	134
Table 15. <i>Term average in each legal English learner corpus</i> .....	162

Table 16. <i>Geographic distribution of FLAX LC users. Reprinted from Wu, S., Fitzgerald, A., Yu. A., &amp; Witten, I.H. (2019). Developing and evaluating a learner-friendly collocation system with user query data. International Journal of Computer-Assisted Language Learning and Teaching, 9(2), pp.53-78.</i> .....	176
Table 17. <i>Database usages and user preferences by country. Reprinted from Wu, S., Fitzgerald, A., Yu. A., &amp; Witten, I.H. (2019). Developing and evaluating a learner-friendly collocation system with user query data. International Journal of Computer-Assisted Language Learning and Teaching, 9(2), pp.53-78.</i> .....	178

## *List of Figures*

<i>Figure 1</i> Traditional Key-Word-In-Context (KWIC) concordance output for “design” via WebCorp .....	13
<i>Figure 2</i> Googlesque autocomplete search function in the FLAX LC system .....	14
<i>Figure 3</i> Open data from Wikimedia and WordNet linked to the search terms “design process” in the Physical Sciences sub-corpus of Academic Collocations in English (ACE) collections .....	15
<i>Figure 4</i> A history of openness. Reprinted from Peter, S., & Deimann, M. (2013). On the role of openness in education: A historical reconstruction. <i>Open Praxis</i> , 5(1), 7-14. doi: <a href="http://dx.doi.org/10.5944/openpraxis.5.1.23">http://dx.doi.org/10.5944/openpraxis.5.1.23</a> CC-BY .....	24
<i>Figure 5</i> Tweets to the hashtag: #openwashingnominee.....	25
<i>Figure 6</i> A framework for design ethnography in information systems. Reprinted from Baskerville, R.L., & Myers, M.D. (2015). Design ethnography in information systems. <i>Information Systems Journal</i> , 25, 23-46. ....	36
<i>Figure 7</i> Generic model for conducting design-based research in education. Reprinted from McKenney, S., & Reeves, T. (2012). <i>Conducting Educational Design Research</i> . London and New York: Routledge.....	46
<i>Figure 8</i> Micro-, meso- and macro-cycles in educational design-based research. Reprinted from McKenney, S., & Reeves, T. (2012). <i>Conducting Educational Design Research</i> . London and New York: Routledge.....	47
<i>Figure 9</i> Full text case study document featuring adjective collocational phrase parsing in the BAWE Life Sciences collection.....	52
<i>Figure 10</i> FLAX Related Words Android mobile application featuring an activity from EThOS PhD abstracts collection .....	55

<i>Figure 11</i> New interface for FLAX LC <i>Related Collocations</i> function .....	59
<i>Figure 12</i> Old interface for FLAX LC <i>Related Words</i> function .....	60
<i>Figure 13</i> Concept map and key derived from automated content analysis of the complete qualitative dataset .....	77
<i>Figure 14</i> Concept map and key derived from automated content analysis of the knowledge organisations' sub-dataset .....	79
<i>Figure 15</i> Concept map and key derived from automated content analysis of the researchers' sub-dataset .....	81
<i>Figure 16</i> Concept map and key derived from automated content analysis of the knowledge users' sub-dataset .....	83
<i>Figure 17</i> Role and courses taken by survey respondents 2014-2016 .....	119
<i>Figure 18</i> Age bands of survey respondents .....	119
<i>Figure 19</i> Educational background of survey respondents .....	120
<i>Figure 20</i> Employment status of survey respondents .....	120
<i>Figure 21</i> Wikification user query pathway for concept definition and related topics in Wikipedia in the CopyrightX MOOC collection .....	127
<i>Figure 22</i> Collocation user query pathway for top 100 collocations in the CopyrightX collection displaying “summary judgment” .....	128
<i>Figure 23</i> Collocation user query pathway for consulting the term “judgment” in the auxiliary FLAX LC system .....	129
<i>Figure 24</i> Learner feedback on the Cherry Basket feature in the English Common Law MOOC collection 2016 .....	129
<i>Figure 25</i> Keyword search user query pathway for “common law” in the English Common Law MOOC collection .....	131

<i>Figure 26</i> Learner feedback on the searchability of the FLAX ContractsX MOOC collection 2016 .....	131
<i>Figure 27</i> Concept map and key of themes indicating native and non-native English speakers' motivations to search for subject-specific terms and to browse linked OERs .....	133
<i>Figure 28</i> FLAX user experience for non-formal online learners, average on a scale of 9 .....	136
<i>Figure 29</i> Further comments and overall reactions to the positive and negative features of FLAX .....	138
<i>Figure 30</i> Licensing information for source material in the FLAX CopyrightX collection .....	145
<i>Figure 31</i> Licensing information for source material in the FLAX ContractsX MOOC collection .....	145
<i>Figure 32</i> Keyword search for “creative” in the CopyrightX collection .....	154
<i>Figure 33</i> Most frequent Academic Word List items in the CopyrightX collection .....	156
<i>Figure 34</i> Preview of some of the top 100 collocations in the British Law Report Corpus (BLaRC) displaying “relevant to the question” .....	157
<i>Figure 35</i> Related collocations for the word “relevant” linked in from the FLAX LC system ...	158
<i>Figure 36</i> Lexical bundles function in the English Common Law MOOC collection .....	158
<i>Figure 37</i> Wikify glossary function in the English Common Law MOOC collection .....	159
<i>Figure 38</i> F-Lingo highlighted phrases in FutureLearn MOOC video transcript .....	173
<i>Figure 39</i> F-Lingo phrase examples for “machine learning method” derived from MOOC course content and FLAX Wikipedia collection .....	174
<i>Figure 40</i> F-Lingo MOOC concept examples for “machine learning” mined with the Wikipedia Miner Toolkit .....	174

## ***Glossary***

My supervisor, Steven Shaw, joked at the beginning of my oral PhD thesis defence that he thought he had been reading Tolstoy's *War and Peace*, but it turned out to be my thesis instead. No doubt the joke was in reference to the tome-like thesis format encountered by examiners in the typical doctoral thesis examination, and no doubt the joke was intended to lighten the seriousness of the examination atmosphere. Later while compiling this glossary and the following list of acronyms that appear frequently in this thesis, I was reminded of the reference to voluminous Russian novels and how they provide a glossary in the front matter of the books with all the names, diminutives and familial affiliations for each of the characters that will appear in the stories as a handy reference for readers. The following glossary items and acronyms reflect important and frequently used terms in this thesis from the fields of computer science, linguistics, law, education, and the various open movements present in this interdisciplinary research:

### **All Rights Reserved**

“A copyright formality indicating that the copyright holder reserves, or holds for its own use, all the rights provided by copyright law.” (Wikipedia, 2019).

### **Automated Content Analysis**

An automated text analysis method for qualitative research designed to increase coding validity and to visualise the lexical co-occurrence information extracted from natural language into semantic or conceptual patterns.

### **Blended Learning**

“An approach to education that combines online educational materials and opportunities for interaction online with traditional place-based classroom methods.” (Wikipedia, 2019).

### **Computational Linguistics**

“An interdisciplinary field concerned with the statistical or rule-based modeling of natural language from a computational perspective, as well as the study of appropriate computational approaches to linguistic questions.” (Wikipedia, 2019).

### **Copyright**

“A legal right, existing in many countries, that grants the creator of an original work exclusive rights to determine whether, and under what conditions, this original work may be used by others.” (Wikipedia, 2019).



**Corpus Linguistics**

“The study of language as expressed in corpora (samples) of "real world" text.” (Wikipedia, 2019).

**Data-Driven Learning**

An approach to language learning where language is treated as data and learners as researchers undertaking guided discovery tasks.

**Design-Based Research**

A research methodology that comprises a series of approaches to solve real-world educational problems by iteratively producing and testing design interventions for the purpose of generating new theories, design principles, artefacts, practices and reforms in education.

**Design Ethnography**

“Ethnographic practice in design to ultimately understand more of the user’s perception of the object, environment, system, or service the user is engaged with.” (Genzuk, 2003).

**Digital Commons**

The creation and distribution of informational resources and technologies that have been designed to stay in the digital commons using various open licenses, including the GNU Public License and the Creative Commons suite of licenses.

**Digital Library**

“An online database of digital objects that can include text, still images, audio, video, or other digital media formats.” (Wikipedia, 2019).

**Digital Humanities**

“An area of scholarly activity at the intersection of computing or digital technologies and the disciplines of the humanities.” (Wikipedia, 2019).

**Domain-Specific Terminology**

“Words, compound words or multi-word expressions that in specific contexts are given specific meanings, including conceptual meanings—these may deviate from the meanings the same words have in other contexts and in everyday language.” (Wikipedia, 2019).

**English for Academic Purposes**

A sub-field of English for specific purposes. It usually refers to supporting students enrolled on formal higher education degree programs with using academic language appropriately for study.

**English for Specific Purposes**

A sub-field of English as a second or foreign language. It usually refers to teaching features of domain-specific terminology for academic or professional purposes.

**Formal Learning**

Education normally delivered by trained teachers in a systematic intentional way within a school, college or university.

**Lexicogrammar**

“A term peculiar to systemic functional linguistics. It was coined by Michael Halliday, the father of systemic functional linguistics, to describe the continuity between grammar and lexis.” (Wikipedia, 2019).

**Informal Learning**

“Any learning that is not formal learning or non-formal learning, such as self-directed learning or learning from life experience.” (Wikipedia, 2019).

**Knowledge Mobilisation**

The movement of available knowledge, usually from formal research, into active use. It usually refers to knowledge brokering, knowledge transfer, knowledge translation and knowledge utilisation between research institutions, research producers and research users or practitioners.

**Machine Learning**

A sub-field of Artificial Intelligence that develops algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead.

**Non-formal Learning**

Non-formal learning is the activity of understanding, gaining knowledge or acquiring skills outside the remit of being a registered student with a formal educational institution.

**Open Access**

“A mechanism by which research outputs are distributed online, free of cost or other barriers, and, in its most precise meaning, with the addition of an open license applied to promote reuse.” (Wikipedia, 2019).

**Open Data**

A mechanism by which data are “freely available for everyone to use and republish as they wish, without restrictions from copyright, patents” (Wikipedia, 2019) or other instruments of control.

### **Open Education**

“Education without academic admission requirements that is typically offered online. It broadens access to the learning and training traditionally offered through formal education systems.” (Wikipedia, 2019).

### **Open Educational Resources**

Freely accessible, “openly licensed course materials, lesson plans, textbooks, games, software” (The Cape Town Open Education Declaration, 2007) and other digital assets that can be retained, reused, repurposed, remixed and redistributed for teaching, learning, and assessing as well as for research purposes.

### **Open Educational Practices**

“A broad range of practices that are informed by open education initiatives and movements and that embody the values and visions of openness” (Koseoglu & Bozkurt, 2018).

### **Open Gratis**

Signifies freely available or read-only resources.

### **Open Libre**

Signifies flexible and customisable resources that can be re-appropriated and retained/revised/remixed/repurposed/redistributed by multiple stakeholders.

### **Open-Source Software**

“A type of computer software in which source code is released under a license in which the copyright holder grants users the rights to study, change, and distribute the software to anyone and for any purpose.” (Wikipedia, 2019).

### **Openwashing**

“Having an appearance of open-source and open-licensing for marketing purposes, while continuing proprietary practices.” (Watters, 2014).

### **Natural Language Processing**

“A subfield of computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to

program computers to process and analyse large amounts of natural language data.”  
(Wikipedia, 2019).

### **Remix**

“A piece of media which has been altered from its original state by adding, removing, and/or changing pieces of the item. A song, piece of artwork, books, video, or photograph can all be remixes. The only characteristic of a remix is that it appropriates and changes other materials to create something new.” (Wikipedia, 2019).

### **Reuse**

“The action or practice of using something again, whether for its original purpose (conventional reuse) or to fulfil a different function (creative reuse or repurposing).”  
(Wikipedia, 2019).

### **Systems Design**

“The process of defining the architecture, modules, interfaces, and data for a system to satisfy specified requirements. Systems design could be seen as the application of systems theory to product development.” (Wikipedia, 2019).

### **Text and Data Mining**

“The process of discovering and deriving high-quality information that is hidden in unstructured textual data.” (Wikipedia, 2019). High-quality information is typically derived through converting this unstructured textual data into structured data and deriving patterns so that it can be analysed and presented to users in concise and useful ways.

## *Acronyms*

### **API**

Application Programming Interface

### **ACA**

Automated Content Analysis

### **ACE**

Academic Collocations in English corpora

### **BAILII**

British And Irish Legal Information Institute

### **BAWE**

British Academic Written English corpus

### **BLaRC**

British Law Reports Corpus

### **BNC**

British National Corpus

### **BOAI**

Budapest Open Access Initiative

### **CALL**

Computer Assisted Language Learning

### **CORE**

COnnecting Repositories

### **DDL**

Data-Driven Learning

### **DE**

Design Ethnography

### **EAP**

English for Academic Purposes

### **ECL**

English Common Law

### **ESP**

English for Specific Purposes

**ESAP**

English for Specific Academic Purposes

**EThOS**

Electronic Theses Online Service

**FLAX**

Flexible Language Acquisition project

**FLAX LC**

FLAX Learning Collocations system

**HEA**

Higher Education Academy (UK)

**HEI**

Higher Education Institution

**HEFCE**

Higher Education Funding Council England

**JISC**

Joint Information Systems Committee (UK)

**KWIC**

Key-Word-In-Context

**LMS**

Learning Management System

**MALL**

Mobile Assisted Language Learning

**MOOC**

Massive Open Online Course

**NLP**

Natural Language Processing

**OA**

Open Access

**OER**

Open Educational Resources

**OEP**

Open Educational Practices

**OSS**

Open-Source Software

**OTA**

Oxford Text Archive

**OU**

Open University (United Kingdom)

**POS**

Part of Speech

**QMUL**

Queen Mary University of London

**R&D**

Research and Development

**SCORE**

Support Centre for Open Resources in Education (UK OU)

**TESOL**

Teaching English to Speakers of Other Languages

**TIRF**

The International Research Foundation for English language education

**UNESCO**

United Nations Education, Scientific and Cultural Organisation

**UX**

User Experience

**VLE**

Virtual Learning Environment

## Chapter 1: General Introduction

*“The future is already here — it's just not very evenly distributed.”* —William Gibson (1999)

The research presented for discussion in this doctoral thesis, although highly relevant to the field of applied corpus linguistics for second language education, owes its greatest recognition to open movements that are of particular relevance to the field of educational technology concerning open educational practices for the reuse of research and pedagogic content in higher education; namely the open access, open data, open-source software, and open education movements. My collaboration with the FLAX language project has emphasised affordances from these movements for designing, developing and evaluating an open data-driven infrastructure with relevant stakeholders concerning the reuse of research and pedagogic content to support the learning of domain-specific terminology in higher education.

The first section of this introductory chapter presents the new paradigm for open data-driven language learning systems design in higher education that I am proposing in this thesis. The following section provides an overview of my doctoral research vistas and presents the relevant research groups in computer science at the University of Waikato that have contributed to the development of the FLAX project. The following section introduces recent reforms in UK copyright law as they pertain to text and data mining in this doctoral research and situates these reforms within the wider context of mining modern languages from the research into applied corpus linguistics. The final section of this introduction provides an overview of the relevant literature with respects to developments with data-driven language learning, academic corpora, and openness in higher education.

### ***A new paradigm for open data-driven language learning systems design in higher education***

A basic premise underpinning the new research paradigm presented in this thesis, and demonstrated by the FLAX project, 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 language classroom, this approach requires that the tools and interfaces, and indeed the 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 alternatively assuming training to, hopefully, compensate for this lack of knowledge and experience.

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 reuse 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. Indeed, the open corpus development work presented in this thesis would not have been possible had it not been for the campaigners for copyright reform, the Internet activists, the open policy makers, the open-source software developers, and the advocates for open access, open data and open education that have made these resources available for reuse 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 and design ethnography. 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. This approach will also be necessarily 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 knowledge of this thesis, 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.

The focus on domain-specific terminology learning support via data-driven approaches is of course also decidedly different from the current 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.

Thus, this is also a new paradigm based on DDL approaches, driving domain-specific terminology 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 toward the non-specialist end-learner, but also hopefully help re-establish the centrality of language specificity to the field of EAP. The new paradigm is necessarily rooted in greater inter- or multi-disciplinarity. Given the goal of facilitating, in particular, the increasing number of learners who are coming online, 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, linguists and EAP practitioners with expertise in corpus-based and DDL approaches, and policy makers in knowledge organizations (libraries, universities, archives).

## ***Research vistas***

### *Open Educational Resources (OER) research fellowships*

Over the period 2011-2014 my doctoral research resided within two open education fellowships with the UK OU, the first of which focused on EAP and open academic practice in collaboration with Durham University, and a further fellowship in international open education with the University of Oxford, which focused on the reuse of Oxford-created OER and Oxford-managed corpora (Fitzgerald, 2013a; Fitzgerald, 2013b; Fitzgerald, Wu & Witten, forthcoming). During the period 2009-2012, the Higher Education Funding Council of England (HEFCE) would invest a total of £15,000,000 in OER projects and fellowships for the formal university sector. Toward

the end of my UK OER fellowships, from 2013 to the present, my research vistas expanded to include the non-formal higher education sector with the FLAX project's research into developing automated domain-specific terminology learning support for MOOCs, the great majority of which are invariably offered in English with no dedicated language learning support (Wu, Fitzgerald & Witten, 2014; Fitzgerald, Wu, König, Witten & Shaw, forthcoming). Early work into devising MOOC linguistic support was carried out in collaboration with the OER Research Hub at the UK Open University with a further OER research fellowship in 2014 funded by the Hewlett Foundation.

These OER fellowships, which provided the momentum for this thesis research, were intended to combine both my doctoral research practice for designing, developing open pedagogical DDL systems, and my former academic practice in EAP teaching and programme management. The research and development of open pedagogical DDL systems with the FLAX project would travel far; spanning four continents, collaborating directly with eleven universities around the world, participating in over fifty international conference events, and harvesting digital language content and metadata from a variety of open datasets managed by leading knowledge organisations to create text and data mined collections for supporting learning and teaching with domain-specific academic English. My awareness-raising role through these OER fellowships was two-fold; bringing awareness of text and data-mining approaches with the open source FLAX project for developing automated domain-specific terminology support to the open education community with a particular emphasis on the MOOC space, and bringing awareness of open educational practices with open access content to the formal EAP community with a particular emphasis on data-driven language learning approaches. Access to the open education community was afforded by the many conference and fellowship impact events as part of the HEFCE and Hewlett Foundation OER programmes in addition to the many active international OER communities online. Access to the EAP community came primarily through UK-based Professional Issues Meetings and conferences organised by 'BALEAP – the global EAP forum', and the growing presence of informal online EAP communities.

In recognition of this interdisciplinary doctoral thesis research on open DDL systems development and evaluation, the FLAX project team has won awards from the British Library and the Open Knowledge Foundation for the reuse of open digital collections for domain-specific language learning in higher education. The work has also been well received and supported by

the English language education research community with individual doctoral dissertation and graduate research grants awarded by The International Research Foundation for English language education (TIRF) and the TESOL International Association.

*The open source Flexible Language Acquisition (FLAX) language project*

The FLAX research team is unique because although there is great interest internationally in automated tools for language learning, other research groups do not have our combination of advanced computer science skills, existing support software in areas such as digital libraries and data mining, and expertise in open education for developing automated data-driven language learning systems for deployment and evaluation across formal and non-formal higher education contexts.

The Waikato Digital Libraries research group, of which the FLAX project belongs, is an acknowledged international leader in its field. FLAX is an extension of the Greenstone digital library system ([www.greenstone.org](http://www.greenstone.org)), which is widely used open-source software (OSS) that enables end users to build large collections of documents and metadata that are searchable and browseable, and to serve them on the Web (Wu, Franken & Witten, 2009). The Greenstone Software produced by the Digital Libraries Group is widely used internationally with the interface having been translated into fifty languages. In 2000 a partnership was established with UNESCO, which is centrally concerned with the dissemination of educational, scientific, and cultural information throughout the world. UNESCO and other world aid organizations use Greenstone to distribute humanitarian information in developing countries to accelerate the use of information technology for the social and economic benefit of citizens and communities.

Another well-known open-source software from the University of Waikato's Machine Learning Group is Weka (Waikato Environment for Knowledge Analysis), a data-mining tool and probably the world's most widely-used machine learning workbench. Professor Ian Witten is the acclaimed computer scientist and lead behind both of the Greenstone and Weka projects with a research career that spans 40 years. His best-known publication is the book *Data mining: Practical machine learning tools and techniques*, now in its fourth edition (2016). Equally influential was *How to Build a Digital Library* (2009). Ian Witten's vision for the FLAX project is a pragmatic one that grew out of his work with the Waikato Digital Libraries and Machine Learning groups:

In my work with Greenstone, I was lucky enough to give courses and workshops in a lot of developing countries. You know, from Cuba to Argentina. From Nepal to Vietnam. From Trinidad to Fiji. It was lovely to go around the world and give courses with people working with technology in developing countries and seeing some of their problems. And, one of the things I learned from that experience was the incredible value placed on knowing the English language. You know, if you're being brought up in Nepal and you can speak English, or you can get some facility with using English, then that puts you in an entirely different category from those people who can't. Learning English - I don't approve of the fact that this should be universal - it should be other languages perhaps, and I'm not here to promote English in any way. It's just that from a practical point of view having facility with English is incredibly important in the developing world, and also in the developed world of course. So, we started a project on trying to assist second language learners with written academic English. (Witten, 2017)

FLAX works entirely automatically, without any human input, and can be applied to any collection of academic text. The pedagogical design of the FLAX system is principled and underpinned by two theories: noticing hypothesis (Robinson, 1995; Schmidt, 2001) and inductive (discovery) learning (Bernardini, 2002). First, noticing is facilitated through input enhancement and enrichment that have been proven to be effective in learners' recognition and recall of language components found in academic texts, including: academic words, key concepts, and multi-word units such as lexical bundles and collocations. Of central importance to the noticing hypothesis and inductive learning theory underpinning the design is the collocation learning system embedded within the design of FLAX with the intent purpose of enabling learners to recognise and produce language accurately and fluently (Wu, 2010). External resources (Wikipedia, Wiktionary, WordNet) augment these academic English language components to give students opportunities to encounter them in various authentic contexts, and repeatedly (Wu, Li, Witten & Yu, 2016; Wu & Witten, 2016; Wu, Fitzgerald, Yu & Witten, 2019). FLAX also uses the Wikipedia Miner toolkit of machine learned approaches to detect and disambiguate Wikipedia concepts within a document to provide learners with associated words and phrases related to a search query (Milne & Witten, 2013). Simple interfaces have been developed so that students can use the information discovery techniques (e.g., searching and browsing) that they have become familiar with through search engines for information retrieval (e.g. Google, Bing) to discover and study the language features of their academic and professional interests (Chinnery,

2008; Wu, Franken & Witten, 2009; Boulton, 2012a; Boulton, 2015; Wu, Fitzgerald, Yu & Witten, 2019).

### ***Text and data mining (TDM)***

Many of the academic English language corpora in FLAX that will be presented for discussion in this thesis are derived from UK research content. The Hargreaves report, which was commissioned by Prime Minister David Cameron in 2010, resulted in a breakthrough limitation and exception to UK copyright law in 2014 that permitted TDM for non-commercial research and educational purposes. The key point that I wish to draw my readers' attention to here is the emphasis on non-commercial reuse of content for research and educational purposes. One of the aims of this research has been to bypass the commercial English language publishing industry with the intent of developing automated DDL systems that reuse authentic and relevant open content that is effectively off-limits for commercial reuse. A further aim is the emphasis placed on developing user-friendly DDL systems that focus on specificity in the language and discourse of the content used to create academic English corpora that reflect communication norms from different disciplines across the academy (Stevens, 1988; Hyland, 2002).

The overarching goal of TDM is to discover and extract knowledge that is hidden in free text, and to convert this unstructured textual data into structured data (Hearst, 1999) so that it can be analysed and presented to users in concise and useful ways (Ananiadou et al., 2010). Broadly speaking, TDM utilises natural language processing applications and analytical methods. Part-of-Speech (POS) tagging is a common NLP application that identifies syntactic patterns within a text, for example collocational phrases such as noun + noun (*data mining*), verb + noun (*visualise data*), and so on. Text mining requires preliminary processing steps, however, that lead up to the data mining process. This requirement is due to the unstructured nature of natural language data that is most often encountered in e.g. journals, books, documents, and in the body of e-mails, web pages and word-processed documents. TDM, according to Ananiadou et al., "comprises three major activities:

- (1) *Information retrieval*. Gathering of relevant texts.
- (2) *Information extraction*. Looking within the retrieved texts to identify, extract and structure a range of specific types of information or facts.

(3) *Data mining*. Finding associations among the pieces of information extracted from many different texts.” (Ananiadou et al., 2010, p. 3831).

### *Opening the way for text and data mining*

We have sought never to lose sight of David Cameron’s “exam question”. Could it be true that laws designed more than three centuries ago with the express purpose of creating economic incentives for innovation by protecting creators’ rights are today obstructing innovation and economic growth? The short answer is: yes. We have found that the UK’s intellectual property framework, especially with regard to copyright, is falling behind what is needed. Copyright, once the exclusive concern of authors and their publishers, is today preventing medical researchers studying data and text in pursuit of new treatments. (Hargreaves, 2011, p.1)

By way of providing a wider international backdrop for the issues surrounding TDM, under US copyright law TDM falls under the fair use doctrine and is considered a legal transformative practice rather than an act of copyright infringement that supplants an original work. For example, in the landmark case where the Author’s Guild sued Google for copyright infringement over Google’s digitisation project of in-copyright books, the court rejected the suit and ruled the Google Books Project lawful because of the greater public interest that the project served in addition to the transformative use that resulted from digitisation, including TDM; thus making something different and new from the original work and therefore justifying the digitisation of the books as an act of fair use (McSherry, 2015).

As an approach, TDM has been successfully and extensively employed to assist researchers in comparing their results with those across the literature to advance research, for example, in chemistry, pharmaceuticals, and biomedicine (see Ananiadou & McNaught, 2006; Gonzalez et al., 2016). As both a knowledge searching and a knowledge generating approach, TDM can synthesize research evidence (Natarajan et al., 2006), extract frequent tentative research hypotheses for developing new lines of inquiry (Malhotra et al., 2013), and assist with systematic reviews of the research literature (Ananiadou et al., 2009) to ensure a strong evidence base, which is viewed as vital for informing policy and practice (Chalmers, 2003). It can also assist with scanning for statistical errors, and for plagiarism across large bodies of research (e.g., Nuijten et al., 2015).

In this thesis, I argue that knowledge organisations (universities, libraries and archives, open access aggregation services) are providing an opening onto a rich seam of authentic linguistic data from the tranche of research and pedagogic content relevant to higher education research, learning and teaching that can now be mined with computational tools and applications to extract and combine information “at speeds and in ways that the human brain cannot” (Swan, 2012, p. 28). One of the priorities of this research has been to bring relevant stakeholders in higher education research, learning and teaching up to the coal face, as it were, of TDM as it applies to the various open movements with a particular focus on open NLP tools and corpus-based systems for DDL.

### *Mining modern language corpora for pedagogical purposes*

Modern language corpora have been mined for linguistic analyses and then applied to language education since at least 1969 with early work carried out by Peter Roe at Aston University (McEnery & Wilson, 1997 p. 12). This is despite corpus linguistics getting off to a somewhat wobbly start with the Chomskyan revolution that swept through much of linguistics research in the 1950s and 1960s, privileging competence over performance data, with the rise in the theory of generative grammar (McEnery & Wilson, 2001; McEnery, Xiao & Tono, 2006).

The emergence of digital tools and language corpora from the late 1960s, and their increasing prevalence and power to support corpus linguistics research, would nevertheless lead to indirect corpus-informed pedagogical applications becoming a mainstay in the development of course books and reference grammars by the commercial language education publishing industry. Further advancements with digital language corpora, and the tools developed to observe and query language data, gave impetus to Tim John’s call in the early 1990s to “attempt to cut out the middleman as far as possible and to give the learner direct access to the data”, resulting in a new pedagogical approach for direct applications with language data known as DDL (Johns, 1991, p.30). Importance is placed on empirical data when taking a corpus-informed and data-driven approach to language learning and language teaching. Moving away from subjective conclusions about language based on an individual’s internalized cognitive perception of language and the influence of generic language education resources, empirical data enable language teachers and learners to reach objective conclusions about specific language usage based on corpus analyses.



Johns' oft quoted words about cutting out the middleman in his data-driven vision for language learning still inspire debate over the feasibility of such a vision; where teacher intuitions about language and generic language learning materials are put aside in favour of powerful NLP and text analysis tools that would provide learners with direct access to some of the most extensive language corpora available, the same corpora that lexicographers draw on for making dictionaries, to discover for themselves how language is used across a variety of authentic communication contexts. As with many brilliant visions for impactful educational change, however, his also appears to have come before its time.

### ***Overarching review of the literature***

The three areas to be covered in this synthesis of the literature feed broadly and directly into this doctoral research, including developments with: DDL in language education, the reuse of artefacts of the academy in language corpus building, and openness in higher education.

#### *Developments with Data-Driven Learning (DDL) in language education*

The cornerstone technology associated with DDL, the concordancer, has confounded and furthered the existence of a meddling middleman getting in the way of language learning, this time the technology itself rather than the language teacher or the generic language learning resource. The traditional concordancer technology, and the oftentimes overwhelming raw linguistic data that concordance output lines present to end users, are frequently cited in the literature as the main obstacles to applying corpus linguistics research in second language education with DDL approaches (Widdowson, 2000; Flowerdew, 2009 Ädel, 2010). Difficulties with employing concordancers in classroom teaching led to some proponents of DDL opting for paper-based solutions that presented modified concordance data to get around the problem of the technology (Willis, 1998; Boulton 2010a; Boulton 2010b); with Johns conceding that this paper-driven approach still constituted DDL (Johns, 1993).

Language teachers and learners are often confused by what constitutes a corpus and the different ways corpora can be mined for their language data and presented for querying purposes. Anthony (2014) in a keynote lecture for the Teaching and Language Corpora (TaLC) conference demonstrated how leading experts from the field of applied corpus linguistics have also managed at times to confuse and conflate definitions for corpora with the tools used for querying them

(Tognini-Bonelli, 2001; Bernadini, 2004; Sinclair, 2004a; Hunston, 2002).

This view of corpora as language data has been echoed throughout the research into applied corpus linguistics (Hunston, 2002; Sinclair, 2004b; McEnery, Tono & Xiao, 2006). However, as a pedagogical approach, Johns was more cautious in his view that DDL would only be successful if educators were “prepared to put a great deal of work into implementing the methodology and sharing our experiences and those of our students with it.” (Johns, 1993, p. 8). Despite these concerns, the educational practice of DDL has been advanced with relative success as a means for language teachers and learners to obtain, organize, and study authentic language data derived from corpora in language education, and has been well documented in the research literature (for instance, Boulton & Thomas, 2012; Boulton & Pérez-Paredes, 2014; Chang, 2014; Cobb & Boulton, 2015; Vyatkina, 2016; Boulton & Cobb, 2017). There remains a persistent lack of exposure to and use of corpus-based systems and 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)

Nevertheless, leading researchers within the teaching and language corpora community have suggested that a point of maturity in applied corpus linguistics for DDL is nearing obtainment with few problems remaining (Reppen, 2010; O'Keefe, McCarthy & Carter, 2007; Biber, 2006). This claim is made despite data-driven methods still remaining somewhat of an exclusive research endeavour rather than a popular sport with classroom-based language teaching. Römer (2010, 2011) carried out reviews of the literature in an attempt to define the state-of-the-art with corpus-based applications in second language education and to determine where specific trends might be leading the field. Tribble (2015) reported on an ongoing series of surveys (distributed in 2001 and revised for redistribution in 2008 and 2012) to try and capture why language teaching practitioners, teacher trainers and researchers do and, in more cases, do not employ corpus resources. User-friendliness and free access are reported to be two major factors in influencing the willingness of respondents to use corpora, while “don’t know how to”, “are not familiar with” are among the reasons for not using corpora (Tribble, 2015). An extensive meta-analysis of the

literature was carried out into corpus use in second language learning by Boulton and Cobb (2017) with findings that do suggest, contrary to some opinion, that a solid base of research resulting in successful DDL praxis has nonetheless been established in second language education.

DDL researchers have also reported several factors that may hinder corpus use, including requirements of metalinguistic knowledge to formulate queries, unfamiliarity with complex search interfaces and functions, overwhelming results, and difficulties in locating and interpreting target language features in concordances, mostly in the form of keyword-in-context (KWIC) fragments and incomplete sentences (Yoon & Hirvela, 2004; O'Sullivan & Chambers, 2006; Yeh, Li, & Liou, 2007; Chen, 2011; Rodgers et al., 2011; Boulton, 2012b; Chang, 2014; Geluso & Yamaguchi, 2014; Daskalovska, 2015). For example, Chang (2014) asserts that the differing interfaces and functions of various corpus tools further increases the technical challenge whereby learners generally need to learn a new system in order to access a different corpus. A more recent study with language educators in Spain and the UK supports these findings, revealing that there was only nominal familiarity with and marginal use of freely available NLP technologies by more qualified language teachers who held PhDs. Compared with those less qualified teachers holding MAs and BAs that were even less familiar with and therefore less likely to use a wider range of freely available corpus-based NLP tools and systems (e.g. corpora, vocabulary profilers, lemmatizers, part-of-speech taggers and parsers, word lists and frequency counts) beyond the popular everyday use of online dictionaries and spell checkers (Pérez-Paredes et al., 2018).

Boulton has written most extensively on DDL, and in many ways has taken up John's baton to not only push research in the area forward but to make appeals to the corpus research and tools development community to create more accessible language learning systems for DDL, for example, with his paper entitled: "Wanted: Large corpus, simple software. No timewasters" (Boulton, 2013). The FLAX project has responded in kind by directing our research and development focus toward just that: large corpora derived from open datasets, and user-friendly tools by way of the open-source Greenstone software. In the same vein, the SKELL project has developed another dedicated tool for learners that bears the closest resemblance to our system in a design departure away from traditional concordancers (Mark Davies' Brigham Young Corpora (BYU), Wordsmith Tools, AntConc, etcetera), which have been the most widely used tools according to Tribble's survey (2015) and those reported in the DDL literature.

To exemplify this design difference, we present the traditional KWIC concordance output from the *WebCorp*<sup>1</sup> project in Figure 1, which reveals language snippets either side of the keyword search term *design*. Through this search, *WebCorp* has harvested web resources from a variety of websites with the first sample of language taken from Wikipedia as shown in Figure 1.

1) [https://en.wikipedia.org/wiki/Design\\_process](https://en.wikipedia.org/wiki/Design_process)

Text, Wordlist, text/html, UTF8 (Content-type), 2018-04-26 (Server header)

```

1:                                     Design From Wikipedia, the free encyclopedia (Redirected
2: Wikipedia, the free encyclopedia (Redirected from Design process) Jump to: navigation, search This
3:   this template message) For other uses, see Design (disambiguation). Design is the creation of a
4:   For other uses, see Design (disambiguation). Design is the creation of a plan or convention for the
5:   circuit diagrams, and sewing patterns).[1] Design has different connotations in different fields
6:   different connotations in different fields (see design disciplines below). In some cases, the direct
7:   engineering, management, coding, and graphic design) is also considered to use design thinking.
8:   and graphic design) is also considered to use design thinking. Designing often necessitates considerin
9:   and sociopolitical dimensions of both the design object and design process. It may involve
10:  dimensions of both the design object and design process. It may involve considerable research,
11:  even methods or processes of designing.[2] Thus "design" may be a substantive referring to a categorical
12:  abstraction of a created thing or things (the design of something), or a verb for the process of
13:  by grammatical context. Contents 1 Definitions 2 Design as a process 2.1 The rational model 2.1.1
14:  The action-centric model 2.2.1 Descriptions of design activities 3 Design disciplines 4 Philosophies
15:  model 2.2.1 Descriptions of design activities 3 Design disciplines 4 Philosophies and studies of design
16:  Design disciplines 4 Philosophies and studies of design 4.1 Philosophies for guiding design 4.2
17:  studies of design 4.1 Philosophies for guiding design 4.2 Approaches to design 4.3 Methods of
18:  Philosophies for guiding design 4.2 Approaches to design 4.3 Methods of designing 5 Terminology 5.1
19:  4.3 Methods of designing 5 Terminology 5.1 Design and art 5.2 Design and engineering 5.3 Design
20:  designing 5 Terminology 5.1 Design and art 5.2 Design and engineering 5.3 Design and production 5.4
21:  Design and art 5.2 Design and engineering 5.3 Design and production 5.4 Process design 6 See also 7

```

Figure 1 Traditional Key-Word-In-Context (KWIC) concordance output for “design” via WebCorp

To demonstrate the different types of user interface encountered in the FLAX system, I will now turn briefly to features of the digital library software, Greenstone, upon which FLAX is built to mimic typical web search behaviour (Wu, Franken & Witten, 2009). For example, the FLAX Learning Collocations<sup>2</sup> (FLAX LC) system provides a dynamic user interface with Googlesque autocomplete features as shown in Figure 2 for searching one- and two-word combinations of collocational phrases across several corpora: the British National Corpus (BNC), the Wikipedia corpus, the Academic Collocations in English (ACE) corpus split into four academic sub-corpora, and the British Academic Written English (BAWE) corpus. Comparisons for how collocations are used in context from across the corpora are further enhanced by offering examples of the search terms in expanded language context (Charles, 2012), and by linking open resources into the same user interface, including: related (co-occurring) words mined from different articles

<sup>1</sup> <http://www.webcorp.org.uk>

<sup>2</sup> <http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=collocations&if=flax>

across the Wikipedia corpus using the Wikipedia Miner toolkit (Milne & Witten, 2013), definitions from Wiktionary, synonyms and antonyms from the open thesaurus, WordNet, and related topics from Wikipedia as shown in Figure 3.

**Learning Collocations** My Cherry Basket Activities

Search in contemporary English (Wikipedia)

**Search results for design**

Family Words | Synonyms | Related Words | Definitions

*designed designer designers designing designs*

**design used as a noun** | **design used as a verb**

Category	Collocation	Count	Collocation	Count
design + noun	design destruction	1006	design work	839
	design designer	787	character design	696
	design destroy	677	design elements	623
	design description	604	design competition	544
	design describe	529	design firm	507
	design desire			
	design descent			
	design designation			
	design destination			
	design desert			
adjective + design	design destroyer	2464	new design	2134
	design desperate	1543	intelligent design	1282
	design descendant	848	own design	815
	design desk	740	basic design	718
	design desirable			

Figure 2 Googlesque autocomplete search function in the FLAX LC system

**Search results for design**

Family Words | Synonyms | **Related Words** | Definitions | Related topics

*design process designer problem-solving philosophy graphic solution creativity purpose art production goal product industrial guide define distinction functional user object specification depend method situation re-design contrast poster question approach advertisement aesthetically artifact principle science utilize problem use-centered solve janus-like term*

**design used as a noun** | **design used as a verb**

Category	Collocation	Count	Collocation	Count
noun + design	material design	341	design process	313
	design parameters	157	analog circuit design	156
	design phase	155	process design	147
	supply chain design	142	design space	128
				120
				>>> more

**design process**

- It allowed the design team to approach the **design process** without preconception or reticence.
- The briefing and **design process** spanned a large and complex client group incorporating community and indigenous consultation.
- Views from the street played an important role in the **design process**, as did views from the library into the adjacent park.
- The briefing and **design process** involved extensive consultation with the Building Committee, faculty representatives and senior university management.
- The result of this intensive planning and **design process** is a building characterised by flexibility, airy openness and simple, clear definition.
- It is developed in a user driven **design process** where a library, sports facilities, auditorium and teaching are weaved into one consistent building.
- Every stage of the architectural **design process** has been carried out through extensive monthly public consultations; and several workshops with various focus groups have been held.
- For the Client and its users the building represents exceptional quality and value for money both in its capital costs and its life cycle costings which were analysed throughout the **design process**.
- The historic context has thus been the main structural idea in the **design process**, ensuring the keen observer will discover a chapter of history in every corner of the yard and every peeling of the wall.

*Figure 3* Open data from Wikimedia and WordNet linked to the search terms “design process” in the Physical Sciences sub-corpus of Academic Collocations in English (ACE) collections

*Developments with reusing artefacts of the academy in corpus development*

Reuse of and analyses with academic text are acknowledged to be of considerable value in higher education, and many pedagogical implications have arisen from studies of academic text, including written and spoken genres of the academy e.g. reports, theses, lectures, seminars etcetera. In this section, I will explore the development of corpora, or collections, that are comprised of reusable authentic digital texts, or artefacts, of the academy to assist learners in coping “with a bewildering array of registers, not only to learn academic content, but also to understand course expectations and requirements” (Biber, 2007). Over the years, corpora, also referred to as collections in the terminology from the digital humanities, have been developed by researchers and teachers to investigate linguistic features that are present in academic genres, to help them identify problem areas in student academic reading, writing, speaking and listening. Some are built from highly graded university assignments, in a range of disciplines and across different genres—essays, reports, critiques, theses etcetera. Some are built from pedagogic content—lectures, seminars, textbooks, etcetera. Some are built from scholar-to-scholar communications—research articles, academic monographs, etcetera.

To provide an overview of the state-of-the-art in academic corpora or collections, the Michigan Corpus of Upper-Level Student Papers<sup>3</sup> contains 830 A-graded papers (2.6 million words) and the Michigan Corpus of Academic Spoken English (MICASE)<sup>4</sup> contains transcribed speech acts from the University of Michigan (1.8 million words). The TOEFL 2000 Spoken and Written Academic Language corpus (T2K-SWAL) comprises diverse spoken and written university registers (2.7million words), and the International Corpus of Learner English<sup>5</sup> is made up of 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. The British Academic Written English (BAWE) corpus<sup>6</sup> includes 2860 graded

---

<sup>3</sup> <http://micusp.elicorpora.info/>

<sup>4</sup> <https://quod.lib.umich.edu/m/micase/>

<sup>5</sup> <http://www.uclouvain.be/en-cecl-icle.html>

<sup>6</sup> <http://www2.warwick.ac.uk/fac/soc/al/research/collect/bawe/>

assignments (6 million words) and the British Academic Spoken English (BASE) corpus<sup>7</sup> includes 160 recorded and transcribed lectures and 40 recorded and transcribed seminars. The PhD Abstract collections<sup>8</sup> in FLAX contain upwards of 50,000 examined doctoral theses abstracts from the open access Electronic Theses Online Service (EThOS)<sup>9</sup> at the British Library (9.8 million words), while the Academic Collocations in English (ACE) collections in FLAX<sup>10</sup> are derived from 135 million peer-reviewed open access research papers and metadata via the CORE (COncnecting REpositories)<sup>11</sup> aggregation service and application programming interface (API) at the UK Open University's Knowledge Media Institute. The development of the BAWE collections, the PhD Abstract collections and the ACE collections in FLAX will be discussed alongside further open academic corpora developed from artefacts of the academy in this research in Chapters 3 and 4. Further corpora are under development, such as the Cambridge Corpus of Academic English<sup>12</sup> to complement the 400 million words of spoken and written English already included in the multi-billion-word general Cambridge English Corpus<sup>13</sup>, and the Corpus of Academic Learner English at Universität Bremen.

Students and teachers can interact with most of the corpora outlined above through accessing online user interfaces, using standard concordance tools, or by downloading entire collections. For example, the Michigan Corpus of Upper-Level Student Papers 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. However, some of the corpora outlined here are closed in terms of access. For example, the T2K-SWAL and Cambridge English Corpus are tied to the respective commercial interests of the high-stakes Test of English as a Foreign Language (TOEFL) by ETS and English language teaching coursebook publications by Cambridge University Press.

With respects to accessibility and openness, corpus developers have some of the greatest technological expertise in computational linguistics and the digital humanities and are no

---

<sup>7</sup> <https://warwick.ac.uk/fac/soc/al/research/collections/base/>

<sup>8</sup> These collections can be viewed at <http://flax.nzdl.org>

<sup>9</sup> <https://ethos.bl.uk>

<sup>10</sup> <http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=accollocations&if=flax>

<sup>11</sup> <https://core.ac.uk/>

<sup>12</sup> <http://www.englishprofile.org/camcae>

<sup>13</sup> <http://www.englishprofile.org/cambridge-english-corpus>

strangers to the meaning and practice of openness for ensuring their research outputs are accessible to other researchers. More recently, there has been an identifiable effort among some corpus researchers to engage in open practices with the development of, for example, open standards, open metadata, open collections and open tools. The CLARIN UK<sup>14</sup> infrastructure project for digital language resources and tools is one such example. Compare this technical expertise and knowledge of open standards within the corpus linguistics and computational linguistics research community with the training in materials development and coursebook adaptation that language teachers commonly receive from formal teacher training bodies. Language teachers are unlikely to be trained in the difference between open and proprietary tools and resources and would most likely find different aspects of the digital infrastructure required in corpus building to pose insurmountable barriers.

*The need to scale data-driven language learning solutions in higher education:* The open-source software (OSS) ethos has much to offer language resources developers, teachers and learners by way of the grounding development principle for enabling distributed communities to collaborate on software outputs. Duolingo<sup>15</sup> is perhaps one of the most successful data-driven and crowd-sourced application-based language learning systems available for free. Although the Duolingo system is not open source it follows the OSS principle of software development with an active worldwide community that incubates new language learning resources through an on-going beta testing phase of not only the software itself but the languages under development for Mobile Assisted Language Learning (MALL). The success of Duolingo demonstrates the global need for intelligent and adaptive language learning resources that can be met with data-driven solutions at no cost to the user.

There also exists a growing need for flexible, automated domain-specific language learning tools and resources along with an expanded open infrastructure to respond to the reality of a rapidly expanding global higher education industry that is increasingly privatised, online, and unregulated (UNESCO, 2017-2018). According to the UNESCO Education for All Global Monitoring Reports (2008; 2015; 2017-2018) there has been a boom in the number of students

---

<sup>14</sup> <https://www.clarin.ac.uk/>

<sup>15</sup> <https://www.duolingo.com/>



registered in higher education programmes around the world with an estimated 213 million students in 2015, an increase of 75 million from 2005 and 120 million from 1999.

Awareness has also been raised around a rapidly increasing number of learners worldwide who will be qualified to pursue studies in higher education but for whom access to traditional higher education to earn formal credentials will not be realistic given the enormous amount of investment required to create enough universities in time to match these growing student numbers. The modest estimate of upwards of 250 million students in 2025 is based on thirty percent of the world's population being under the age of 15 in 2013, and the already estimated 213 million students enrolled in tertiary education in 2015 (Uvalić-Trumbić & Daniel, 2011; Daniel, 2013; UNESCO, 2015; 2017-2018). Beyond the mobile elite who are the target audience for most formal EAP programmes, estimated at more than 2.5 million international students worldwide (Altbach, Reisberg & Rumbley 2009), open educational resources and practices make it conceivably possible for open NLP tools and collections to be employed not only in language schools and university language support centres, but also in online open and distance education as a means to bridging informal and non-formal learning in higher education.

Research into how data-driven learning systems can be scaled to meet the needs of a growing number of learners internationally who are coming online via informal learning and non-formal learning with, for example, MOOCs, and require assistance with domain-specific terminology for academic and professional purposes has so far been off the radar. This is despite the innovative research from corpus and computational linguistics into the Web as corpus and Web search strategies for DDL applications in classroom-based teaching (Kilgariff, 2003; Biber, 2007; Shei, 2008; Wu, Franken & Witten, 2009; Boulton, 2015). Online learning environments are arguably closer to providing a natural 'home' for the diffusion and uptake of NLP technologies and DDL methods, compared with the barriers identified from the literature from formal higher education contexts for using DDL technologies and approaches in classroom-based language learning and teaching.

OERs are predominantly created in the world's lingua francas with English-medium OERs being the most prevalent. As English is currently the international lingua franca of academic study, research and publishing, and will be for the foreseeable future, there is great demand internationally for high quality English-medium OERs that reflect the best in teaching and research practices from higher education institutions. Indeed, the position of English as

international lingua franca is wholly dependent on its use and ownership by non-native speakers of English (Graddol, 2006). This doctoral research will demonstrate findings for how English-medium open access content and OERs can be enhanced with TDM approaches so that they are more linguistically accessible, easily discoverable and adaptable for reuse in formal and non-formal higher education contexts.

Unlike traditional copyrighted materials, OERs are educational materials that are created by the educational community to be freely used, and often changed or adapted by other educators and students. They are usually online resources or e-learning materials. Although they were originally considered to be more valuable for informal online learning, there is growing evidence that they can be useful in formal university settings for fostering open academic practices (McGill, Beetham, Falconer & Littlejohn, 2010; Borthwick & Gallagher-Brett, 2014; Littlejohn & Hood, 2017). A successful example of OERs for developing open academic practices with English writing provision in higher education is WritingCommons.org<sup>16</sup>. Since launching in 2011, WritingCommons.org has hosted 6,315,882 users who have reviewed over 11 million pages (Moxley, 2018). Of interest to the discussion presented in this section on reuse, are the origins of the project of which the learning content behind WritingCommons.org started out as a print-based college writing textbook in 2003 with Pearson, which the publisher failed to promote and was therefore a flop. Pearson returned the copyright to the author, Joe Moxley, and he built his textbook into the online writing commons resource it has become today with the growing collaboration of the greater college writing education community.

### *Developments with openness in higher education*

Open has come of age it seems, with pathways to courses, the sharing of courseware code and access to research becoming increasingly free and open to learners; and with models for educational delivery and accreditation being experimented with on an almost daily basis by educators and institutions.

Openness made its entry into the formal higher education sector in the 1960s, and originally indicated that admissions barriers had been lifted from entry to formal study. This understanding of openness is still true today with formal higher education institutions operating around the globe and specialising in open and distance education, e.g. Athabasca University in Canada and

---

<sup>16</sup> <https://writingcommons.org/>

the OU in the United Kingdom with almost 200,000 registered paying students coming from a variety of traditional and non-traditional backgrounds. Nonetheless, this original definition of openness is still ground-breaking when we consider that most of the brick ‘n’ mortar higher education institutions of the world, including those with online and blended learning offerings, still maintain strict admissions policies based on entrance examinations and prerequisites. Open has come to mean much more than this, however, with the rapid ascension of OERs and MOOCs in response to the growing culture of the digital commons. Once again, the OU has been no stranger to this rise in non-formal education offerings as demonstrated in their longstanding work with the BBC, and in leading a bevy of wide-reaching open education projects including OpenLearn<sup>17</sup> and now FutureLearn<sup>18</sup>. The definition of open in higher education still remains blurred, however, when we compare the openly-licensed content of OpenLearn with FutureLearn content, for example, the latter of which is free to access but licensed all-rights-reserved. I will address this distinction between open gratis (free) and open libre (open to retain, reuse, repurpose, remix, and redistribute) in more detail in chapter 2 of this thesis. For the purpose of clarification, I will in this section, however, discuss the use and often perceived misuse, of the term ‘open’ in and beyond education in relation to principles from the commons in contrast with commonly held market principles.

Open innovation accesses and utilizes both internal and external ideas beyond the boundaries of any particular organisation, ultimately extending to include a wide variety of participants and society in general (Chesbrough, Vanhaverbeke & West, 2006). Current activity with openness in online higher education can be characterised as having reached a beta phase of maturity. In much the same way that software progresses through a release life cycle, beta is the penultimate testing phase, after the initial alpha-testing phase, whereby the software is adopted beyond its original developer community. Open education came to the attention of the mainstream press with the advent of MOOCs and has increasingly come to the attention of traditional formal higher education with an increase in funding and the adoption of open policies by government ministries of education and universities that favour the cost-saving benefits of OERs and open textbooks as evidenced recently in the US (SPARC, 2018). The participating masses can be likened to beta testers of these newly opened ways of educating. As with many recent software hits from Internet

---

<sup>17</sup> <http://www.open.edu/openlearn/>

<sup>18</sup> <https://www.futurelearn.com/>

giants such as Google (e.g. Gmail), it is highly likely that open education will remain in a state of ‘perpetual beta’ development and testing, as the higher education community investigates and measures the impact of openness in formal, non-formal and informal modalities of higher education.

Over a decade ago, and in keeping with Raymond’s OSS development philosophy from his seminal text, *The Cathedral and the Bazaar*, “Release early. Release Often. And listen to your customers” (Raymond, 1997), publisher and open source advocate, Tim O’Reilly, positioned ‘the perpetual beta’ stage of software development as the new norm:

Users must be treated as co-developers, in a reflection of open source development practices (even if the software in question is unlikely to be released under an open source license.) The open source dictum, ‘release early and release often’, in fact has morphed into an even more radical position, ‘the perpetual beta’, in which the product is developed in the open, with new features slipstreamed in on a monthly, weekly, or even daily basis. It’s no accident that services such as Gmail, Google Maps, Flickr, del.icio.us, and the like may be expected to bear a ‘Beta’ logo for years at a time. (O’Reilly, 2005)

*The need for higher education to reclaim the future of education narrative:* Running parallel to the aforementioned characterisation of open education as being in a state of perpetual beta, where users are treated as co-developers, is a growing tension around historical narratives on the future of education (Watters, 2015), and who gets to re-story these stories of prophecy. A grand narrative (Lyotard, 1984) currently exists in the field of educational technology and is applied broadly to the context of higher education. It echoes loudly for how the formal education system is broken, how the university has had no part in technological innovation – by ignoring the many contributions of the academy to Internet research and development – and how, inevitably, corporate interests and their technologies will step in to fix education. This predictive meta-narrative echoes from the chambers of Harvard Business School and venture-capital-fuelled EdTech start-ups in Silicon Valley as the theory of disruptive innovation (Christensen, 1997; Christensen, Raynor & McDonald, 2015).

The hype and decline in mainstream press coverage of the MOOC phenomenon with Silicon Valley start-ups in consortia with elite universities (Kovanović, Joksimović, Gašević, Siemens & Hatala, 2015) speaks, however, to a surrounding mistrust of this meta-narrative of disruptive

innovation in education. Drinking the Kool-Aid of the MOOC technofix story has not yet resulted in disrupting formal higher education nor has it displaced wicked problems with differentiated access to education globally. Drawing on Rittel and Webber's terminology from social policy planning in the 1970s, problems in global education can be classified as 'wicked' in the sense of being complexly resistant to resolution due to incomplete, contradictory and changing requirements, as opposed to the more 'tame' and resolvable problems that have often been the focus of technological innovation for bringing new products to market (Rittel & Webber, 1973).

Professor Clayton Christenson of Harvard Business School famously and explosively predicted in 2012 at a range of speaking venues that half of all US universities would be bankrupt in a decade. He similarly predicted that the iPhone would not be a success and that Tesla electric cars would not make it to market. Despite the unlikelihood of his prediction about US universities coming to bear, Christenson appears to be doubling down on his belief that higher education will be irrevocably disrupted (Lederman, 2017). Proponents of disruptive innovation theory will argue that it is too soon to say whether higher education, given time, will not be disrupted by technological innovation and that, for example, the MOOC phenomenon is still in its infancy. Across the Harvard University campus at the Department of History, Professor Jill Lepore deflates the grand narrative of Christenson's disruptive innovation theory, however, as merely a "theory about why businesses fail". In a striking New Yorker article, she exposes the shaky evidence and fail-safe boundaries of the theory, which takes credit from all predictions, proved and disproved, rendering it a "very poor prophet" in the business world, and sorely misplaced in "public schools, colleges and universities, churches, museums, and many hospitals, all of which have been subjected to disruptive innovation":

If an established company doesn't disrupt, it will fail, and if it fails it must be because it didn't disrupt. When a start-up fails, that's a success since epidemic failure is a hallmark of disruptive innovation. [...] When an established company succeeds, that's only because it hasn't yet failed. And, when any of these things happen, all of them are only further evidence of disruption. (Lepore, 2014)

In critiquing narratives on the future of education, Audrey Watters, in her keynote address at the Open Education 2013 conference, proposes communities rather than technology markets as the saviours of education:

Where in the stories we're telling about the future of education are we seeing salvation? Why would we locate that in technology and not in humans, for example? Why would we locate that in markets and not in communities? What happens when we embrace a narrative about the end-times — about education crisis and education apocalypse? Who's poised to take advantage of this crisis narrative? Why would we believe a gospel according to artificial intelligence, or according to Harvard Business School, or according to Techcrunch...? (Watters, 2013)

A chapter in the book, *The Battle for Open* (Weller, 2015), has been dedicated to the 'education is broken and the Silicon Valley narrative', and echoes concerns for the distortion of key principles for openness in education (Wiley, 2013); as being sold downstream through the imposed economic value system of a booming online education market. The open-washing of the open education movement, in favour of capitalising on 'open' education at a massive scale, is being viewed by critical open educationalists in much the same way as green activists view the green-washing of the green movement, with our world's most pressing environmental problems playing second fiddle to the big business of so-called green solutions – cloth shopping bags, for example – that are mismatched to the actual scale of the wicked problems the world faces.

It may be useful to look at how historical perspectives contribute to understanding the issues and challenges faced in the open education movement today with respects to open-washing. Peter and Diemann (2013) offer a historical reconstruction of the role of openness in education as shown in Figure 4 that sidesteps the hyperbole on technological innovation and disruptive innovation theory that is currently centre-stage, and instead steps backstage to moments in history where tensions existed in the philosophical underpinnings of openness as it played out in society and education with the advent of, for example, the Gutenberg press, portable books, public lectures and universities from the middle ages. The authors caution against assumptions that certain movements with openness will prevail as originally intended and direct their readers' attention to their observations that "[a]fter a period of open movements many times there have been slight but important shifts from "pure" openness towards "pretended" openness, i.e., some aspects have been modified to offer more control for producers and other stakeholders." (Peter &

Diemann, 2013, p. 12). The first of two examples of “pretended” openness is the elite self-education societies of the late 1700s and early 1800s that formed off the back of open-to-anyone coffee houses from the mid 1600s to mid 1700s. The second example from the 21<sup>st</sup> century is the early connectivist MOOC model of Downe’s and Siemens’s Connectivism and Connective Knowledge MOOC (CCK08) in 2008 that included openly licensed and aggregated content that would in turn become eclipsed by the scaled provision from MOOC start-ups such as Udacity and Coursera from 2011 onwards of read-only open access xMOOCs with All Rights Reserved content.

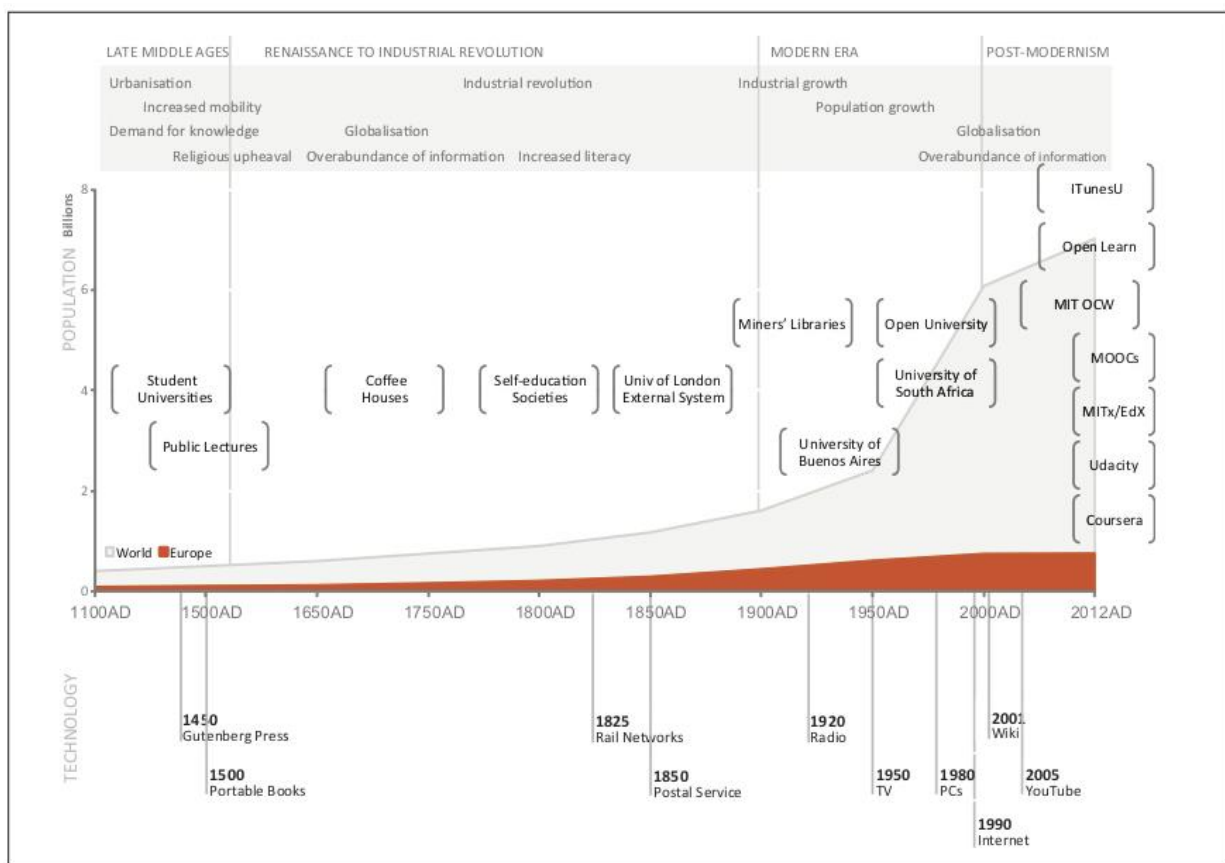


Figure 4 A history of openness. Reprinted from Peter, S., & Deimann, M. (2013). On the role of openness in education: A historical reconstruction. *Open Praxis*, 5(1), 7-14. doi:

<http://dx.doi.org/10.5944/openpraxis.5.1.23> CC-BY

A tweet from Audrey Watters in 2014 that provided a definition for openwashing garnered a lot of attention in the open education community: “Openwashing: n., having an appearance of

open-source and open-licensing for marketing purposes, while continuing proprietary practices.” (Watters, 2014). The tweet has inspired a Twitter hashtag #openwashingnominee to report and share instances of openwashing as can be seen in Figure 5.



Figure 5 Tweets to the hashtag: #openwashingnominee

In Watter’s OpenCon14 keynote address, she provided examples of how the open agenda was being appropriated by big business, and what the implications of openwashing would mean for the wider education community:

It was a subtweet, if you will, a reference to the learning management system Blackboard’s acquisition of Moodlerooms and Netspot, two companies that help provide support and deployment services for schools that use the open-source LMS Moodle. "Ours is no mere dalliance with open source," the company said. "Openwashing," I muttered under my breath. In education technology — my field, that is — I can list for you any number of examples of companies and organizations that have attached that word "open" to their products and services: OpenClass, a learning management system built by Pearson, the largest education company in the world and one of the largest publishers of proprietary textbooks. I don’t know what "open" refers to there in OpenClass. The Open Education Alliance — an industry group founded by the online education start-up Udacity. I don’t know what "open" refers to there in the Open Education Alliance. The start-up Open English, an online English-language learning site and one of the most highly funded start-ups in the last few years. I don’t know what "open" refers to there in Open English. All these append "open" to a name without really even trying to append "openness," let alone embrace "openness," to their practices or mission. Whatever "openness" means. (Watters, 2014)



Watters is, arguably, one of the most open critics of the field of educational technology and of the educational technology vendor industry. Her self-appointed position as an educational hacker who posts her critiques of the field and related industry on her Hack Education blog<sup>19</sup> rather than in scholarly journals is perhaps more radical than those traditional academic positions occupied by critical pedagogues working within formal higher education following the conventions of academic publishing. Nonetheless, critical pedagogues have also taken aim at open education and educational technology more broadly, and more specifically at the ‘learnification’ (Biesta 2010) model of higher education that over-emphasises technology and de-emphasises teaching to mere facilitation of self-directed learning (see Selwyn, 2015, and the *Learning, Media and Technology* special issue on a critical approach to open education by editors, Bayne et. al, 2015).

---

<sup>19</sup> <http://hackeducation.com/>

## **Chapter 2: Research Methods**

This chapter provides a discussion of the methodological approaches of design ethnography and design-based research as they are applied to the different research and development contexts with relevant stakeholders in direct reference to the concept and practice of knowledge mobilisation. The final section of this chapter provides an overview of the different research sites and summarises the original contribution to knowledge of the three manuscript-based studies in this thesis.

### ***Knowledge mobilisation***

The research participants in this doctoral research have been categorised according to terminology and definitions from the theory and practice of knowledge mobilisation in education (Levin, 2011) to highlight the knowledge brokering, knowledge transfer, knowledge translation and knowledge utilisation between three different types of actors in this research. The following list identifies the different participant group categories in this research and the chapters in this thesis that correspond to each participant group:

- Knowledge organisations that produce, manage and curate research artefacts
  - libraries and archives; open access aggregation services, and universities in collaboration with MOOC providers (Chapter 3: Study 1 and Chapter 4: Study 2)
- Interdisciplinary researchers engaged in R&D projects who utilise these research artefacts e.g. knowledge and results, in the design, development and dissemination of open innovative prototypes and systems for uptake and evaluation in education
  - converging from the fields of open education for language learning, corpus linguistics, and computer science with a focus on automatic natural language processing and text and data mining (Chapter 3: Study 1, Chapter 4: Study 2 and Chapter 5: Study 3)
- Knowledge users in education who are, on the one hand, educators and students in formal higher education, and on the other hand, informal and non-formal learners from the general public, for which both groups are accessing, utilizing and evaluating the same open innovative prototypes and systems via the Internet in their respective educational contexts

- formal higher education - language learners, teachers, and programme managers from formal academic English language and translation studies university programmes (Chapter 3: Study 1 and Chapter 5: Study 3)
- non-formal higher education - MOOC learners, learning technologists, and academics in the role of subject matter experts from non-formal online learning programmes (Chapter 4: Study 2)

There are numerous terms and definitions for what in essence knowledge mobilisation as a research activity is that vary across different sectors and disciplines. The underlying goal of making research more meaningful in practice and policy for organisational and system improvement remains the same, however, whether it is characterised as knowledge translation in the design and health sectors or as knowledge management in the business sector.

A central proposition of this thesis with publications is that where language corpora have been deployed in the research for linguistic analyses by researchers, the knowledge generated has often failed to translate into the design of openly accessible pedagogical applications for DDL. Instead what we have witnessed is corpus systems that have been designed and developed primarily by and for corpus linguists for research purposes. This failure in knowledge translation is due in no small part to the following issues: copyright restrictions with the texts in corpus building that inhibit text data mining and sharing; subscription costs with NLP and text analysis software tools that restrict access; and complex user interface designs of NLP and text analysis tools that limit uptake and utilisation by non-expert users, namely language teachers and language learners.

Table 1 below 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 project.

Table 1. *Open corpora in FLAX: Content and collaborators*

<p><b>Learning Collocations System<sup>20</sup> in FLAX (2009 - 2019)</b></p>
---

<sup>20</sup> <http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=collocations&if=flax>

Content	<ul style="list-style-type: none"> <li>• Wikipedia corpus of contemporary English derived from three million Wikipedia articles comprising three billion words (Wu &amp; Witten, 2016)</li> <li>• British National Corpus (BNC) of 100 million words (BNC Consortium, 2007)</li> <li>• British Academic Written English (BAWE) corpus of 2500 pieces of assessed university student writing from across the disciplines (Nesi, Gardner, Thompson &amp; Wickens, 2007; Nesi &amp; Gardner, 2012)</li> <li>• Academic Collocations in English (ACE) corpora of harvested open access content and metadata from 135 million articles residing in open journals and open repositories</li> </ul>
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) <sup>21</sup> team, UK Open University (ACE corpora)
Researchers	FLAX team (Wu, 2010; Wu, Franken & Witten, 2010; Wu, Witten & Franken, 2010; Wu, Franken & Witten 2012; Franken, 2014; Wu, Li, Witten & Yu, 2016)
Knowledge Users	Waikato University computer science students (Wu, Fitzgerald, Yu & Witten, 2019); Durham University EAP teachers and students (Fitzgerald, 2013a); University of Oxford OER International stakeholders (Fitzgerald, 2013b)
<b>British Academic Written English (BAWE) Collections<sup>22</sup> in FLAX (2012)</b>	
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 (Wu & Witten, 2016)

<sup>21</sup> <https://core.ac.uk/about#mission>

<sup>22</sup> <http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=BAWELS&if=flax>

Knowledge Users	Durham University (Fitzgerald, 2013a); University of Oxford (Fitzgerald, 2013b)
<b>British Law Report Corpus (BLaRC)<sup>23</sup> in FLAX (2014)</b>	
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) <sup>24</sup> aggregation website.
Knowledge Organisations	BAILII
Researchers	Universidad Murcia (Marín 2014; Marín & Rea, 2014); FLAX team
Knowledge Users	Law MOOC learners
<b>MOOC<sup>2526</sup> / Micro-Networked Course<sup>27</sup> Collections in FLAX (2014–2016)</b>	
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; LACELL group, Universidad Murcia
Knowledge Users	MOOC learners and MOOC subject matter experts (Fitzgerald, Wu, König, Witten & Shaw, forthcoming); legal English translation studies teachers and students at the University of Murcia (Fitzgerald, Marín, Wu & Witten, 2017; Marín, Orts & Fitzgerald, 2017)
<b>PhD Micro-abstract corpora<sup>28, 29</sup> with FLAX mobile<sup>30</sup> activities (2014-2015)</b>	
Content	Domain-specific micro abstract corpora e.g. in the areas of Law, and Water Politics and Tourism Studies. Developed in collaboration

<sup>23</sup> <http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=BLaRC&if=>

<sup>24</sup> <http://ials.sas.ac.uk/digital/bailii>

<sup>25</sup> <http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=contractlaw&if=>

<sup>26</sup> <http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=englishcommonlaw&if=>

<sup>27</sup> <http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=copyrightlaw&if=>

<sup>28</sup> <http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=flaxc383&if=flax>

<sup>29</sup> <http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=flaxc404&if=>

<sup>30</sup> <https://play.google.com/store/apps/developer?id=FLAX%20TEAM&hl=en>

	with EAP teachers at Queen Mary University of London for use on summer EAP pre-sessional courses. Developed with web-based and mobile language learning activities using the suite of mobile applications for Android from FLAX
Knowledge Organisations	British Library Labs <sup>31</sup> and EThOS <sup>32</sup> at the British Library
Researchers	FLAX team
Knowledge Users	EAP teachers and learners at Queen Mary University of London (Fitzgerald, Wu, & Barge 2014)
<b>PhD Abstract Corpora<sup>33</sup> in FLAX (2015–2016)</b>	
Content	9.8 million-word corpus derived from the metadata, including the abstracts, of 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 (Wu, Fitzgerald, Yu & Witten, 2018)
Knowledge Users	EAP teachers and managers at Queen Mary University of London; Current research with MOOC learners via F-Lingo <sup>34</sup> Chrome extension and FutureLearn platform
<b>Academic Collocations in English (ACE) <sup>35</sup> Collections in FLAX (2018-2019)</b>	
Content	Harvested open access content from open journals and open repositories divided into four sub-collections: Arts & Humanities, Social Sciences, Life Sciences, Physical Sciences
Knowledge Organisations	CORE (COncnecting REpositories) team, UK Open University
Researchers	FLAX team
Knowledge Users	<ul style="list-style-type: none"> <li>Planned user query data analysis research with the FLAX LC system learners worldwide</li> </ul>

<sup>31</sup> <https://www.bl.uk/projects/british-library-labs>

<sup>32</sup> <http://ethos.bl.uk/Home.do>

<sup>33</sup> <http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=PAAH&if=flax>

<sup>34</sup> <https://chrome.google.com/webstore/search/flingo>

<sup>35</sup> <http://collections.flax.nzdl.org/greenstone3/flax;jsessionid=F00696E7EDD1AABF0B7BED4E30A88931?a=fp&sa=collAbout&c=accollocations&if=flax>

	<ul style="list-style-type: none"> <li>Planned research with MOOC learners via F-Lingo Chrome extension and FutureLearn platform</li> </ul>
--	---

## ***Design Ethnography (DE)***

*“I am not a teacher; only a fellow traveller of whom you asked the way. I pointed ahead— ahead of myself as well as of you.”* —George Bernard Shaw (1913)

John Huges is renowned for pioneering the use of ethnography for systems design, and for attracting a cohort of sociologists and software engineers who founded what later became known as the Lancaster School. The school’s primary concern was to address:

the turn to the social that occurred in the late 1980s as the computer moved out of the research lab and into our collective lives, and the corresponding need that designers had to find ways of factoring the social into design. (Crabtree, Rouncefield & Tolmie, 2012, p. 7)

Huges et al. (1992) characterise early approaches for engaging in design ethnography (DE) as, “faltering from ethnography to design” whereby ethnographers started to engage with people in the field but in a tentative way that often lacked “a kind of *sociological sensitivity*” (Crabtree, Rouncefield & Tolmie, 2012, p. 13). Baskerville and Myers (2015) propose a framework for enabling the design ethnographer to move beyond tentative engagement with users in the field toward active engagement with people in the field. In this chapter section, I will apply Baskerville & Myers’ (2015) framework for DE for the purpose of charting the progression of my doctoral research as it has iterated toward open educational practices for open DDL systems design.

The Cape Town Open Education Declaration (2007) stated more than a decade ago that:

... open education is not limited to just open educational resources. It also draws upon open technologies that facilitate collaborative, flexible learning and the open sharing of teaching practices that empower educators to benefit from the best ideas of their colleagues. (Cape Town Open Education Declaration, 2007).

Becoming an open educational practitioner requires that you travel far from traditional educational and research practices, often into uncharted territory. This journey can be both compelling and challenging as I have experienced first-hand in my endeavour to bring an awareness of new approaches to DDL systems design that draw on open educational practices and resources; to those working in traditional face-2-face (f2f) EAP and to those learning and delivering learning support in non-formal online education (MOOCs). How did I travel this far to become a postdoctoral research fellow attached to the Department of Computer Science at Waikato where I am currently doing the final edits on my thesis? Indeed, how does a former classroom English language teacher who always found DDL systems to be too overwhelming to use with her students, end up working with some of the world's leading computer scientists in the collaborative design of innovative open DDL systems? The short answer is that computer scientists engaged in educational software R&D need applied social scientists to devise feasible educational applications for the software they are developing, and educators need computer scientists to design technologies to mitigate real-world educational problems.

The new research paradigm for open DDL systems design presented in this thesis can be seen to articulate with wider organisational frameworks and domains of activity. As mentioned earlier, open data-driven language learning systems design as an approach is learner-centric and operates with the interface to the learner. The learning modalities by which the learner interfaces with the FLAX system are of central importance to this design research. Formal EAP provision is looking at predominantly classroom-based academic practices in traditional brick and mortar universities that cater to the mobile elite (Altbach, Reisberg & Rumbley, 2009). Whereas, open education provision is looking at the whole gamut of formal, non-formal and informal digital scholarship practices (Weller, 2011). In addition to the open online resources that can be leveraged to provide educational opportunities for anyone with an Internet connection, including the estimated upwards of 100 million learners currently seeking access to the formal post-secondary sector (Uvalić-Trumbić, S & Daniel, J., 2011).

#### *A framework for design ethnography*

We can summarize the nature of DE by saying it is a form of ethnographic research that is more than just immersed, and more than just participative, but one in which the researcher is actively intervening in changing the subject area – the context – in which the researcher is researching. The researcher is



actively engaged with others in a future-oriented way: designing, creating, innovating and improvising artefacts that may affect the cultural and social values under study. (Baskerville & Myers, 2015, p. 30)

Design ethnography differs from traditional ethnography in the fields of sociology and anthropology. The former, which has also been referred to as design anthropology (Gunn & Donovan, 2013; Gunn, Otto & Smith, 2013), involves prescriptive elements where the design process is viewed as an intervention and the designer ethnographer is designing artefacts to support or change the everyday life activities of the communities being researched. The latter involves purely descriptive elements where the ethnographer is immersed in the everyday life activities of the communities being researched (Blomberg, 1993; Salvador, Bell & Anderson, 1999; Otto & Smith, 2013). Two further differences have been identified with respects to temporality and materiality. Shorter time frames are typical for DE and have been referred to as “rapid ethnography” (Bichard, 2010, p. 45) in order to gain insights into users’ everyday life activities while meeting the time constraints placed on industry-based design projects. DE is often conceptualised as being *in correspondence* with a future orientation and as being open-ended in its pursuit (Gatt & Ingold, 2013). It is also viewed within the wider field of design as being a “proscriptive action, that is actively reflecting within a present moment on future action and contingency” (Wakkary, 2005, p. 66). The material dimension is also centralised in DE with practices for conceptualising, visualising and prototyping (Otto & Smith, 2013; Baskerville & Myers, 2015). The aforementioned details for differences observed with DE with regards to temporality and materiality contrast with traditional ethnographies that are typically carried out over years of work in the field and where the material dimension is not central to the work.

### *Ethnographic toolkit*

Examples of ethnographic tools applied to this thesis research include fieldwork and immersion, design workshops (Emery & Devane, 2007), think aloud designing (Eaglestone, Ford, Brown & Moore, 2007), co-planning and co-designing (Kilbourn, 2013), blogging and design diaries (Naur, 1983), design thinking (Lugmayr, Stockleben, Zou, Anzenhofer & Jalonon, 2014), interviews and focus discussions, work emails and professional discussion-list postings, and conversation analysis (Salvador, Bell & Anderson, 1999). Study 1 in particular speaks to most of the ethnographic tools used in this thesis research.

In the following sub-sections, I will present an assemblage of thick descriptions (Geertz, 1973) or ethnographic accounts (LeCompte & Schensul 1999, p. 17; Clifford 1990, pp. 51-52), which are part narrative and part design diary, and which are interspersed with the adopted framework for design ethnography by Baskerville and Myers (2015); to stop the clock as it were and to reorder the past that has been observed and jotted down; to surface, contextualize and assemble the activity of this design research into open educational practices and resources for designing open DDL systems in higher education.

Following the work of Whitaker (1996), ethnography is approached more contingently in this research, “as a form of learning, rather than absolutely, as a form of representation.” (Whitaker, 1996, p.1). The DE presented in this chapter section will refer to ethnographic methods that contribute to the design of artefacts in addition to reflecting on, “the design process itself as a subject of ethnographic analysis” (Baskerville & Myers, 2015, p. 27). Here, I draw on my personal accounts of moving into different design research contexts, and evaluations made by the different social actors in this research concerning open educational practices and the re-use of open access research and pedagogic content in DDL systems design and development. As part of the reflexive writing process, (see Study 1), I have re-storied the stories of participating individuals and institutions, placing them in chronological sequence and providing causal links among themes and concepts generated from the automated content analysis of qualitative data collected over the years of this DE work with the FLAX project. Moreover, these accounts continue to inform the design of open-source digital library software for developing flexible open English language learning and teaching collections in the FLAX system.

Figure 6 shows separately the interactive DE framework by Baskerville & Myers, 2015 for design activities that [1] lead to the production of ethnographies and [2] ethnographies that result from design activities. The composite phases that make up this framework will be discussed in the following sub-sections of this chapter. Where necessary, I will point to further sections in this thesis that better represent some of the stages in Baskerville & Myer’s framework for DE

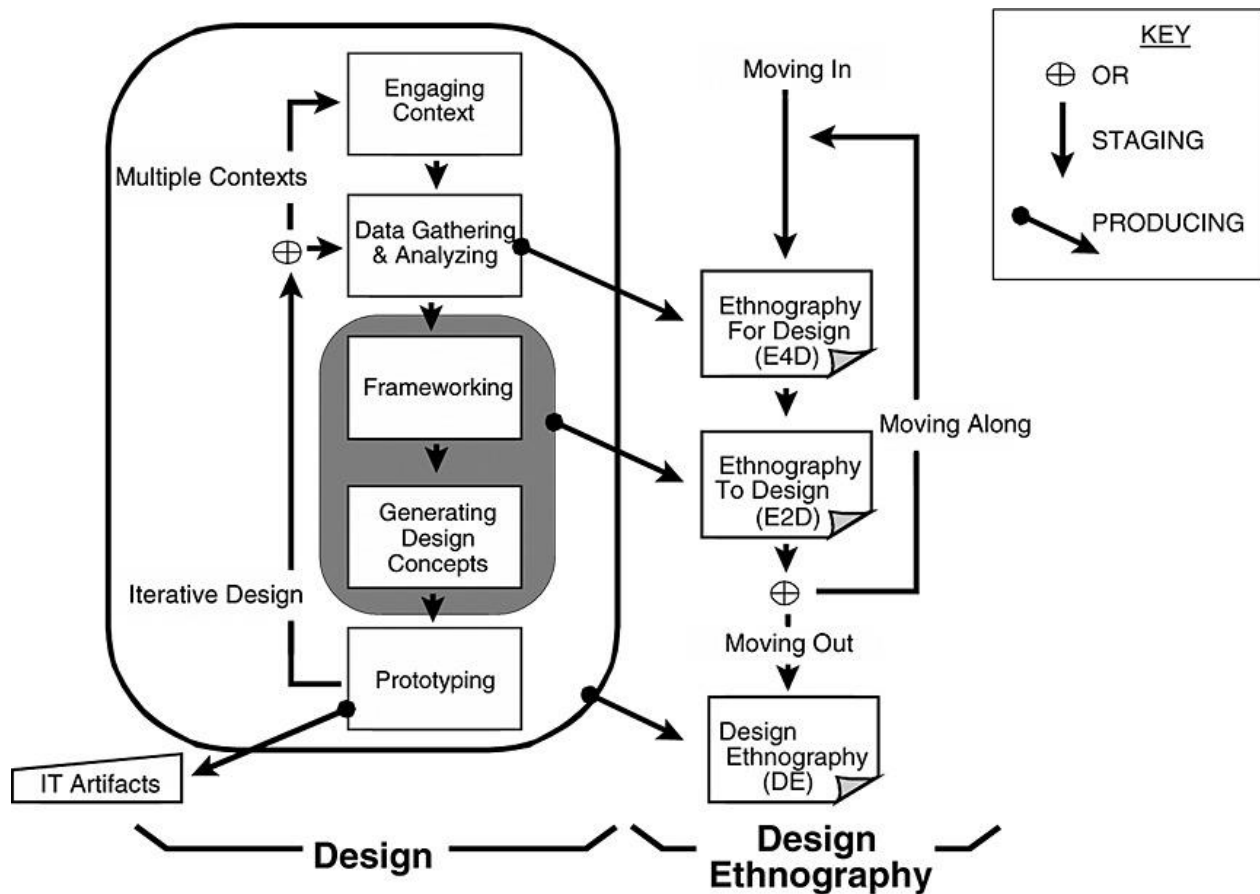


Figure 6 A framework for design ethnography in information systems. Reprinted from Baskerville, R.L., & Myers, M.D. (2015). Design ethnography in information systems. *Information Systems Journal*, 25, 23-46.

### Engaging context

Baskerville & Myers (2015) refer to the first stage in the DE framework as “establishing the context for the forthcoming design activities... [as a process] ...of engagement setting” (Baskerville & Myers, 2015, p. 32). This stage may also require the pre-emption of any possible problems that may arise due to the sometimes-conflicting roles of being a researcher and a designer in the DE process (Rapoport, 1970).

I have worn different hats in this research: those of researcher and knowledge user based on my current research and development experience with the open FLAX project, and my past experience as an EAP teacher and manager. The phenomenon of wearing more than one hat is shared with the other participating researchers in this research who also have a background in

language teaching (see Study 1). One criticism levelled at the designer ethnographer and those educators engaged in design-based research methods is the often-dual role of researchers who act at the same time as promoters of the technologies and systems they are engaged in designing and evaluating (Sanjek, 2004; Anderson & Shattuck, 2012). In Study 1, myself and another FLAX researcher will reflect on our involvement with the project and the challenges that arise with CALL research, which has been characterised, not-altogether positively, as being superficially multifaceted wherein actors, technologies, methods and theories from different disciplines are frequently converging but without producing much in the way of unique and contributing theories from the field of CALL (Levy, 1997; Colpaert, 2004; 2018).

Researcher bias is a phenomenon that haunts the research process. Throughout this research, I have encountered a fair amount of resistance to open educational practices and resources. In a blogpost I wrote for a well-known English language teachers' blog, *ELTjam*, in an attempt to try and engage language education practitioners working in traditional English language teaching and publishing in discussions about the encroaching online culture of the digital commons, including the piracy of commercially published English language coursebooks that live a second life online in .PDF format, I was characterised in the comments section as being subversive:

Finally, instead of being someone a print publisher might work with I suggest they see you as a wolf in sheep's clothes. If I were you, I would embrace that calling. (Fitzgerald, 2015).

Openness in education, and in any domain for that matter, can be viewed as both subversive and interventionist - as having a radical agenda for upending the status quo by opening up access to education. These views demonstrate how openness is perceived, often correctly, as an affront to traditional business models in formal and commercial education provision. Norton, in his keynote address at the 2012 Association for Learning Technology conference, raised the issue that often free technologies and the growing communities and practices that have gathered around them, for example, Craigslist, Napster, are mischaracterized by the media and those traditional businesses that have been supplanted as seemingly coming from out of nowhere; that their disruptiveness could not have been anticipated let alone harnessed for viable business opportunities. He refutes this position, using the example of Wikipedia, by underscoring the closed mindset inherent among the echelons of Encyclopaedia Britannica where it was perceived

that nothing of authoritative educational and research value could come from the commons (Norton, 2012). This closed and dismissive view of Wikipedia is still persistent among many academics and teachers. This despite Wikipedia's massive uptake in education and research and its ability to significantly disrupt traditional educational publishing (Bosman, 2012). I mention Wikipedia here as our Wikipedia corpus in FLAX, although the most popular in terms of uptake according to our system log data (Wu, Fitzgerald, Yu & Witten, 2019), has also been met with the most resistance by English language teachers when I have demonstrated its features at conference events.

Leading activity theorist, Engeström, duly notes that sociological research interventions should expect subversion, resistance and struggle from local social actors in response to their intervenors. These subversive actions, writes Engeström, "... are essential core ingredients of interventions, and they need to have a prominent place in viable intervention methodology" (Engeström, 2009, p.319). Merlucci (1996) also points to the importance of resistance in intervention research and states the necessity for, "actors themselves [to be able to] make sense out of what they are doing, autonomously of any evangelical or manipulative interventions of the researcher" (Merlucci, 1996, pp. 388–389).

### *Moving in*

Moving in connotes both a change of life for the researcher and others in the context. [...] rather than beginning with a problem (problem formulation or problem awareness, Argyris & Schön, 1991), DE begins with the immersion of the researcher into the design practices of the subjects (which may be in progress or ongoing). (Baskerville & Myers, 2015, p. 33)

The biggest shift in my practice as a designer ethnographer occurred in late 2010 at the end of a senior EAP management role at Durham University's Foundation Centre, and at the beginning of my first OER for academic practice fellowship that was managed by the OU and based at Durham's English Language Centre. The following design diary extract describes this shift in my design practice as emphasised by moving in toward the open educational practices present in the UK OER community:

The end of my EAP management contract at Durham's Foundation Centre has ended on a high note with funding to present my early research collaboration with the FLAX project at the OER 2010

conference in Cambridge. I have discovered a heady nest of OER practitioners working across UK Higher Education Institutions (HEIs) on innovative OER projects funded by a 3-year programme with HEFCE in collaboration with the Joint Information Systems Committee (JISC), the Higher Education Academy (HEA) and the Support Centre for Open Resources in Education (SCORE) managed by the OU. Discovering the UK OER community has been like discovering the missing link I didn't realise I was looking for between the open-source FLAX project and my work in EAP. Open educational practices and resources seem to be the way forward. Networking with the UK OER community has begun almost immediately, and the momentum is palpable. Off the back of the OER10 conference, I applied for and was offered a SCORE fellowship from the OU with funding from HEFCE.

I think I've met with what must be my first resistance to open education challenge (head on). As the funding for my fellowship needs to be channelled through a UK HEI, I went into the Language Centre at Durham and asked former colleagues from the pre-sessional summer EAP programme to champion the idea to the director, only for him to then reject it as he did not view it as central to the business of the centre. Knowing that I had already secured the funding, and that I had the support of colleagues, I went above the director to the dean who in turn overrode the director's decision. A sign of my commitment to beginning this OER journey, perhaps? What a relief though that I'm in. The feeling is galvanising.

My fellowship, which is managed by the OU, has begun with monthly and sometimes fortnightly journeys by train down the East Coast line from Durham to London King's Cross, then a short walk over to Euston station to catch a connecting train over to Milton Keynes, and this is expected to go on for about a year or so. Already my perception of my teaching and resources development practices is becoming radically altered, and I have started to associate the modern OU campus, with no onsite students and regular meet-ups with academics to discuss openness and online learning and teaching, as journeying toward the future of what higher education is becoming; whereas my return journeys up to Durham, with the monolith of medieval Durham Cathedral being the first to meet my view from the train, as journeying back to the past of what higher education has been. (Alannah Fitzgerald, Josephine Butler College, Durham University, January 2011)

### *Data gathering and analysing*

... anchored more to practical action (such as questionnaires, objective observations and instrumented measurements of material performance) for use in designing the intended artefacts. (Baskerville & Myers, 2015, p. 33).

This stage in the DE framework is covered in significant detail in each of the three studies in this thesis.

### *Ethnography-4-Design (E4D)*

E4D is aimed at a deeper description of the users or consumers of the artefact being designed. It facilitates better designs by trying to obtain a deeper understanding of the future users of the proposed product. It presupposes that this better understanding will lead to more ideal features being incorporated into the design. By learning about the ideas, beliefs, values and behaviours of users and consumers, designers can translate these into useful ideas for design, engineering and marketing. (Baskerville & Myers, 2015, p. 27).

The E4D stage of the DE framework corresponds with the reflection and evaluation micro-cycles of DBR for the FLAX collections designed throughout this doctoral research. These micro-cycles are described more fully in relation to each of the collections in the section of this chapter dedicated to DBR. The E4D stage is also reflected in Studies 1 and 2 where emphasis has been placed on presenting the perceptions of all of the participant groups who have engaged with collections design and evaluation in this research.

Although the findings from this DE research are tied to issues with language as data in corpus and data-driven learning systems development, wider issues pertaining to CALL and blended learning for the reuse and remix of open access content in EAP materials development practices for classroom teaching and blended learning will also be discussed in Study 1. EAP practitioners are confronted on a regular basis with issues surrounding the use of technology and the reuse of real-world language data. In terms of granularity, the collaborative research in Study 1 presents reflections at a macro level on the development of domain-specific corpora that are augmented by massive amounts of data in the form of content and metadata about that content. The research also reflects at the micro level on data as content in the form of individual texts e.g. MOOC lectures licensed as OERs and authentic research articles managed under open access reuse policies. Reflections captured in Study 1 from the qualitative data collected in this research speak to how these texts can be adapted for classroom teaching and blended learning resources development as part of the wider EAP materials development remit.

### *Frameworking*

Such frames are a complex form of sensitizing concepts that provide indicators for design directions, a starting point from which to make design concepts (Bowen, 2006). Frameworking marks the transition point at which the designers transform the foregoing research in the design context into concepts that will drive their design decisions. (Baskerville & Myers, 2015, p. 34).

The frameworking stage in this research has resulted in the identification of two central macro-cycles of collections design and development coursing throughout the collaborative project work with the FLAX team. These two macro-cycles will be described and discussed in detail in the following sub-section of this chapter on DBR.

### *Generating design concepts*

The generative aspect of DE is centred in this process. The engagement of the ethnographic researcher as a participant observer in practical acts of creation, innovation and improvisation of designs is one hallmark of DE. (Baskerville & Myers, 2015, p. 34).

This stage in the DE framework corresponds with the maturing interventions and growing theoretical understandings of the design research over time. Design principles are generated at this stage of DE and will be addressed specifically in the following section in this chapter on DBR.

### *Ethnography-2-Design (E2D)*

... the researcher uses an ethnographic frame to study the cultural and societal aspects of the designers. However, within a DE framework, this is a participative study. The researcher is not only studying the designers but participating by doing a share of the designing. (Baskerville & Myers, 2015, p. 34).

I have been engaged in the co-design of language collections in the FLAX system for approaching a decade now, so working with computer scientists does not feel as foreign to me now as it did at the beginning when I joined the project collaboration.

The following blog post was written in 2013 as part of my OER international fellowship with the University of Oxford, and describes the types of design workshops and developer meetings that are typical of the FLAX project team members:



While back in New Zealand late last year with the FLAX project team at the Greenstone digital library lab at Waikato, every week I would participate in developer meetings with the computer scientists behind the project and one other English language teacher from the Chinese Open University who is also basing her PhD research on the FLAX project. Well-versed in natural language processing and research on current web-base search behaviour, the computer scientists behind the interface designs of the FLAX collections and activities were adept at exploiting available linguistic resources for the development of simple-to-use language learning collections and OSS text analysis tools. I soon picked up what the limitations of the different technologies and resources were. The focus of these design workshops was to develop rapid prototype resources for envisioning and discussing how they could work across different language learning scenarios. I was able to observe and contribute to many iterations of the resources currently under development and I will be bringing these resources to the fore of future blog posts in this series. (Fitzgerald, 2013c)

### *Moving along*

Moving along to other settings can broaden the knowledge scope of the ethnography and develop knowledge that spans multiple contexts. It provides not only an understanding beyond a particular place or configuration but also how actions in different contexts mediate the relationship between materials and knowledge (Kilbourn, 2013). (Baskerville & Myers, 2015, p. 35)

The moving along stage of my DE journey was clearly demarcated throughout my OER fellowship work with the OU and Oxford by moving toward working with knowledge organisations (libraries, archives, universities in collaboration with MOOC providers). The aim behind my decision to move in this direction with the DE was to see how far I could get with pushing at the parameters of policy for the reuse of open access research and pedagogic content deemed useful for learning features of specialised varieties of English. This work with knowledge organisations is covered most explicitly in Study 2 and in reference to current and planned work in Chapter 6. The following entry from 2012 in my design diary describes a shift in realisation toward the end of my first OER fellowship with the OU, that higher education offerings, including those from mainstream MOOCs, had become stuck at the default setting of read-only open access to content on the open education continuum. What myself and the rest of the OER fellows observed in 2012 was a retro-step with openness in higher education following the UK OER funded fellowship period which, coincidentally, also ended in 2012:

The flurry of UK OER funded activity continued until 2012 when it gave way to and was eclipsed by the mainstreaming of the MOOC phenomenon with Coursera, edX and FutureLearn, who were on a mission to sign on esteemed UK HEIs but not necessarily those of us, we now realised, who had developed expertise in openness with UK OER. (Alannah Fitzgerald, The Open University, April 2012)

### *Prototyping*

Prototyping is common to the field of computer science and software development whereby early samples, models or releases of an artefact are introduced somewhat *synthetically* into a research environment (Simon, 1996) to test a concept or process that can be evaluated to develop theory from the design, and to build further iterations of the design. “For the design ethnographer, designing is not a human action that ever completes [and is] ...contingent on future design concepts that are yet unknown (Baskerville & Myers, 2015, p. 35).

Bricolage as a DIY (Do-It-Yourself) research method best characterises the R&D prototyping process with the FLAX project with trying to raise awareness around openness with various social actors in English language education and open education. Toying and tinkering with open linguistic datasets, and trialling and testing these out with research participants across different modalities for learning in higher education became the foci in the prototyping stage of this research. The bricolage process in educational research Kincheloe (2005) denotes ‘playing around’, for learning from and working towards solving problems. From the field of organization studies, Weick (1995, p. 350) identifies “intimate knowledge of resources, careful observation and listening, trusting one's ideas, and self-correcting structures with feedback” as requirements for successful bricolage in organizations. These types of requirements have been played out and explored in this research as micro-, meso- and macro-cycles of DDL systems design, and will be explored in more detail in the following section of this chapter dedicated to DBR.

### *Artefacts*

‘that bundle of material and cultural properties packaged in some socially recognizable form such as hardware and/or software’ (Orlikowski & Iacono, 2001, p. 121).

First, their production is part of the design process. Second, they become part of the context of future designing activity and as such become one of the future sources for data gathering and analysis. (Baskerville & Myers, 2015, p. 36)

Table 1 at the beginning of this chapter identifies all of the open artefacts that have resulted from this DE research and reflects how the FLAX project has managed to sustain itself over the years by engaging in open educational practices for open DDL systems design.

### *Moving out*

...the DE that is produced should be one that is more focused on the wider boundaries of knowledge proceeding from the research (Kilbourn, 2013). However, these boundaries should also extend to the generation of conceptual alternatives to current theory and proposed explanations for future possibilities. (Baskerville & Myers, 2015, p. 36)

The moving out stage of the DE framework speaks to the new paradigm proposed by this doctoral research as foregrounded in the introductory chapter of this thesis. To date, the FLAX project team have published many papers on our work with designing, developing and evaluating the usefulness of our systems for DDL. Each of the studies in this thesis makes a specific original contribution to knowledge, which will be discussed in more detail in the final section of this chapter.

One of the drivers of this research has been to engage with non-specialist end users, namely teachers and learners in higher education, to collaborate in the design and development of open language collections, and the interfaces for searching, browsing and interacting with these collections. Research and development work between the FLAX project and various stakeholders continues to the present day. Design ethnography and design-based research have played, and continue to play, a focal part in this collaborative R&D work, which is presented for discussion here in this chapter on research methods. In particular, Study 1 of this thesis provides an overview of all of the stakeholders engaged in this on-going design-based research with the FLAX project. Studies 2 and 3 of this thesis provide a further lens onto this work with design ethnography and design-based research methods that zooms in on two of the stakeholder groups identified in Study 1.

### ***Design-Based Research (DBR)***

Interest in design-experiments as contributing to a ‘design science’ in educational research can be traced back to the early 1990s (Brown, 1992; Collins, 1992). Soon thereafter in the mid-1990s, came refinements of the concept for a design-experiments method and the foundation of the National Design Experiment Consortium led by Jan Hawkins. At the end of the 1990s the Design-Based Research Collective was established in 1999 by Christopher Hoadley from which the modern term, DBR, is derived.

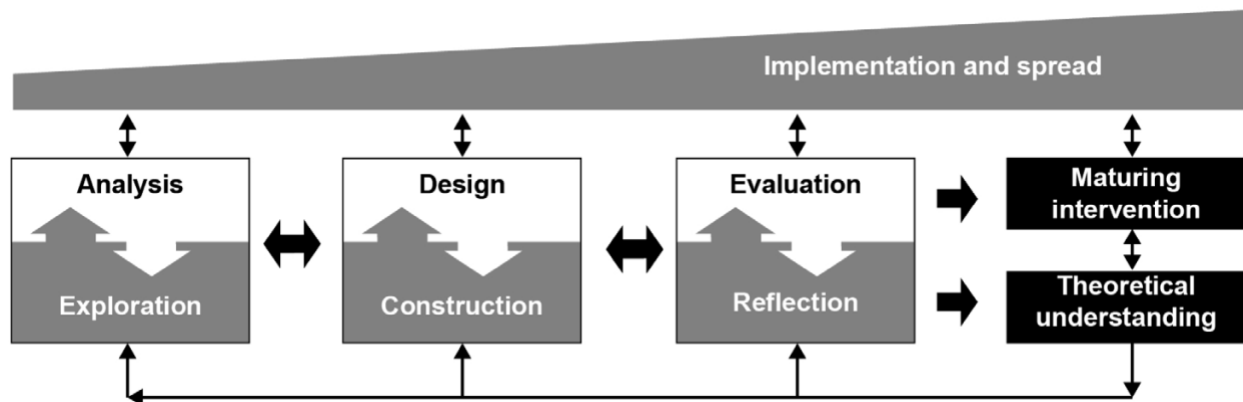
Most language teachers are familiar with action research, which shares many of the same principles as design-based research. Pragmatism is central to both approaches, often employing mixed methods of inquiry to arrive at tangible solutions to educational problems. Normally within action research cycles it is individual teaching practitioners who carry out classroom teaching interventions to observe, record and reflect on the impact of these interventions over time with the aim of informing and improving teaching practice (Reason & Bradbury, 2007). In contrast, what we witness within design-based research is more emphasis commonly placed on educational practitioners working in collaboration with research and design teams (Anderson & Shuttuck, 2012).

Although DBR has sustained great interest from researchers and practitioners in instructional design and educational technology, it is nevertheless a long-term and very resource-intensive exploratory research method with difficult to define goals and outcomes. The literature on DBR attests to “a series of approaches, with the intent of producing new theories, artefacts, and practices” (Barab and 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 at minimum over several weeks but in most cases collected over several months or years as has been the case with this thesis research (Herrington et al., 2007). Knock-on challenges arise with maintaining a collaborative vision and partnership among stakeholders in the research, which in of itself rarely has recourse to sufficient funding to match the years of investment required to sustain the research (Design-Based Research Collective, 2003; Anderson and Shattuck, 2012). Knowing when to stop the research and decide if a designed intervention has succeeded or failed and whether or not it warrants further

investment in research are further challenges with DBR due to a lack of formal criteria and scientific methods to follow (Dede, 2004).

### *Models for understanding and conducting DBR*

McKenney & Reeves (2012) offer a generic model for understanding and conducting DBR in education as pictured in Figure 7. Three key design phases are represented in squares on the model, representing flexible and iterative activities, starting with the initial phase of analysis and exploration, followed by a design and construction phase, and then the final evaluation and reflection phase. The design phases feed into one another and result in outputs of maturing interventions and theoretical understandings as represented by the rectangles in black on the right side of the model. The trapezium at the top of the model represents the gradually increasing implementation and spread of the design research interventions over time as being practice-driven and use-inspired.



*Figure 7* Generic model for conducting design-based research in education. Reprinted from McKenney, S., & Reeves, T. (2012). *Conducting Educational Design Research*. London and New York: Routledge.

Due to the highly iterative nature of DBR interventions in education, McKenney and Reeves (2012) expanded their generic model to include further detail in each of the three key design phases as having sub-components representing different-sized design cycles within each design phase, namely micro-, meso- and macro-cycles as illustrated in Figure 8. For example, the generic design-based research process pictured in Figure 7 is representative of one macro-cycle that would involve numerous micro- and meso-cycles over a long period of time. The refined

model of micro-, meso- and macro-cycles is indicative of the nature of DBR interventions as being comprised of inter-related micro- and meso-cycles taking place simultaneously at each design phase. These cycles lead to greater refinement with theoretical understanding and the maturing of interventions as represented in Figure 7. To provide some foreshadowing of the progressed DBR design interventions that contain micro-, meso- and macro-cycles in relation to Studies 1-3, which will be discussed in more detail in Chapters 3-5, I will briefly outline the exponents of each design phase in the following passages of this chapter sub-section.

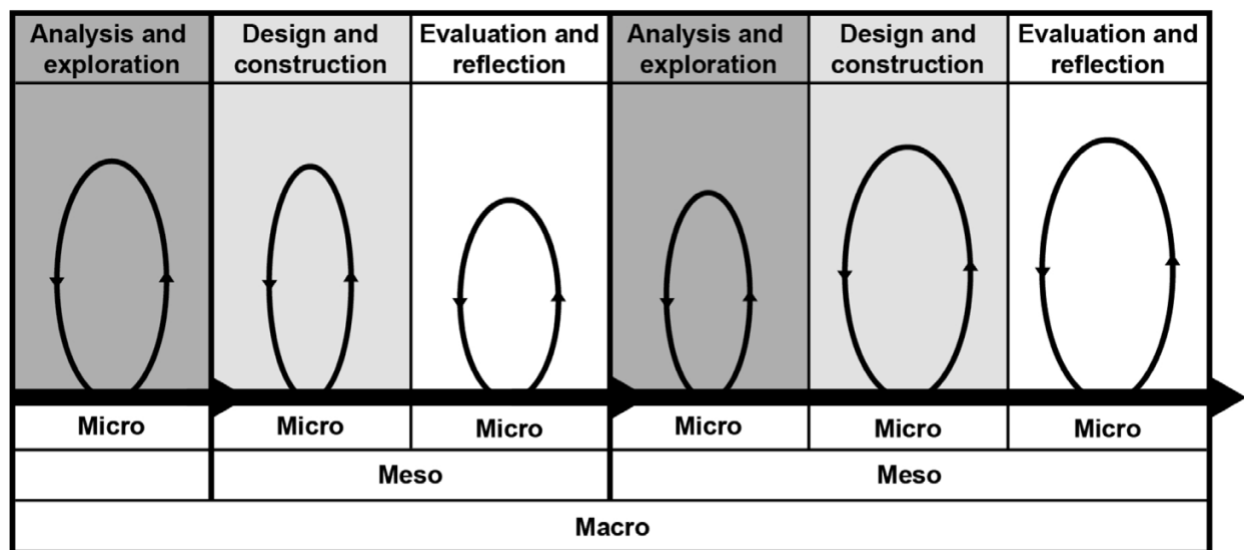


Figure 8 Micro-, meso- and macro-cycles in educational design-based research. Reprinted from McKenney, S., & Reeves, T. (2012). *Conducting Educational Design Research*. London and New York: Routledge.

The first phase of analysis and exploration for all of the open language collections developed in relation to this research (see Table 1) involved defining problems in consultation with knowledge organisations, researchers and knowledge users where I had direct access to research sites and participants. This direct activity is reflected in my consultation with knowledge organisations who manage open access content, in my shared reflections with researchers from computer science, open education and corpus linguistics, and in my shared reflective practice on resources development work with EAP practitioners. Each meso-cycle in this research is related to a particular collection in the FLAX system. This first design phase with micro- and meso-cycles overlapped in real-time across research sites for the co-design and co-development of the

collections. In the case of the MOOC and online networked course contexts where I did not have direct access to the end users for research purposes, I participated in the non-formal courses as a learner to get a better understanding of the learning support designs that the host institutions had devised to augment the MOOC and LMS platform experience. Reviews of the literature into the research from DDL, EAP, corpus linguistics and open online education also informed the research questions and hypotheses formed at this first design phase of analysis and exploration.

The second design phase of design from this research involved the iterative development of open-source software in the FLAX system based on learning design theories, principles and practices drawn from the literature of specific fields (Herrington et al., 2007) in computer science, educational technology, applied corpus linguistics and second language acquisition. For example, the iterative development of the Law Collections in FLAX overlapped in terms of time and implementation of the micro- and meso-cycles included in this design and construction phase of the research into collections building for non-formal online learning.

The third design phase of evaluation and reflection in this research involved the actual and repeated implementation of innovations into the research sites identified in the three studies of this thesis. This phase also includes the reflections and evaluations from stakeholders engaged in the research, and the means for systematic data collection of relevant material (Lincoln & Guba, 1985) over the course of multiple design iterations of interventions that reflect micro- and meso-cycles in the research. Methods for collecting data from stakeholders engaged in the various iterations of the research in different contexts were also devised, refined and implemented at this third design phase with reference to different research methods (qualitative and or quantitative), different modalities (online, face-2-face or blended), and different instruments for collecting perception and performance data (interviews, focus discussions, design diaries, surveys, user query data written to system log files, and student writing). The many iterations with DBR R&D cycles result in a large quantity of data collected which may only lead to small contributions to theory. Dede (2004) has stated that the same results could be achieved by only analysing five percent of the data collected in design research interventions.

The fourth phase includes maturing interventions and theoretical understandings in the form of findings and actual artefacts that may provide solutions to posed problems, and in the generation of design principles that may inform future designs as documented through research outputs. Examples of interventions from this thesis research are the actual open-source software, open

domain-specific language collections and maturing user interface designs based on user evaluations that have been iteratively developed with an eye to making them more accessible and user-friendly to non-specialist users, namely teachers and learners. Theoretical understandings in this research have emerged as design principles for how to scale the open data-driven learning systems developed in this research for greater implementation and spread in higher education. This final phase may also include an awareness of limitations for continuing with different aspects of the research. It may also include an awareness of alternate pathways to lead the research forward in new directions with new design cycles for the development of yet more new systems.

DBR goes hand in hand with pragmatism (Onwuegbuzie & Leech, 2005) and has been supported by the mixed research methods employed within each study (Bereiter, 2002). Drawing on the R&D methods utilised in this research, I will provide an overview of two central iterative macro cycles to the FLAX project for open data-driven language learning systems design in higher education, and their composite meso- and micro-cycles following the framework for educational DBR put forward by McKenney and Reeves (2012) as shown in Figures 7 and 8.

#### *Macro DBR cycle 1: Augmented full-text FLAX corpus design*

The first macro cycle of DBR concerns the on-going development of augmented full-text corpora in the FLAX system. By way of reflecting on the type of corpora being developed at the centre of this research, we have followed recommendations from participating EAP practitioners in this study and recommendations from the literature (Stubbs, 1996; Hyland, 2000) that language should be studied as whole texts. Moving away from the traditional concordancer text analysis interface from the field of corpus linguistics, which only reveals language snippets from complex querying by researchers, the FLAX project has developed simple yet powerful augmented text interfaces for language learners, which present documents in full and are augmented by powerful auxiliary open resources such as Wikipedia and the FLAX LC system.

I will break this macro cycle down into four meso-cycles that correspond with four different collections in FLAX that feature full-text document browsing and wikification affordances in the chronological order that they were developed: [1] the BAWE Collections, [2] the Law Collections, [3] the PhD Abstract Collections, and [4] the FutureLearn MOOC Collections via the F-Lingo Chrome extension by Jemma König. Each meso-cycle contains three categories of



micro cycles for [i] analysis and exploration, [ii] design and construction and [iii] evaluation and reflection. Each micro-cycle speaks to the maturing interventions as represented by each of the collections in FLAX and the theoretical understanding that results from this cumulative design process.

*Meso-cycle 1. FLAX BAWE Collections:* I was engaged in an OER research and academic practice fellowship based at Durham University Language Centre and managed by the Support Centre for Open Resources in Education (SCORE) at the OU in 2011-2012. My work with EAP colleagues at Durham highlighted the need for access to full academic texts, including full texts written by university students, that could be reused in the development of EAP classroom materials and that could be shared as OERs. This need for full texts led to the development of the BAWE collections in FLAX with a design emphasis on displaying full augmented text as shown in Figure 9.

Table 2. *Meso-cycle 1. FLAX BAWE Collections*

<b>Micro-Cycle: Analysis and Exploration</b>	<b>Micro-Cycle: Design and Construction</b>	<b>Micro-Cycle: Evaluation and Reflection</b>
<ul style="list-style-type: none"> <li>▪ Exploring freely available online corpus-based DDL systems with EAP teachers and students at Durham University Language Centre.</li> <li>▪ Consulting the EAP and DDL literature for calls for larger and easy-to-use DDL systems (Boulton, 2013), and the reuse of full-texts in EAP teaching and learning resources development.</li> </ul>	<ul style="list-style-type: none"> <li>▪ Full-text BAWE collections in FLAX, which contain 2860 high-standard student assignments representing different written genre types from across the academy (6 million words) (Nesi and Gardner, 2012).</li> <li>▪ Wikification with the Wikipedia Miner toolkit (Milne &amp; Witten, 2013).</li> </ul>	<ul style="list-style-type: none"> <li>▪ (See Study 1, Chapter 3)</li> <li>▪ Taking the BAWE collections in FLAX around the world in 2012-2013 with my Oxford OER International fellowship showed me how teaching and learning EAP on university language programs in Asia and South America had supplanted the teaching and learning of general conversational English.</li> <li>▪ Reflecting on the design and development process to share</li> </ul>

		with other DDL systems developers.
<b>Maturing intervention and theoretical understanding:</b>		
<p><i>Design principles:</i></p> <ul style="list-style-type: none"> <li>▪ Academic English corpora are increasingly viewed as valuable and desirable to learners and teachers in higher education (Fitzgerald, 2013b).</li> <li>▪ The FLAX design departure away from concordanced interfaces is novel to DDL systems design (Wu &amp; Witten, 2016)</li> <li>▪ The FLAX design departure toward full-text browsing is novel to DDL systems design (Wu &amp; Witten, 2016)</li> <li>▪ The FLAX design departure toward Wikification is proposed as useful for learning related words and topics leading to further open resources e.g. Wikipedia articles, and is novel to DDL systems design (Wu &amp; Witten, 2016)</li> <li>▪ Smaller corpora require augmentation with more powerful corpora e.g. FLAX LC (Widdowson, 2000)</li> </ul>		

The BAWE corpus is managed by the Oxford Text Archive (OTA). Subsequently, I held a UK OER international fellowship with the University of Oxford to promote OpenSpires<sup>36</sup> podcasts, and the work that the FLAX project had done with Oxford-managed corpora (the BAWE and the BNC), to make these collections openly accessible, interactive and pedagogically-focused for data-driven learning with UK and international audiences. A formal request was registered with the OTA to develop the BAWE corpus for non-commercial “research use or educational purposes” (IT Services, University of Oxford, OTA, 2015).

---

<sup>36</sup> <http://openspires.oucs.ox.ac.uk/>

British Academic Written English (Life Sciences)
your name:

About
Search
Browse by Genre
Browse by Discipline
Collocations
Wordlist
LexicalBundles
My Cherry Basket

<-Back to document list
Orthopaedics Patient Portfolio

Original
wordlist
wikify
adjective
noun
verb

Referral information Source of referral and a summary of key information

was admitted to Hospital on the for arthroscopic subacromial decompression of the right shoulder with minor open rotator cuff repair and distal clavicular resection. This was performed on the same day. He has a long-standing history of right shoulder pain and impaired movement. He has suffered from ME for 8 years and as a consequence has had to take early retirement through ill health. History All relevant medical history, co-existing problems, current treatment, significant past medical history and the social and family background. The patient's current presentation and recommendations for treatment. Presenting Complaint.

"Severe" right shoulder pain History of Presenting Complaint. Shoulder Pain.

is a right-handed gentleman with a 16-year history of right shoulder pain. He thinks the pain started as a cumulative result of repetitive heavy straining and lifting whilst doing DIY one weekend at home. Pain was gradual in onset over a period of a week, of moderate intensity and experienced only when lifting the arm out to the side or front. It was focused at the tip of the shoulder and scapula and associated with extensive swelling of the right shoulder going down into the arm. Pain was described as "stabbing" in nature. Since the onset has continued to experience constant "dull, aching" shoulder pain with "progressive worsening in severity". Immediately prior to the current admission intensity was rated at 8/9 out of 10 compared to 5/6 at the time of injury. Pain experienced since the initial episode has radiated from the shoulder to the elbow and hand. He feels pain also radiates upwards causing neck pain and frequent headaches. Pain limits movement of the affected shoulder, this is compounded by stiffness resulting from reduced use over the 16 year period. He is unable to raise his right arm to the side (abduction) by more than approximately 50°, as pain limiting further movement, or to lift his arm above his head or to put it behind his back. Movement in front of his body is less painful but limited to approx. 100°. This makes manual work, carrying shopping and gardening extremely difficult. He also has difficulty brushing his hair and washing his back. Pain is exacerbated by the aforementioned movements but is continually present. Co-codamol and paracetamol provide some relief but this is currently minimal.

first sought medical advice immediately after the original injury. He was told he had "damaged a shoulder ligament" and advised to rest the arm, which provided little relief. Over the past 16 years he has received numerous episodes of physiotherapy, electrotherapy, steroid injections (most recent in 2004) and worn a shoulder sling, which have provided no long-term benefit, "nothing has cured it". In January 2001, he had an arthroscopic washout, which provided both pain relief and improvement in shoulder function for a period of approximately a year.

Figure 9 Full text case study document featuring adjective collocational phrase parsing in the BAWE Life Sciences collection

Meso-cycle 2. FLAX Law Collections: Over the period 2013-2016, my colleagues and I of the FLAX team co-designed and developed various augmented full-text MOOC corpora with universities who had openly licensed their MOOC content with Creative Commons licenses (Wu, Fitzgerald & Witten, 2014; Fitzgerald, Wu, König, Witten & Shaw, forthcoming). The BLARC collection also makes up part of the Law Collections in FLAX (Marín 2014; Marín & Rea, 2014).

Table 3. Meso-cycle 2. FLAX Law Collections

Micro-Cycles: Analysis and Exploration	Micro-Cycles: Design and Construction	Micro-Cycles: Evaluation and Reflection
For MOOC learners, a review of the literature was carried out to identify barriers to successful learning and retention of learner numbers, including language barriers.	Development of full-text MOOC pedagogic collections, and the BLARC corpus (Marín 2014; Marín & Rea, 2014) in FLAX with a focus on domain-specific terminology in the area of legal English.	<ul style="list-style-type: none"> <li>▪(See Study 2, Chapter 4)</li> <li>▪Uptake and evaluation in the MOOC and online networked course contexts with non-formal learners and subject tutors (in the case of CopyrightX with Harvard).</li> <li>▪(See Study 3, Chapter 5)</li> </ul>

		<ul style="list-style-type: none"> <li>▪ Uptake in legal English translation studies with terminology analysis of student writing.</li> </ul>
<b>Maturing intervention and theoretical understanding:</b>		
<p><i>Design Principles:</i></p> <ul style="list-style-type: none"> <li>▪ Greater scalability of Creative Commons-licensed content in the MOOC space is currently unfeasible due to the current business models of mainstream MOOC provision whose Terms and Conditions for All Rights Reserved material default to read-only open access of course content (Study 2, Chapter 4: Fitzgerald, Wu, König, Witten &amp; Shaw, forthcoming)</li> <li>▪ Higher term average usage in student writing was reported as a result of using text and data-enriched MOOC content for reuse in the context of English for Specific Academic Purposes (Fitzgerald, Marin, Wu &amp; Witten, 2017; Marin, Orts &amp; Fitzgerald, 2017)</li> </ul>		

*Meso-cycle 3. FLAX PhD Abstract Collections:* A further OER research fellowship with the Hewlett Foundation-funded OER Research Hub at the OU (2013-2014) included work with EThOS at the British Library and with universities delivering MOOCs. We developed the PhD abstract corpora of 9.8 million words with the British Library and participating EAP teachers and managers from Queen Mary University of London from 2014-2016 (Wu, Fitzgerald, Yu & Witten, 2018). Following the initial exploration with the reuse of open access content that began with the OTA and the BAWE corpus, engagements with further knowledge organisations ensued, including the British Library. I scoped out the abstract metadata of 450,000 PhD theses as being valuable for EAP, which is available via the EThOS toolkit that offers guidance on employing EThOS metadata for “reuse by third parties for not-for-profit purposes” (British Library, n.d.). We discovered a work-around solution utilising TDM for remixing and displaying the full PhD abstract texts due to their status as both content and metadata. However, to go one step further by employing TDM approaches to the full texts of the PhD theses was a step too far due to the mixed provenance in terms of copyright restrictions for the reuse of each doctoral thesis.

Table 4. *Meso-cycle 3. FLAX PhD Abstract Collections*

<b>Micro-Cycle: Analysis and Exploration</b>	<b>Micro-Cycle: Design and Construction</b>	<b>Micro-Cycle: Evaluation and Reflection</b>
--	---	---

Scoping activity with EAP teachers and program managers at Queen Mary led me to contact Sara Gould of EThOS at the British Library to gain access to PhD theses for reuse.	Initially EAP practitioners at Queen Mary developed three micro PhD corpora using EThOS content for the development of interactive game-based collections for use on their pre-sessional programs. Later, the FLAX team developed the more powerful and complete PhD Abstract collections	<ul style="list-style-type: none"> <li>▪(See Study 1, Chapter 3)</li> <li>▪Evaluation and reflection on FLAX collections building by Queen Mary participants in the research.</li> </ul>
<b>Maturing intervention and theoretical understanding:</b>		
<p><i>Design principles:</i></p> <ul style="list-style-type: none"> <li>▪EAP practitioners require time and support with building interactive game-based micro-corpora in FLAX. As a result, we have moved our focus at the FLAX project away from language teachers building their own collections and have re-focused our efforts on building larger more powerful academic English corpora that can be consulted as reference resources via the FLAX website by learners and teachers (Fitzgerald, Wu, &amp; Barge 2014)</li> <li>▪Metadata of academic research publications includes full abstracts, which are useful in the design and development of abstract corpora (Wu, Fitzgerald, Yu &amp; Witten, 2018)</li> </ul>		

The iterative design and evaluation work with the team at Queen Mary had shifted in focus from using full EThOS PhD theses, from three UK universities who had granted the necessary permissions via requests from the British Library, to only using PhD abstracts. It was also decided that the smaller abstract texts better enabled the development of activity-based micro-corpora in FLAX that could be augmented with a much larger Google n-gram corpus in developing automated collocations games for use with the FLAX suite of mobile applications for Android (Wu, Franken & Witten, I.H., 2012; Yu, Wu, Witten & König, 2016) thus avoiding issues with large text scrolling on mobile devices as shown in Figure 10. However, for one of the PhD abstract micro-corpora on water politics and tourism studies where there were not enough abstracts available in these domain areas, the EAP teacher responsible for building this micro-corpus, Chris Mansfield, inadvertently added more abstracts harvested from the entire EThOS repository. When Chris and I presented the micro PhD abstract collections in FLAX at the British Library in June 2015, two of the EThOS curators, Sara Gould and Heather Rosie, were keen to

inform our collaborative project team that the entire dataset of PhD thesis abstracts were considered metadata and therefore available for remixing in the FLAX project. This breakthrough with the reuse of EThOS metadata led to the development of the PhD Abstract Collections in FLAX.

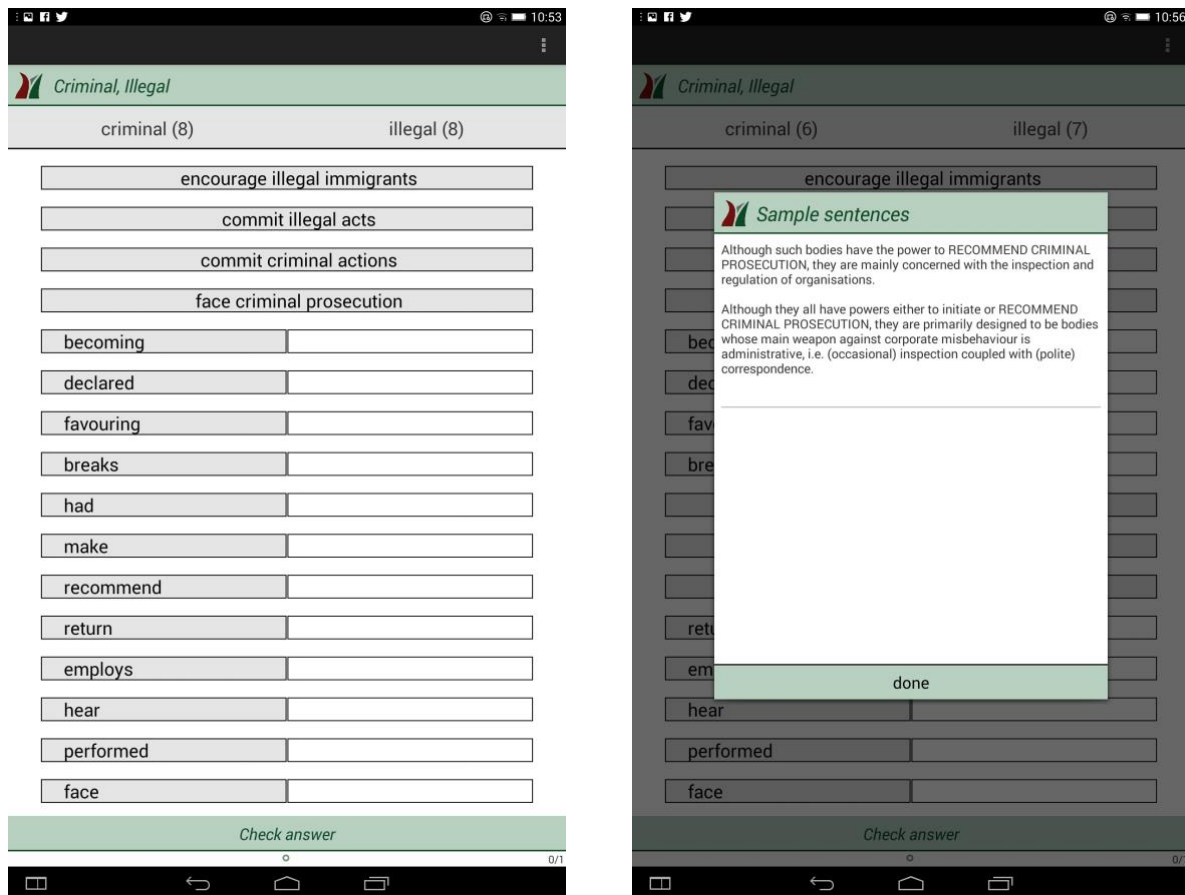


Figure 10 FLAX Related Words Android mobile application featuring an activity from EThOS PhD abstracts collection

#### *Meso-cycle 4. FutureLearn MOOC Collections via the F-Lingo Chrome extension*

##### *Open gratis vs open libre*

The British Library has an Access and Reuse Committee and an established British Library Labs service to encourage research and experimentation with the reuse of their digital collections. The CORE service aggregates unique datasets and provides APIs to conduct research into the reuse of millions of open access publications. In stark contrast, one of the biggest criticisms levelled at the rise of the mainstream MOOC has been the omission of open education policy from commercial

platform providers such as Udacity, Coursera and FutureLearn (see Campbell, 2013) on the reuse of their participating higher education institutions' course content. Instead, what we have witnessed with the big MOOC providers is an apparent emphasis on 'open' as signifying freely and openly accessible resources for philanthropic purposes (open gratis) rather than flexible and customisable resources that can be re-appropriated and retained / revised / remixed / repurposed / redistributed by multiple stakeholders for educational purposes (open libre). Moreover, it is important to note that the majority of MOOC content is licensed All Rights Reserved so this is a real barrier currently where text and data mining reuse of MOOC content in the development of language learning derivatives is concerned. The not-for-profit MOOC provider, edX, has gone some way toward remedying the lack of openness in MOOCs, however, with the development of the edX Creative Commons licensing plugin for their open source platform to enable MOOC host institutions to license their course content openly (Vollmer, 2012; Green, 2015). Nonetheless, the issue of open education policy in MOOCs is an unresolved and ongoing one.

The lack of open education policy in the MOOC space has resulted in knock-on limitations for the development of language learning derivatives from MOOC course content. The theoretical understanding of limitations from this research with MOOCs is coupled with the understanding that the interventions of MOOC language collections developed from openly-licensed course content provided proof of concept for their perceived usefulness by learners. Although this may signify the conclusion of the research of MOOC content with the approach taken in Study 2 of this thesis, current work by the FLAX team has resulted in a radical departure from the FLAX software toward the development of the F-Lingo system by Jemma König that can work around the limitation of All Rights Reserved content by embedding the system in a web browser (with the current iteration as a Chrome extension) with the aim of embedding the system for scaled uptake in MOOC platforms and Learning Management Systems. The F-Lingo system will be introduced in Chapter 6 with reference to current and future work.

Table 5. *Meso-cycle 4. FutureLearn MOOC Collections via the F-Lingo Chrome extension*

<b>Micro-cycle: Analysis and Exploration</b>	<b>Micro-cycle: Design and Construction</b>	<b>Micro-cycle: Evaluation and Reflection</b>
Analysis of barriers to implementing data-driven	Design departure from the FLAX digital library system to	▪(See Chapter 6 for future work)



domain-specific terminology learning support in the MOOC space.	the development of the F-Lingo Chrome extension to use with FutureLearn MOOCs by Jemma König.	<ul style="list-style-type: none"> <li>▪ Doctoral research experiment by Jemma König (forthcoming)</li> <li>▪ carried out with Data Mining FutureLearn MOOCs at the University of Waikato.</li> </ul>
<b>Maturing intervention and theoretical understanding:</b>		
<p><i>Design principles:</i></p> <ul style="list-style-type: none"> <li>▪ Developing integrated language learning support directly into the MOOC platform experience provides a critical advantage over learners having to navigate away from the platform to the FLAX system website (see Study 2, Chapter 4).</li> <li>▪ F-Lingo still requires universities to allow the pre-processing of their course content, so the challenge of reusing copyrighted content still remains. However, this challenge is lessened by the fact that the content will remain in the MOOC space.</li> <li>▪ In order for F-Lingo to be scaled for wider adoption in MOOCs, learning technologists responsible for delivering MOOCs will need to be trained in data-scraping methods to pre-process course content to be traversed by F-Lingo for features of domain-specific terminology (see Chapter 6).</li> </ul>		

### *Macro DBR cycle 2: FLAX Learning Collocations system design*

Anthony (2014) in his keynote address to the Teaching and Language Corpora (TaLC) conference demonstrated the importance of viewing corpora as data (McEnery, Xiao & Tono, 2006; Hunston, 2002; Sinclair, 2004b) by way of providing an overview of the types of tools, many of which are now freely available, and the limitations of those tools, developed so far for uses with language corpora:

The essence of the corpus as against the text is that you do not observe it directly; instead you use tools of indirect observation, like query languages, concordancers, collocators, parsers, and aligners. (Sinclair, 2004b, p. 189)

The FlaxLC system has been designed to mimic the structure of a traditional collocation dictionary after studying the different definitions for collocations in the literature, and investigating the structure, organization, and language items found in traditional collocation dictionaries.



The second macro cycle of DBR concerns the on-going development of the FLAX Learning Collocations (FLAX LC) system design, and the brainchild of Shaoqun Wu (2000). My doctoral research contributions to the FLAX LC system include the scoping out of relevant authentic academic content for the co-design of academic collections that have been added to the FLAX LC where there were none before. I will break this macro cycle down into three meso-cycles that correspond with three different collections in the FLAX LC system which feature affordances to support learner search strategies. The FLAX LC system is linked to all of the collections discussed in the aforementioned macro-cycle of full-text corpus design in the FLAX system. The large databases and novel learning support functions that make up the FLAX LC system serve to boost collocation learning in any FLAX collection, however great or small, by demonstrating how language is used in wider and multiple contexts.

Dedicated learning support in the FLAX LC includes the recent addition of word autocomplete functionality to aid learners with their search queries. Unlike all of the collections presented in the first macro-cycle of design and development from this research, which display full-texts and support browsing strategies, users of the FLAX LC are required to employ search strategies for querying the system. The act of searching requires greater language proficiency in order to be able to formulate queries, so I will speak briefly to the word autocomplete learning support feature in the FLAX LC.

Misspelling is common in search engine queries. What happens when we employ a search engine like Google is that the autocomplete facility compensates for our bad spelling by consulting historical query terms to provide hints while we are typing. However, this approach for reusing historical query terms is not applicable for FLAX LC user queries because learners' language proficiency with vocabulary query items is likely to be more limited; therefore, the misspelling rate would be higher in learners' historical query terms (Wu, Fitzgerald, Yu & Witten, 2019). A dictionary derived from 32,000-word entries extracted from a Wikipedia article corpus of three-billion words were sorted by frequency and inflected forms of a word (e.g. *takes*, *taken*, *taking* for the word *take*). Rare words (i.e. that occur only once in Wikipedia) were omitted to achieve a good user interface response time. Only up to twenty suggestions are given at a time to avoid overwhelming users with too many language choices (see Figure 2).

The corpora in FLAX LC have also been integrated with the Wikipedia corpus and Wikipedia Miner toolkit of machine learned approaches (Milne & Witten, 2013) for the design of additional

learning features in the system, which I will go on to describe in the following sub-section of this chapter.

### *Meso-cycle 1. Wikipedia in the FLAX LC system*











We explored the possibility of using the publicly available and growing Wikipedia corpus of articles to present related words and collocations (Wu, Li, Witten & Yu, 2016). The related words function in the FlaxLC system extends Chen's (2011) idea of retrieving words that are semantically related to the query term. This feature has been designed to help learners expand their word and collocation knowledge, especially in domain-specific areas, or on topics related to what they are studying. First, the best matching Wikipedia article and then the keywords and collocations of that article are retrieved. The collocations are then grouped by the keywords they contain. FlaxLC traverses the Wikipedia corpus with a commonly used metric in information retrieval (called TF-IDF, and described by, for example, Witten, Paynter, Frank, Gutwin & Neville-Manning, 1999). The TF-IDF metric is used to rank words related to the query, so that they can be displayed in descending order of relatedness. Figures 11 and 12 show iterations with the Related Words feature in the FLAX LC for the search term *research*. The positioning and display of the function differ in terms of interface design as shown in Figures 11 and 12 with the latest version in Figure 11 showing the Related Words feature at the top of the web page as a tab in part of a learning support options menu.



Family Words | Synonyms | **Related Words** | Definitions | Related topics

research hypothesis scientific empirical method prediction academic researcher hourglass quantitative journal knowledge criticism evidence vary outcome  
search matter observation publication apply historical definition conceptual subject electronic data necessary depend basic test freely systematic guideline  
analysis accuracy step social information disorganise

>>>> more

**research used as a noun** | **research used as a verb**

research + noun	 research results	3310	 research center	2591
	 research environment	2586	 research visits	1768
	 research assistance	1308	 research question	1186
	 research network	1050	 research support	870
	 research project	840	 research institutes	759
>>> more				

 research environment	2586	 research environments	8
--	------	---	---

adjective + r

- Second, this technology allows truly vast **research environments**.
- The importance of universities as **research environments** was ranked in fourth position (OECD 1982).
- Institutions of higher education have very different teaching and **research environments** depending on their mission.
- Their findings suggest that certain **research environments** and infrastructures can significantly improve one's research standing. A
- If desired, designers can fix the code so that two **research environments** are exactly the same, down to every leaf on every tree.
- Using the already defined hypernodes it becomes possible to not only connect single patents by co-inventors but **research environments** by peers.
- In essence, the underlying logic for this segmentation is that the institutional web of informal norms, formal rules and enforcement characteristics affect the educational and **research environments**.
- Even so, considerable expertise is required to assess the accuracy of data and metadata in these **research environments**, as minute errors in calibration can influence analysis and interpretation significantly.

2537  
2537  
497  
743  
599  
more  
249

research + preposition + noun



 research institute in economics	233	 avenue for future research	157
---	-----	--	-----

Figure 11 New interface for FLAX LC Related Collocations function

This design modification was to increase visibility of the various learning support features in the system. Previously with the old interface design shown in Figure 12, although more aesthetically pleasing, required users to scroll to the bottom of the web page. We believe this may have been reducing the uptake of this and other learning support features based on reviewing user query pathway data from system log files.

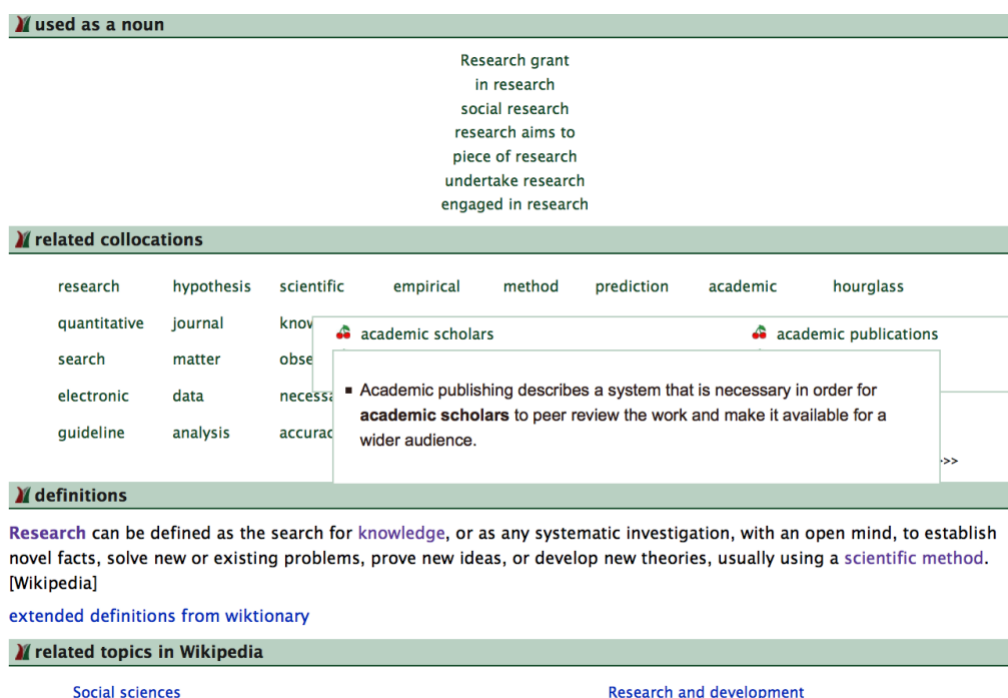


Figure 12 Old interface for FLAX LC Related Words function

Table 6. Meso-cycle 1. Wikipedia in the FLAX LC system

Micro-cycle: Analysis and Exploration	Micro-cycle: Design and Construction	Micro-cycle: Evaluation and Reflection
Analysis and exploration of existing tools and linked open data-sets that could be incorporated into the FLAX LC, including OpenNLP, WordNet, Wikipedia Minter toolkit, dictionary of terms for	Developing the Wikipedia database in FLAX LC to include learning support functions, including Autocomplete search, Definitions, Related Words, Family Words, Related Words –	<ul style="list-style-type: none"> <li>▪ (See Studies 1 &amp; 2, Chapters 3 &amp; 4)</li> <li>▪ OER case study with Durham EAP teachers and pilot study with EAP learners on the affordances of using FLAX LC related words function to support lexical range of</li> </ul>

autocomplete population to support user queries.		domain-specific terminology in essay writing. ■ Evaluation by online non-formal learners and tutors of links from MOOC corpora to FLAX LC with Wikipedia corpus as default.
<b>Maturing intervention and theoretical understanding:</b>		
<i>Design principles:</i> <ul style="list-style-type: none"> <li>■ Search strategies in language learners require additional learning support (Wu, 2010; Wu, Franken &amp; Witten, 2010; Wu, Witten &amp; Franken, 2010; Wu, Franken &amp; Witten 2012; Franken, 2014).</li> <li>■ Open linked data and open tools provide novel design departures for building DDL systems (Wu, Li, Witten &amp; Yu, 2016)</li> <li>■ Larger collections e.g. the ACE collections are more suitable to FLAX LC whereby language snippets only are presented rather than full texts (Wu, Fitzgerald, Yu &amp; Witten, 2019)</li> <li>■ Resistance to Wikipedia as a corpus appears to be diminishing in DDL systems development by researchers in the field (see BYU Wikipedia Corpus by Mark Davies<sup>37</sup>).</li> <li>■ Related Words feature is perceived as highly relevant for domain-specific terms and concepts for raising learner awareness of lexical range. (Fitzgerald, 2013a; Fitzgerald, 2013b)</li> </ul>		

*Meso-cycle 2. The BAWE corpus in the FLAX LC system:*

The BAWE corpus was added to the FLAX LC in 2012 and despite its small size system log data indicates that academic English queries are frequent due to the addition of this corpus (Wu, Fitzgerald, Yu & Witten, 2019). However, we have recently replaced the BAWE corpus with the new ACE collections which are far larger and more powerful academic English corpora, which I will discuss in the next sub-section of this chapter.

Table 7. *Meso-cycle 2. The BAWE corpus in the FLAX LC system*

<b>Micro-cycle:</b> <b>Analysis and Exploration</b>	<b>Micro-cycle:</b> <b>Design and Construction</b>	<b>Micro-cycle:</b> <b>Evaluation and Reflection</b>
--	---	---

<sup>37</sup> <https://www.english-corpora.org/wiki/>

Exploration with EAP teachers and learners that determined there was a need for an academic English corpus to be added to the FLAX LC.	The BAWE corpus was added to the FLAX LC at the same time the full-text BAWE collections in FLAX were developed in 2012.	<ul style="list-style-type: none"> <li>▪(See Study 1, Chapter 3)</li> <li>▪Evaluations on the addition of the BAWE corpus in the FLAX LC were carried out with EAP practitioners around the world during my OER International fellowship with Oxford.</li> <li>▪Analyses of user query data from the FLAX LC BAWE corpus were carried out over the period of one year.</li> </ul>
<b>Maturing intervention and theoretical understanding:</b>		
<p><i>Design principles:</i></p> <ul style="list-style-type: none"> <li>▪Due to the addition of the BAWE corpus in the FLAX LC, the system was deemed more valuable as an academic collocations consultation resource by EAP practitioners (Fitzgerald, 2013b).</li> <li>▪Although the BAWE corpus in the FLAX LC has proven to be popular it has been replaced by the new ACE collections to better support the increased demand for academic English collocation learning support (Wu, Fitzgerald, Yu &amp; Witten, 2019).</li> </ul>		

*Meso-cycle 3. The ACE corpora in the FLAX LC system:*

The British Library directed me to the CORE (COnnected Repositories) open access harvesting and aggregation service at the OU where they are developing useful services and APIs for working with open access data from upwards of 135 million open access articles. CORE's mission (Knoth & Zdrahal, 2012) is perhaps the closest yet to the original Budapest Open Access Initiative (BOAI) definition, where they “offer seamless access to millions of open access research papers, enrich the collected data for text-mining and provide unique services to the research community.” (CORE, n.d.). Our most recent collections development work with CORE has resulted in the Academic Collocations in English (ACE) collections in FLAX.

Table 8. *Meso-cycle 3. The ACE corpora in the FLAX LC system*

<b>Micro-cycle: Analysis and Exploration</b>	<b>Micro-cycle: Design and Construction</b>	<b>Micro-cycle: Evaluation and Reflection</b>
--	---	---

Following the work with EThOS at the British Library where we reached the limit of full-text PhD thesis reuse (abstracts only), we scoped out further content and metadata via the CORE services at the OU.	<ul style="list-style-type: none"> <li>▪ Dirty collections due to the high amount of OCR content in CORE.</li> <li>▪ Huge collections which are more suitable to the FLAX LC system whereby language snippets only are presented rather than full texts.</li> </ul>	<ul style="list-style-type: none"> <li>▪ (See Study 1, Chapter 3 &amp; Chapter 6)</li> <li>▪ Cleaning up the ACE corpora will require more development work.</li> <li>▪ The work with the ACE collections in FLAX is part of my current and future postdoctoral research.</li> </ul>
<b>Maturing intervention and theoretical understanding:</b>		
<p><i>Design principles:</i></p> <ul style="list-style-type: none"> <li>▪ The larger ACE collections, which have been derived from the content and metadata of an aggregation of 135 million open access journal articles are more powerful and suitable for domain-specific term querying. However, they are also messier in terms of bugs appearing in the collections' language output.</li> <li>▪ Design challenges still remain with the CORE datasets being comprised of completely unstructured data with some OCR formatted open access content present.</li> </ul>		

### ***Overview of research sites and original contributions to knowledge***

This doctoral research presents three empirical design research intervention studies with the FLAX project that report on the processes of iteratively designing, developing, implementing, and evaluating new open data-driven language learning systems with participating knowledge organisations, researchers and knowledge users. With the use of academic English language corpora derived from open access content as a uniting factor in the three studies presented herein, this thesis aims to advance the fields of applied corpus linguistics and educational technology by demonstrating how traditional tools for querying language corpora can be improved upon and scaled by adopting an open infrastructure in collaborative data-driven language learning systems development; where the focus with end user designs has been deliberately shifted away from research applications toward pedagogical applications in both formal and non-formal higher education contexts.

In response to the deficit in accessible open language corpora and the user-friendly tools needed to analyse and exploit them for DDL, the on-going design research interventions

presented herein with the FLAX project have made an original contribution to knowledge by proposing a new paradigm for designing open data-driven language learning systems in higher education. This research has emphasised an infrastructure of open educational practices that are pushing at the parameters of policy for the reuse of research and pedagogic content in the development of automated open DDL systems for support with learning features of domain-specific terminology in formal and non-formal higher education contexts. This research has engaged knowledge organisations such as libraries, archives, aggregation services, and universities working with MOOC providers, all of which are providing increased open access to a tranche of invaluable linguistic data for teaching and learning features of domain-specific terminology (Wu, Fitzgerald & Witten, 2014; Fitzgerald, Marin, Wu & Witten, 2017; Marin, Ortis Llopaz & Fitzgerald, 2017; Wu, Fitzgerald, Witten & Yu, 2018; Fitzgerald, Wu, König, Witten & Shaw, forthcoming).

Study 1 of this thesis provides a qualitative inquiry into reflections from the different stakeholder groups engaged in participatory design ethnography research interventions with the FLAX project. The research in Study 1 is characterised by emergent goals that have arisen from design cycles for employing TDM and NLP methods for the reuse and remix of open access linguistic content in the development, enactment and redevelopment of open data-driven learning systems for academic English. This research is guided by the vision of the as-yet-unrealised potential for scaling the reuse of open access artefacts of the academy for the development and deployment of data-driven language learning systems across all modalities of higher education provision: formal, non-formal and informal. The research in Study 1 has pushed at the parameters of policies adopted by knowledge organisations to tease out affordances and barriers as perceived by stakeholders in this research with regards to the reuse and remix of open access content for non-commercial research and educational purposes. The design research interventions and findings captured in Study 1 are further supported and evolved by the mixed methods of research inquiry employed in Studies 2 and 3 of this thesis.

Study 2 fills an existing research gap in the DDL literature by reporting on data-driven language support in non-formal higher education learning contexts (MOOCs). In a mixed methods study, system log data is triangulated with user studies by way of self-reported learner and teacher perceptions from survey perception data. An evaluation of the input enrichment and enhancement of MOOC pedagogic content to create corpora that have been developed to support

domain-specific terminology learning in minimally-guided online learning contexts with first and second language users is presented for discussion in relation to Study 2 of this thesis. In online learning, and in MOOC provision specifically, the digital library affordances in FLAX of being able to search and browse through course content that has been augmented with further open resources (e.g. Wikipedia, documents in the public domain, the FLAX collocations database etc.) serve to enhance the functionality of the typical experience with the closed LMS and mainstream MOOC platforms. MOOC platform designs have drawn heavily on the content and learning management designs of the standard LMS, essentially the same LMS designs that have steered the educational technology vendor industry for decades (Watters, 2016).

Efficacy with open educational resources (OERs) from the digital commons has focused almost exclusively in the literature on their cost-saving value. Study 3 in this thesis offers methods for digitally enhancing OERs to render them linguistically accessible in addition to being accessible in terms of removing or reducing cost barriers. A quasi-experimental empirical intervention and quantitative analysis of learner performance data from the context of formal language and translation studies at a university in Spain is presented for Study 3. Student informants were divided into two groups: an experimental group and a control group. The experimental group was assigned to the exclusive use of the English Common Law MOOC pedagogic corpus in FLAX for support in completing an essay from a series of assigned topics on the English common law system. The control group were assigned the same essay topics and were advised to use any information source from the Internet to complete the essay assignment. Results from Study 3 indicate higher levels of implementation of domain-specific terminology in the essays of the experimental group than in the essays of the control group. These findings have pedagogic implications for second language writing for academic and professional purposes where OERs have been enhanced by TDM and NLP methods, resulting in increased awareness in learners for domain-specific term use in course communications and assessments.



## ***Introduction to Study 1***

In light of the current digital era, copyright has become increasingly viewed by many actors in the various open movements as a pre-digital tool, at times wielded bluntly against innovation and the public good for the benefit of protecting publishers' revenues (Okerson, 1991; Willinsky, 2002; Tennant, et al., 2016). The simple act of downloading an article to read it is an act of copying. Digital capabilities for reusing digital content therefore make it very easy to breach copyright. A moral distinction has been drawn by prominent Internet activists whereby breaching copyright and thereby breaking the law is viewed as technically illegal but not immoral in advancing the cause of the open access movement. In legal philosophy such an act would be considered as *mala prohibita* compared with those acts which are considered *mala in se*, which translates from the Latin as "bad in themselves".

Aaron Swartz of the early guerilla open access manifesto (Swartz, 2008) who systematically downloaded hundreds of thousands of J-STOR articles, and Alexandra Elbekyan of Sci-Hub<sup>38</sup> who provides access to millions of paywalled open access journals and books, are two renowned open access activists who have been charged with wire fraud, computer fraud and abuse, and copyright infringement. Both have paid a high personal price: with Swartz's arrest in 2011 and the threat of a maximum prison sentence of 35 years and a \$1 million fine resulting in his suicide in 2013; and Elbekyan's current life in hiding at the time of writing this thesis. In both cases, the technical expertise of Swartz and Elbekyan has outstripped the paywall systems put in place from commercial academic publishers. The law-breaking side of open access has been seen as both helping and hindering the movement, however. Appendix A provides a more in-depth overview of major historical milestones in the progress of the open access movement and open access publishing.

The networked course, CopyrightX, from Harvard Law School and the Berkman Klein Center for Internet & Society features in the FLAX project's research into automated language support in non-formal learning. The following lecture excerpt from Professor Fisher of CopyrightX reflects on the power of criminal sanctions as they correspond to copyright law and Swartz's legal nightmare following the severity of charges he faced in response to his open access activism that led to his suicide:

---

<sup>38</sup> <http://sci-hub.tw/>

In short, the methods that Swartz chose to pursue his vision may well have been wrong. But there's a big difference between misguided idealism and the sort of self-serving piracy at which the criminal statutes are primarily aimed. Perhaps some sort of criminal penalty was warranted in this case, perhaps a deferred prosecution agreement, which would have been effective in preventing Swartz from engaging in similar conduct in the future. Perhaps. But certainly not six months in jail. In short, the prosecutors in this case failed to exercise their power wisely. I know and respect one of those prosecutors. He's not a cruel person. But he and his colleagues acted irresponsibly, and the result was tragedy. From that tragedy, at least two lessons can be drawn.

First, criminal sanctions are both formidable and dangerous. They have important social functions, but their power makes them risky. The hazard that they will be imposed in appropriate circumstances is exacerbated by the large and increasing diversity of the sets of circumstances and the kinds of technologies implicated by copyright law and the kinds of activities that may constitute copyright infringement. It's impossible for legislators to anticipate all of those circumstances and to differentiate them on the basis of the severity of the harms they threaten and, consequently, the severity of the sanctions they merit. It's thus imperative that the people who control the machinery of the criminal law exercise their power sensitively and wisely.

The second, broader point is that the copyright system as a whole is an extraordinarily complex and powerful machine. As I hope you now see, it affects myriad dimensions of the global economy and culture. It seeks simultaneously to advance many different social goals and to protect many different rights and freedoms, some of which are in tension. Effectively operating a machine this complex and important requires care and, again, wisdom. When tuned intelligently and deployed thoughtfully, copyright has enormous and growing benefits. If it is out of tune or deployed thoughtlessly, it can cause great harm. My ambition, in this lecture series, has been to provide you the information and analytical tools you need not just to understand the copyright system as it currently exists but to participate in the ongoing project of adapting that machine to deal responsibly with changing social and cultural circumstances. I hope you have found the lectures helpful in this regard. Thank you for your patience and attention. (Fisher, 2014a)

Many of the corpora in this study have been derived from content created in the UK. It is important to note that my PhD research has not only benefitted from but has been sustained as a direct result of innovative reforms in UK copyright law. The Hargreaves independent assessment and review of the UK's legal framework for intellectual property rights, commissioned by David

Cameron's government in 2010, was preceded by six such reviews conducted in the space of four years none of which had resulted in any significant reforms where copyright law was concerned (Edwards et al., 2012). Nonetheless, in June 2014 a significant amendment to UK copyright law deemed a limitation and exception would follow the recommendation of the Hargreaves review to allow TDM of copyrighted content for non-commercial research purposes.

The received climate in UK higher education at the time of conducting this research followed on the heels of the so-called Academic Spring with the growing online Cost of Knowledge campaign that led to the Elsevier boycott in 2012 of “thousands of researchers complaining of profit taking by scientific journals at their expense” (Epstein, 2012), and signing a declaration not to publish or engage in peer-review and editing with any Elsevier outlet. Identified as “the worst offender” by many mathematicians (Cost of Knowledge, n.d.), the Elsevier boycott and Cost of Knowledge protest campaign was preceded by nine mathematicians at the University of Oxford who resigned in 2006 from the editorial board of the Elsevier journal, *Topology*, in protest of Elsevier's publishing and pricing policies as being damaging to the mathematical research community (Shapiro, 2006). The academic research community continues to push and renegotiate terms for scholarly publishing with commercial publishers, Elsevier being the largest commercial publisher of scholarly journals. Recently, Germany, Sweden, Peru and Taiwan have declared their countries as No Elsevier Deal zones.

The University of California in the US declared in February 2019 that they had reached a similar impasse with Elsevier in trying to seek:

... sustainable cost controls as well as a novel transformative agreement in which our Elsevier authors would retain their copyrights, their articles would become completely and immediately open access, and the payments for open access publishing would offset our Elsevier subscription expenditures (University of California Academic Senate, 2019).

Unable to move beyond the impasse with Elsevier, the University of California decided to terminate all journal subscriptions with the publisher. Jeffrey MacKie-Mason, university librarian and economics professor at UC Berkeley, and co-chair of the University of California's negotiation team commented that the “prices of scientific journals now are so high that not a single university in the U.S. — not the University of California, not Harvard, no institution — can afford to subscribe to them all” (University of California Office of the President, 2019).

### *Open access as the content reuse default in higher education*

Attempts to define openness have been numerous as trends in openness have been observed in a wide number of sectors, including government, research, education, publishing, software, standards, and services. Tensions between stakeholders are changing relationships in all of these sectors as social, economic and legal factors are taken into account for understanding the impact and reach of openness and the growth of the commons paradigm (Bollier, 2007; Benkler, 2007; Kelty, 2008). Richard Stallman's (2002) famous distinction from the "free software" movement that open is more akin to free speech than free beer is perhaps one of the most enduring understandings of openness where greater success can be observed with policies and services for the reuse of open access content and open data in research and with the reuse of source code in open-source software in industry. Far less success can be observed with open policy for the reuse of pedagogic content in education, however (Weller, 2015). Read-only open access has become the content reuse default in higher education with free rather than open courses in, for example, the MOOC space, and with read-only access of research articles and books in digital format. The research presented in this next chapter points to the reuse potential that lies within TDM and NLP approaches for linguistically enhancing research and pedagogic content so that it can be searched, browsed and augmented with further open resources to support language learning for specific academic and professional purposes. Formatting issues do continue to present problems, however, with the greater amount of research publications currently being in PDF format rather than the preferred XML format (Extensible Markup Language), thereby hampering TDM technologies from 'seeing':

... most of the literature at the moment. Access to abstracts and bibliographic details is not enough: these tools need to be able to 'read' the full text of a research article, including any data within it and supporting it. (Swan, 2012, p. 17).

### Chapter 3: Study 1

## Reflections on remixing open access content for data-driven language learning systems design in higher education

### Abstract

This qualitative study mines the concept of open educational systems and practices, which have unique characteristics and challenges with regards to diffusion, uptake and integration.

Reflections spanning 2012-2019 will be presented from an ongoing multi-site design-based research study with the open source FLAX project (Flexible Language Acquisition [flax.nzdl.org](http://flax.nzdl.org)) into design and dissemination considerations for remixing domain-specific open access content. The successive design iterations carried out over the course of this research have resulted in an automated data-driven corpus-based system for applications with learning aspects of domain-specific terminology in formal and non-formal higher education. Primary stakeholders in the research collaboration include:

*Knowledge organisations* that provide open access to content – libraries and archives including the British Library and the Oxford Text Archive, universities in collaboration with MOOC providers, and the CORE (COncnecting REpositories) open access aggregation service at the UK Open University;

*Researchers* who mine and remix content into corpora and open data-driven language learning systems – converging from the fields of open education, computer science, and applied corpus linguistics;

*Knowledge users* who reuse and remix content into open educational resources (OER) for blended learning – English for Academic Purposes (EAP) practitioners from university language centres.

Automated content analysis (ACA) was carried out on a corpus of interview and focus-discussion data with the three stakeholder groups in this research. Themes arising from the ACA point to affordances as well as barriers with the adoption of open policies and practices for remixing open access content for data-driven language learning applications in higher education against the backdrop of different business models and cultural practices present within participating knowledge organisations.

**Keywords:** automated content analysis; blended learning; design-based research; design ethnography; English for academic purposes; massive open online courses (MOOCs); open access; open educational practices; open educational resources

## ***Introduction***

The story of the research presented in this chapter was made possible due to developments with the open access movement, which in itself is intrinsically tied to developments with the Internet and online publishing. From the 1990s onward, the culmination of an old tradition wherein researchers and scholars engage in peer-review and publish in scholarly journals without payment was converging with a new technology, the Internet (Laakso et al., 2011). These two phenomena would coalesce in a defining moment in 2002, with the coining of the term “open access” as it 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).

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 lecture and course reading content from the world's leading universities and institutions with an expanding provision in MOOCs. A further example is open access to a growing corpus of over 450,00 PhD theses from universities across the UK with the British Library's Electronic Theses Online Service. 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 reuse 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. This chapter will take a look behind the scenes, as it were, to explore which openings in the research and development journey enabled the collaboration with the FLAX project to advance, and which roadblocks needed careful navigation to keep the collaboration with stakeholders moving forward.

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

For this project, the first author has scoped out and thrown a lasso around 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. Particularly at a time when the proliferation of generic EAP teaching and learning resources from commercial English language education publishers is at an all-time high. In this chapter, we will share reflections on our work with 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 with 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 reuse of digital research content following an independent government report (Hargreaves, 2011).

We will discuss the perceived value that EAP researchers, teachers and managers place on the efficacy of utilising authentic 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 reuse of digital open access content and collections for non-commercial research and education purposes.

With the FLAX project, we have placed particular emphasis on co-designing and co-creating a language learning system for pedagogic purposes rather than for corpus linguistics research purposes. Drawing on the concept of knowledge mobilization (Levin, 2011) our goal is to engage relevant stakeholders in moving available knowledge from research in corpus linguistics, computer science (NLP and TDM), and open education toward 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 with the FLAX project.

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?

### ***Research materials***

Intermediaries working in knowledge organisations have acted as brokers and OER champions in this research by way of creating access to knowledge artefacts that are valued for reuse in EAP via initiatives in open access policy and reforms in copyright law. Table 1 in the previous chapter provides an overview of our work to date, and 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 project. Although the findings from this research are tied to issues with designing and developing open access content into data-driven learning systems, wider issues pertaining to blended learning vis-à-vis the reuse and remix of open access content in language materials development practices will also be discussed as they apply to both modalities of blended learning: classroom teaching and online learning.



### *Open data in Computer Assisted Language Learning (CALL)*

Colpaert (2016) distinguishes between different uses for data in CALL as falling into two main categories depending on divergent goals for reuse: 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, while 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. The reuse of data in CALL is a nascent and under-researched area in the field, and the XIXth International CALL Research Conference in Bruges in 2018 was dedicated to exploring this theme of data reuse in language education.

### ***Research methods***

The first author will draw on design principles from her direct engagement and placements with stakeholders in the research using multiple methods to collect a variety of data types (Kuper, Lingard, et al., 2008; O'Brien et al., 2014; Santiago-Delefosse et al., 2016). Methods for collecting data from different participant groups in different locations (Santiago-Delefosse et al., 2016) over a period of years included: focus-discussions, interviews, and email exchanges stemming from project meetings on observations and evaluations shared in this situated research that comprise a corpus of just over 50,000 words. Automated content analysis 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 different 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 in this qualitative research to create opportunities for participants to comment on transcripts and emerging findings, and to confirm thematic and conceptual findings in the datasets as they pertain to reflections on the iterative design processes for designing open data-driven systems for academic English (Elliott et al., 1999; Herrington et al., 2007; O'Brien et al., 2014; Santiago-Delefosse et al., 2016; Tong et al., 2007).

### *Design-Based Research in the context of Design Ethnography*

Barab and Squire define design-based research (DBR) as:

... not so much an approach as it is a series of approaches, with the intent of producing new theories, artefacts, and practices that account for and potentially impact learning and teaching in naturalistic settings. (2004, p.2)

With a discernible amplification of the educational research process, DBR involves collaboration between researchers and participants (Anderson & Shattuck, 2012; Cobb et al., 2003) engaged in design and evaluation iterations of multiple research interventions rather than a single intervention carried out by an individual researcher (Design-Based Research Collective, 2003; Amiel & Reeves, 2008; Anderson & Shattuck, 2012). Design ethnography is increasingly carried out in educational settings where there are multiple stakeholders involved in satisfying critical social and organisational requirements for the success of systems design (Crabtree, Rouncefield & Tolmie, 2012), and where it is necessary to explore in whose interests the designer anthropologist operates to navigate the perceived openings and closings that determine the course of the design research (Bell, 2004).

### ***Results and Analysis***

In this section, we look through the analytical lens offered by ACA at key themes and the concepts that make up these themes from each of the three participant groups. 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. Where we present a summary of results from all three sub datasets, themes and concepts will be italicised.

#### ***Automated Content Analysis (ACA)***

Our reasons for employing the Leximancer ACA software to analyse the qualitative datasets were two-fold: to increase validity and to visualise the lexical co-occurrence 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 designed 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. Conceptual analysis is concerned with measuring the presence and frequency of concepts in a document set by extracting words, phrases or collections of words that represent a concept. Relational analysis is concerned with measuring the co-occurrence of concepts within a document set, extracting these co-occurring concepts and visualising them 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, 2000b, 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. As shown in Figure 13, 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 inter-related 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 being 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 of the key themes (Campbell, Pitt, Parent, & Berthon, 2011; 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 on-going design-based research with the FLAX project and will provide a basis for the discussion section of this chapter.

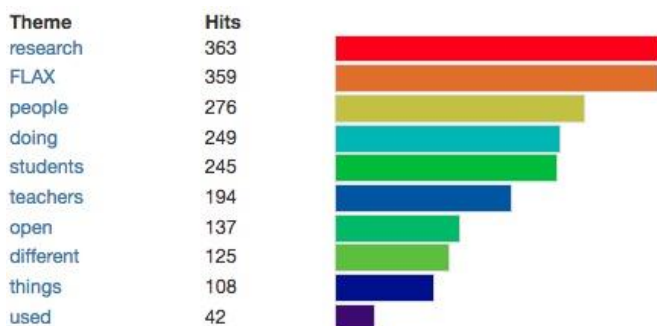
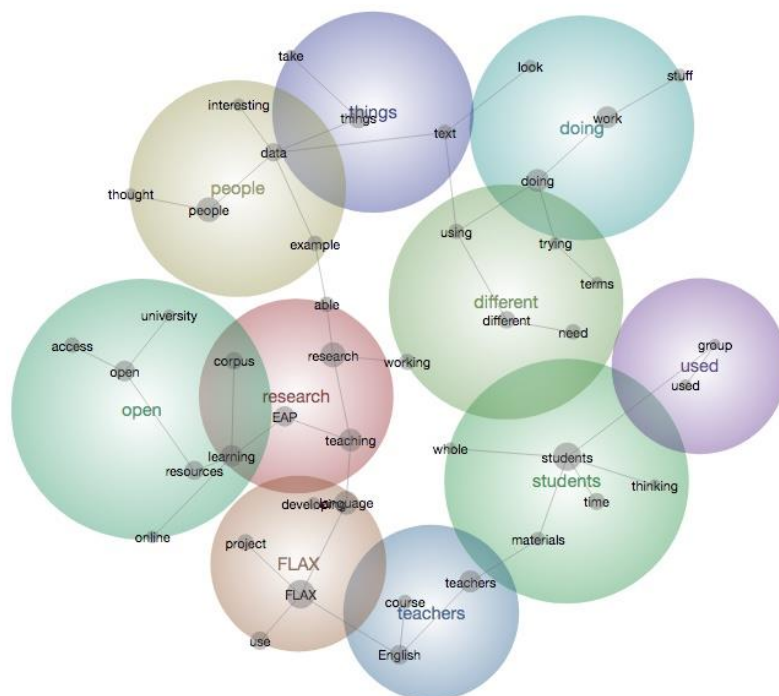
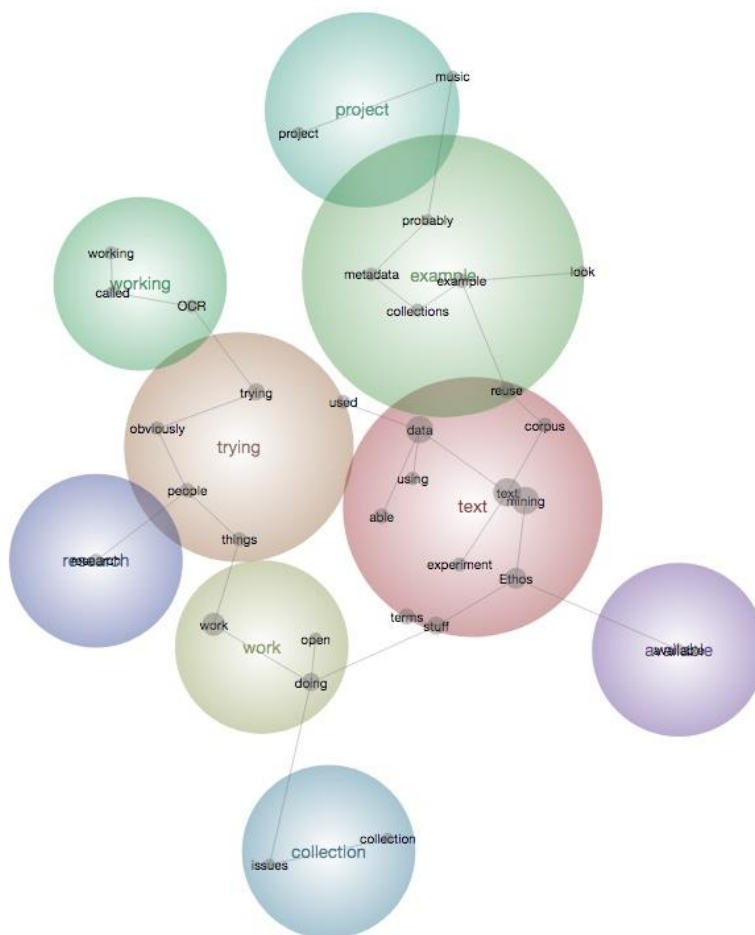


Figure 13 Concept map and key derived from automated content analysis of the complete qualitative dataset

### Knowledge organisations

The Leximancer analysis of data from the knowledge organisations group reveals *text* as the major theme as is indicated by the red thematic circle on the heat concept map and corresponding bar chart in Figure 14. 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

*reuse*, and what you are *able* to do when *using* texts with *text* and *data mining*. The second most prominent theme is *work* with concepts therein 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 and nearby 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*. *Reuse* is the concept shared between the overlapping *text* and *example* themes. Also apparent in the *example* theme are the key interlinked concepts of *example*, *collections* and *metadata* for what can *probably* be *looked* at with respects to research and development that focus on the *reuse* 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 reuse in this research with knowledge organisations.



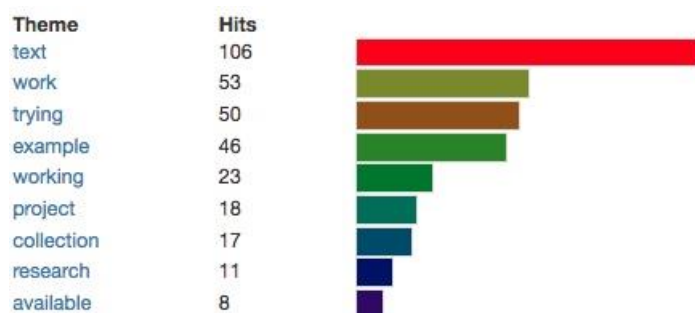


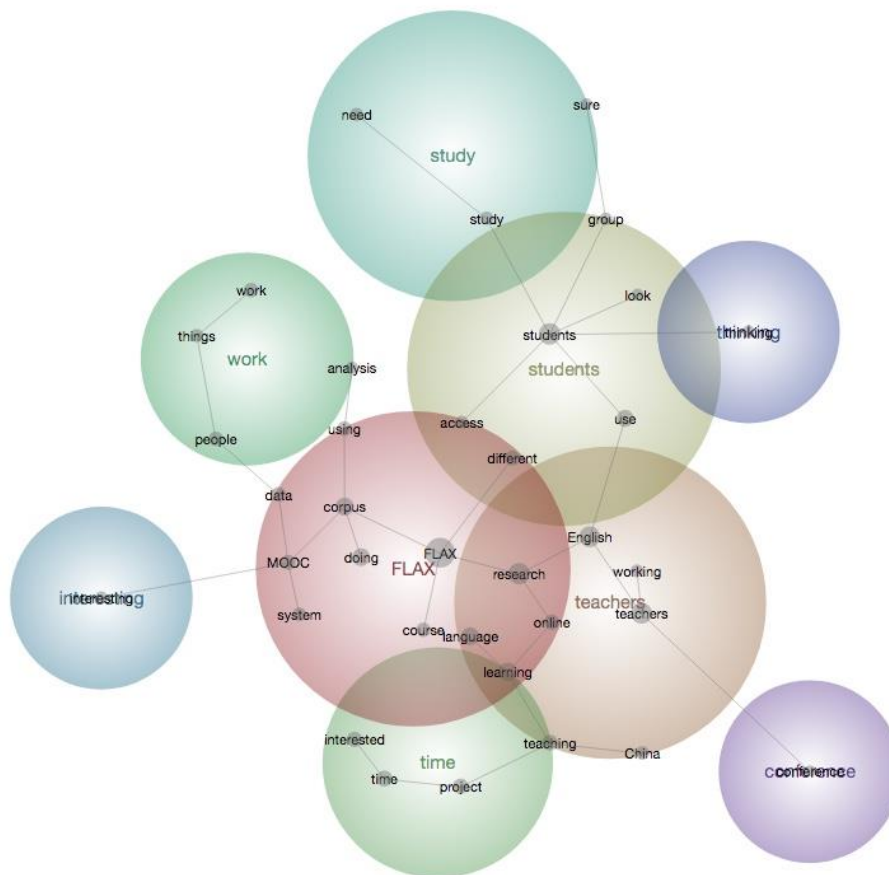
Figure 14 Concept map and key derived from automated content analysis of the knowledge organisations' sub-dataset

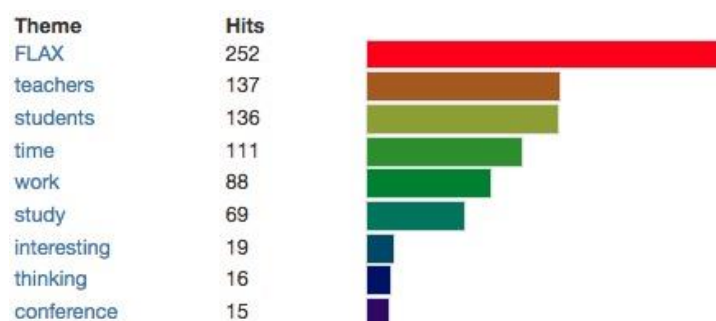
### Researchers

We now turn to interview data between the first author and two further researchers who have worked with the FLAX project. The first researcher interviewed was Maria José Marín, 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 and were made available with an open access government licence from the British and Irish Legal Institute (BAILI). Maria José Marín later worked with the first author on a reuse study with the English Common Law MOOC collection in FLAX for uptake with legal English translation students at her university in Spain, which is the basis for Study 3 in this thesis. The second researcher interviewed was Liang Li, who has carried out doctoral research into lexical bundles with the FLAX project (Li, Franken & Wu, 2017) with a particular focus on the Chinese and New Zealand EAP contexts.

When we look at the Leximancer concept map in Figure 15 for the researcher group, of note are four prominent and overlapping themes: *FLAX*, *students*, *teachers*, and *time*. What is more, 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 provide a summary of the findings from these concepts that appear within the overlapping thematic circles on the heat map, and 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 on the concept map, 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, which appears in all four overlapping themes on the concept map, 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 *use studies* on the uptake of *FLAX*. In addition, the issue of *access* is further expressed with 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.





*Figure 15* Concept map and key derived from automated content analysis of the researchers' sub-dataset

### *Knowledge users*

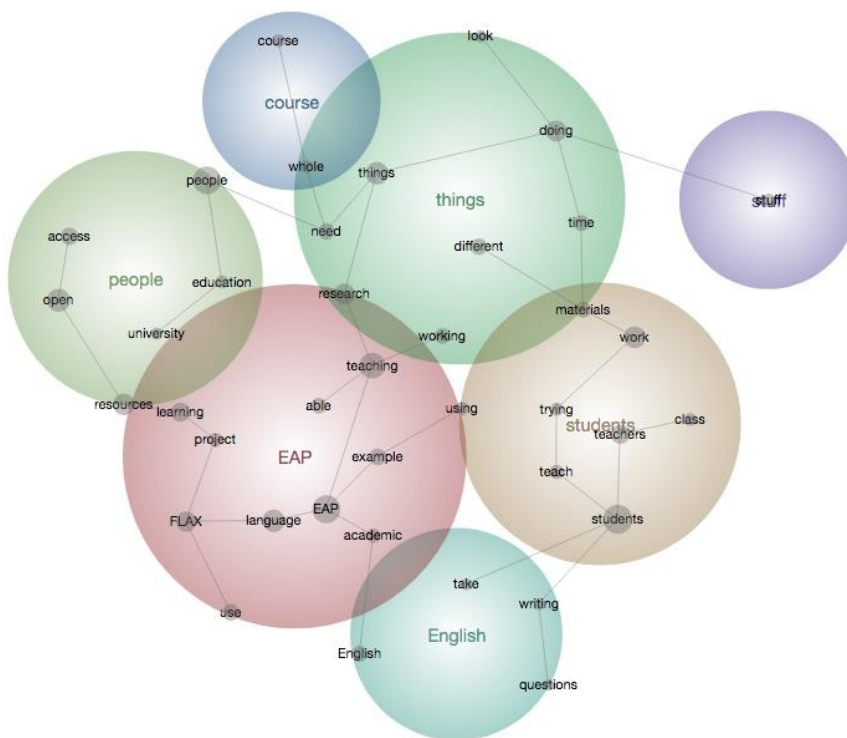
Of the eight EAP practitioners who took part in the research only one, Chris Mansfield of Queen Mary University of London (hereafter referred to as QMUL), had extensive experience with using corpus tools in his classroom teaching, namely the Sketch Engine<sup>39</sup> suite of tools for querying and sketching corpora. The three other participants at QMUL, Martin Barge, William Tweddle and Saima Sherazi, all had a background in Computer Assisted Language Learning (CALL) for developing free online EAP resources for blended learning, most notably Academic English Online<sup>40</sup>. The three EAP teachers at Durham University who were former EAP teaching colleagues of the first author, Terri Edwards, Jeff Davidson and Clare Carr, were early adopters and advocates for using open-source software and/or open educational resources in their classroom teaching as a means of ensuring that their students had access to free and open online teaching and learning resources after their courses had finished that the participants considered to be efficacious. In addition to access beyond their institution's closed Virtual Learning Environment (VLE), the British equivalent to the LMS that is the widely adopted terminology in North America. This motivation to adopt open educational practices as they apply to academic practice in higher education was expressed by the EAP practitioners in this study as a motivating factor for participating in the research with the FLAX project. Learning effectiveness, learner and faculty satisfaction, access and flexibility, and cost effectiveness have been identified as key motivators for educators to engage in blended learning approaches (Graham, 2012).

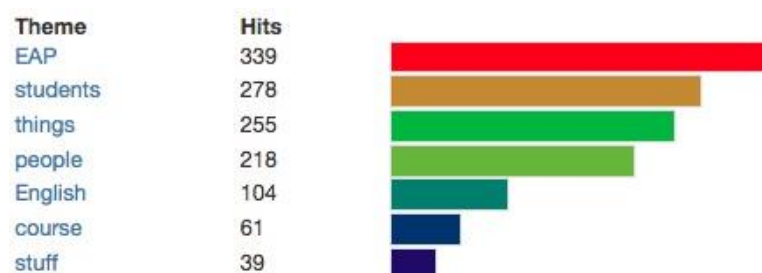
<sup>39</sup> <https://www.sketchengine.eu/>

<sup>40</sup> <http://aeo.sllf.qmul.ac.uk/>



The dominant themes arising from the Leximancer analysis of interviews and focus discussions from project meetings with knowledge users – EAP teachers and course managers – are *EAP* followed closely by *students*, *things* and *people* as shown in the concept map and key in Figure 16. 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 is 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 chapter on knowledge users.





*Figure 16* Concept map and key derived from automated content analysis of the knowledge users' sub-dataset

The work at Durham in 2012 took the form of an OER cascade training project with the participating EAP practitioners and their students that introduced them to four data-driven text analysis language learning systems online: Lextutor<sup>41</sup>, AntConc<sup>42</sup>, Word and Phrase<sup>43</sup> and FLAX. This OER cascade training work led by the first author of the FLAX project also led to collaborative evaluations and further development iterations of the FLAX LC. This work included the addition of the open access BAWE corpus managed by the OTA for a specific focus on academic English collocations (Fitzgerald, 2013a). 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 from 2014-2016 focused on design collaborations with open access PhD thesis abstract content managed by 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, Wu & Barge, 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, Fitzgerald, Yu & Witten, 2018). Table 9 shows the number of abstracts in the PhD Abstract collections: the number of running words, the average length of an abstract, and disciplines in each area. We built digital library collections (sets of electronic documents) for each of these four disciplinary areas as they pertain to the PhD Abstract collections.

<sup>41</sup> <https://www.lexutor.ca/>

<sup>42</sup> <http://www.laurenceanthony.net/software.html>

<sup>43</sup> <https://www.wordandphrase.info/>

Table 9. *Number of abstracts and disciplines in each area of the PhD Abstract Corpora.*

*Reprinted from:* Wu, S., Fitzgerald, A., Witten, I.H. & Yu, A. (2018). Automatically augmenting academic text for language learning: PhD abstract corpora with the British Library. In B. Zou, M. Thomas (Eds.), *Integrating Technology into Contemporary Language Learning and Teaching*, pp. 512-537. IGI Global.

Area	Abstracts	Running words	Average 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.

## ***Discussion***

In this section, we provide discussion on prominent themes and interrelated concepts from the ACA of the datasets. We drill further down into the data to present relevant transcriptions of data from a variety of data collection methods for capturing reflections with participants in the research. Where we present actual data for discussion, themes and concepts will be italicised.

### *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 reuse 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, Sara Gould, and the BL Labs project manager at the British Library, Mahendra Mahey. The progress with policy development for open access and reuse 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 where there has been less progress made with the reuse of educational content. Open access, in most cases, to read-only research publications and, in lesser cases, to pedagogic content, has become the default reuse position of most universities and of mainstream MOOC providers.

The original vision for MOOCs, which would later become known as connectivist or cMOOCs by Downes (2007) and Siemens, included openly licensed content to reflect the ‘O’ for “open” in MOOCs, drawing on principles from connectivist pedagogy (Siemens, 2005). With the rapid ascent of mainstream MOOCs with large platform providers such as Coursera and Udacity came the arrival of another type of MOOC, the xMOOC, which Siemens differentiates as being focused on “knowledge duplication” rather than the cMOOC focus on “knowledge creation and generation” (Siemens, 2012). However, no open policies exist across the broad spectrum of mainstream xMOOC provision by industry frontrunners such as Coursera, edX and FutureLearn where the majority of content is licensed as All Rights Reserved making open access read-only the default user experience.

Once again, it is those individuals who are already open education practitioners, for example, educational technologist, Pat Lockley, of the English Common Law MOOC, or subject academics, Vincent Raceniello at Columbia University (of the Virology MOOCs with Coursera), and William Fisher at Harvard (of CopyrightX, formerly with edX), whom openly license their educational resources with Creative Commons licenses that enabled the FLAX team to develop derivative language 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. Of note, 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 maximize 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, 2014b, p. 17).

In an interview with Pat Lockley, the developer of the open source OER repository and search engine, Solvonauts, and learning technologist responsible for delivering the English Common Law MOOC at the University of London with Coursera, we discussed external platforms for hosting openly-licensed MOOC content, including the FLAX website for MOOC language support resources. Upon reflection on the most significant change in the English Common Law MOOC resulting from his participation as an open education practitioner, he responded that the open educational practice of creating “multiple download formats, open formats and cross hosting sites, basically putting stuff in as many places as you can” would be his legacy with this MOOC (Interview with Pat Lockley, via email, November 2015). Discussion of this point about reuse and redistribution resurfaced on the OER-Discuss online forum, whereby the first author invited Pat to elaborate on and share the nuts and bolts of their interview with colleagues in relation to this issue of hosting open MOOC content externally to MOOC platforms. He drew on an encounter at a MOOC conference:

It was the Coursera conference at Senate House (2013 or ‘14) ... I think I asked about the logic of having a list of Coursera videos outside of the course platforms that people could use. The response from Koller or Ng [founders of Coursera] was that it didn't seem to fit the business models of universities. I spoke to Penn [Pennsylvania State University] afterwards who do a lot of OER and they thought it was a good idea. It might be worth noting here that after three years of using the Coursera VLE [or LMS] the only visible interface changes are on the analytics side and a little bit on asset management. Most of the work has been on the on-demand side. Perhaps they see no benefit to openness and have a business to run? Perhaps it might be easier to level this criticism at FutureLearn? (Lockley, 2015).

The view reported and shared here from the Coursera MOOC founders does seem to run at cross purposes with what we are seeing as evidence from the literature, for example, coming from edX MOOC completers, many of whom are educators who have a vested interest in the learning

content to see how MOOCs deliver subjects they themselves are teaching, reflecting “the diversity of possible, desired uses of open online courses beyond certification” by the larger education community (Ho et al., 2015, p. 2; Chuang & Ho, 2016).

The participating knowledge organisations in this research differ with respects to policies and practices around reuse. It could be argued that work in the digital humanities and in public collaborative projects like those from Wikimedia for the reuse of digital content and collections from galleries, libraries, archives and museums (GLAM)<sup>44</sup> has a longer history with openness than in higher education institutions, where access to knowledge is part of their mission:

The digital humanities can be dated to 1949, when IBM partnered with Roberto Busa, a Jesuit priest, to create a concordance of the complete works of St. Thomas Aquinas. A thirty-year text-digitization project, it is now available online. (Borgman, 2015, p. 162)

British Library Labs (BL Labs) is an Andrew Mellon Foundation funded initiative, which supports the remixing and reuse of the British Library’s digital collections and data for research and educational purposes. In an interview with Mahendra Mahey, the project manager of BL Labs, we discussed the FLAX project research with the EThOS dataset for the development of the PhD Abstract collections wherein he identified four pillars, which enabled the reuse of this dataset that can be broadly applied to the reuse 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 *reuse*? 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* license?”

(Interview excerpt with Mahendra Mahey, British Library, October 2016).

---

<sup>44</sup> [https://en.wikipedia.org/wiki/GLAM\\_\(industry\\_sector\)](https://en.wikipedia.org/wiki/GLAM_(industry_sector))

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 reuse 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 reuse 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 respects to TDM, whereby collections are only available for non-commercial reuse purposes on a request-only basis. The BL Labs manager, Mahendra Mahey, 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 collections in an effort to anticipate affordances and hindrances with conducting research:

Alannah: Can you just gloss for potential readers, what dogfooding is?

Mahendra: Dogfooding is if you're *trying* to promote something, so, for *example*, in our project something with an *experimental reuse of data and collections*. We feel if we're promoting that we should actually eat our own dogfood. We should really do it ourselves. So, we're really *trying* to understand what the issues are. Without doing that you can't understand. So, that's what we do a lot of and when we spoke to Sara, we had an internal *workshop* just with the *EThOS* team and Labs, and we *looked* at the *data*, and we said: "Right, what *experiment* would you like to do on this data?" Because we can use this as a model then to do with, you know, researchers. We can go out into the community and do a similar kind of thing. But we will do it first, so we will know what the pain points are, what will *work*, what won't *work*. Will, you know, a spreadsheet load up in Excel, you know, for *example*? It kept on crashing, for *example*, we learnt that. So, you know, *obviously* we have to give it in digestible chunks if *people* want to be *able* to do basic *things*.

Alannah: And, do you share your code on GitHub?

Mahendra: Yep, absolutely. Part of our Mellon agreement is that everything is *openly* available. Basically, everything from that internal workshop the idea came about, which was, okay, *metadata* for PhDs doesn't always have the funder, or the supervisor. And, we thought, okay, let's do a *text and data mining experiment* on the acknowledgement pages of all the PhDs.

Alannah: Interesting.

Mahendra: So, in order to do that you need to get access to all the *text*, okay? So, that's the *experiment*. So, we're literally at this very moment we have access to about 150,000 full *texts* of PhDs. Some of them have been OCRed [Optimal Character Recognition] and some of them are post 2009, which means they're born digital, so we don't have issues about OCR. We are going to... we are deciding on a little *experiment* on *mining* the acknowledgements pages to see if we can extract useful information on them to...

Alannah: ...to augment the *metadata*?

Mahendra: Yes, to augment the *metadata*.

[...]

Mahendra: Okay, so I have four pillars. I'm just *trying* to remember them all. So, yes, in order to *work* with a collection, yes. 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.

Alannah: Sorry to interrupt you there, but were there any issues around *ETHOS*?

Mahendra: 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 straight forward 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.

Alannah: Is that right? So, it's not always automatically the student's *work*? I thought it was?

Mahendra: All I know is that some *work*, some PhD *work*, is embargoed because it has commercial sensitivities in there. So, for *example*, somebody might...

Alannah: Because they've been funded by...?

Mahendra: Yeah, because they've been funded by Panasonic, for *example*.

Alannah: Yeah, I get that.

Mahendra: 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 *reuse*.

Alannah: 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*.



Mahendra: Exactly.

Alannah: And, that's when somebody puts in a request, for *example*. We want to *reuse* these *texts* for these purposes, and this is what the end result will *look* like kind of.

Mahendra: 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*.

Alannah: 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* in FLAX 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.

Mahendra: 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 Mahendra Mahey, British Library, October 2016).

### *Researchers*

The ACA of the entire qualitative dataset reveals a direct link between the knowledge organisations and researchers' sub-groups with the overlapping themes of *access*. Put simply, access to digital collections that can be reused 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 reuse.

We turn first to a discussion on the perceived affordances of reusing and remixing open access publications for open data-driven learning in EAP. The first author interviewed Maria José Marín who created the BLaRC of 8.85 million words (Marín, & Rea, 2014), which is derived from open access documents licensed with a government license and available from the BAILII online service. Marín developed the BLaRC due to the lack of relevant authentic resources for teaching the specific area of legal English in EAP. The first author 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.

Upon completion of her corpus, Maria José had contacted different corpus projects for enabling online access to the BLaRC and Tom Cobb of the Lextutor added it to his website. She also gave the corpus to the commercial Sketch Engine project, but it was not made freely available for querying purposes on their website. The first author interviewed Maria José 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:

*Alannah:* 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)?

Maria José: 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 [...]

Maria José: 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.

*Alannah*: *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 Maria José Marín, via Skype, August 2015)

Liang Li's experience of trying to carry out research with FLAX and 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 dataset. Her greatest challenges were with securing *access* to research sites with *students* and *teachers* in China to test out the efficacy of the FLAX LC system and the lexical bundles function in the FLAX system. She and the first author, who both 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 that appears strongly on the concept map in Figure 11. This finding from the data is supported by Colpaert's (2018) renewed call to CALL research and practice teams place greater emphasis on transdisciplinary approaches to create new knowledge to remedy the clay feet syndrome present in the field whereby:

...the CALL field remains vulnerable to absorption by other disciplines due to its feet of clay. Its weak point is its very foundation: the lack of CALL knowledge in terms of its own theories, methods, models, frameworks and concepts based on accepted findings.” (Colpaert, 2018, p.1)

*Alannah*: They talk a lot about *user studies* in computer science, don't they?

*Liang*: Yeah, but those *user studies* are only to prove that the tool works.

*Alannah*: Right, the focus is not to prove that *learning* has occurred with use of the tool.

*Liang*: 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.

*Alannah*: That's why we as educational researchers are engaged in this *project* to see if these tools do indeed help with *learning* and *teaching*.

(Interview excerpt with Li Liang, University of Waikato, December 2015)

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.

### *Knowledge users*

Collaborative work with Durham University (2012) and Queen Mary (2014-2016) has revealed that data-driven approaches are not embedded within materials development and classroom teaching practices at these two UK university language centres, although online corpus-based resources have a valued place as supplementary EAP materials at QMUL. 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. This observation differs with findings from an iterative survey-based study investigating the use, or lack thereof, of corpora in language teaching and learning, indicating that almost twenty percent of respondents (N=560) reported “corpus data being used for the preparation of... paper-based classroom materials”, almost on a par with those reporting corpora use by students and teachers as a reference resource at twenty-one and twenty-two percent respectively (Tribble, 2015, p. 53). Tribble does temper these findings from his survey data, however, with the caveat that the survey was largely circulated among DDL and corpus linguistics community discussion lists by leading corpus linguists (Ibid, pp. 45-6).

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 pay attention to discrete language items, and the infeasibility of shepherding large groups of students in developing and mining personalised domain-specific corpora for focused help with, for example, dissertation and thesis writing. This is despite some promising findings from research into DDL approaches with smaller more tailored EAP classes for building Do-It-Yourself corpora with students to help with PhD thesis writing (Charles, 2012 & 2015), so as to maximise the benefits of the two modalities present in blended learning: face-to-face and online (Graham, 2006).

Saima Sherazi, in-sessional EAP programme manager at QMUL, during one of our focus group discussions raised the issue of moving beyond merely introducing corpus-based systems to students for DDL:

Saima: I mean we can *take* students to the water, but we can't make them drink. There actually needs to be a *research project*, probably, where we ascertain how much of what we introduce to them - because this is all that we are doing, we're introducing them to WordSmith<sup>45</sup> or Sketch Engine or introducing them to FLAX - whether they actually use any of them.

(Saima Sherazi, focus group discussion excerpt, Queen Mary University of London, April 2015).

It may be useful, however, to examine the business models behind many EAP programmes where current practices place very little value on researching the design, development, evaluation and impact online resources and classroom teaching materials have on actual teaching and learning. Arguably, there has been far greater provision in the distribution of generic EAP course books by commercial publishers and the uptake of these materials for implementation on an increasing number of EAP programmes. Where evaluations on the impact of materials on teaching and learning do exist, they are often inaccessible to the wider education community:

The aspect of materials development which has received the most attention in the literature is evaluation. Much of what has been written on evaluation focuses on procedures for evaluating materials and on the development of principled criteria. Very little of it presents the findings of actual evaluation of materials for the obvious reason that most evaluations are confidential to publishers, to Ministries of Education or to institutions. (Tomlinson & Masuhara, 2010, p. 7).

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*. The following excerpt from Martin Barge, 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 reuse of authentic language content for classroom teaching purposes:

Martin: 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

---

<sup>45</sup> <https://www.lexically.net/wordsmith/>

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.

(Martin Barge, focus group discussion excerpt, Queen Mary University of London, April 2015).

From the same focus discussion, the pre-sessional course director at QMUL, William Tweddle, discusses the barriers to people working in universities from openly sharing EAP materials across institutions as being 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 reuse 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.

William: 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.

(William Tweddle, focus group discussion excerpt, Queen Mary University of London, April 2015).

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 as reflected in the following extract:

*Chris*: 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 reuse the artefacts of the academy. This really bucks some *people* in academia.

*Alannah*: 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.

*Chris*: 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.

[...]

*Chris*: Now, yes, I need *people* within this higher education environment [Queen Mary] to reuse 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 Chris Mansfield, Cutty Sark pub in Greenwich, London, June 2016)

*A crisis in EAP identity*: An emerging tension in formal EAP is the issue of EAP practitioner identity in the neoliberal university (Hyland, 2002; Hadley, 2015; Ding & Bruce, 2017). 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). Maintaining alignment with teaching aspects of specificity as they pertain to language and discourse norms from across the academy is Hyland's (2018) defence of EAP for supporting students in becoming more critical of the academy:

EAP's pragmatism leaves it open to criticism, these views are seriously reductive and ignore the variety of commitments, contexts and discourses that fall under the EAP umbrella. Indeed, I argue that

EAP can play an important role in assisting students to unpack textual norms to take a more critical view of the academy. (Hyland, 2018, p. 383)

There has been an upswing in commercially produced EAP publications with a notable shift in focus toward 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). Received definitions and understandings from the literature 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) appear to be conflated and confused as the popularity of EAP textbooks and programmes continues to rise and distance itself from the nomenclature and meaning of specificity. Gillett (2018) raises cause for concern that the specific language and discourse needs of EAP learners are not being met by a growing number of EAP practitioners and commercial EAP publications that do not demonstrate the understanding that EAP is a type of ESP:

[that] involves research into the needs of the learners and the nature of the practices and language involved. I think this is particularly important in EAP, as if EAP is not seen as belonging to ESP, then this essential research may be ignored or thought unnecessary and EAP will mean simply using a textbook with EAP in the title, without any clear knowledge or thought of the needs of the students. We can do better than that for our students. (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 between the first author and former teaching colleagues at Durham, Terri Edwards and Jeff Davidson, reflections turned toward collaborative work in developing an OER case study for the UK Higher Education Academy in 2012, which involved trialling corpora and data-driven approaches for EAP (authors). The discussion drew comparisons between the explicit focus on teaching language specificity in EAP against a growing perception that the culture and practice of EAP is moving away from a focus on language toward generic skills, and the implications that this shift in focus might have for teachers and students:



Terri: I think one major, major, 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*.

Alannah: I mean we didn't do any, there was no *time* in the timetables for *language*, right?

Terri: No, for *language*, nothing. It's all just skills.

Jeff: I couldn't believe it when I started *teaching EAP*.

Terri: Skills and process. And, this is so deeply concerning when they don't have the *language* to express their ideas.

Jeff: 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 Terri Edwards & Jeff Davidson, Café Nero, Durham UK, April 2015)

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 corpus-informed. However, there are many more coursebook publications that appear to fly in the face of evidence-based approaches to materials writing for meeting the demands of commercial publishers seemingly driven by market research first and foremost rather than research into whether or not materials have positively influenced teaching, learning and language acquisition.

The following excerpt between *EAP* teacher, Chris Mansfield, and the first author highlights some of the issues with *EAP* materials writing with commercial publishers, resulting in materials that do not always draw on evidence of how language actually works yet are widely marketed for sales distribution:

Chris: What I saw with him [*EAP* materials writer with Oxford University Press] was, with his presentation at IATEFL [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*.

*Alannah*: 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.

[...]

Chris: 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.

*Alannah*: Yes, that always wins out, doesn't it?

Chris: 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, Cutty Sark pub in Greenwich, London, June 2016)

## **Conclusion**

This chapter has presented reflections on different research contexts for exploring open educational practices with relevant stakeholders in resource revision, remix and redistribution with open access content that goes beyond the often-held misconception that the open education movement is primarily concerned with making learning material accessible online (Knox, 2013). The research findings presented in this paper cut across the range of findings found in a recent systemic review of the literature on open educational practices with respects to two major strands: those researchers who “discuss OEP in the context of open educational resources, mostly in terms of open educational resource creation, adoption and use, and those who discuss OEP in relation to other areas, including open scholarship, open learning, open teaching or pedagogy, open systems and architectures, and open-source software” (Koseoglu & Bozkurt, 2018, p. 441).

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 full-text BAWE collections, the EThOS PhD abstract corpora with participating EAP practitioners from Queen Mary University of London, the legal English BLaRC collection by Maria Jose Marín from the University of Murcia, and the ACE corpora with the CORE aggregation and API services at the Open University. There is a growing sense that 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 reuse 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 terminology in non-formal education that has reuse value in formal EAP education. However, findings from our research point to a current problem with the scalability of developing derivative resources 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 open MOOC and networked learning pedagogy. Rather, current MOOC business models appear to focus on paying 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 Jemma König of the FLAX project team with the development of F-Lingo, a Chrome extension, which will be discussed further in Chapter 6 of this thesis with respects to current and planned research. Nonetheless, this work with F-Lingo would still require higher education institutions to allow the reuse of their course content for research and development into domain-specific language learning support in the MOOC context.

The observed absence of data-driven approaches to support blended learning in EAP at two UK university language centres, 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. The absence of investment for measuring the impact on language acquisition of materials used in blended learning in many formal EAP programmes has also been raised here in this chapter. Design-based

research in collaboration with various relevant stakeholders is presented in this chapter as a means of fostering innovative and evidence-based open educational practices with the development of EAP materials, and their implementation in both the classroom and online modalities of blended learning, including those practices supported by data-driven learning systems and approaches. 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 EAP.

This research has also argued for greater access to and reuse 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 academic English language stakeholders place on pedagogic, professional and research texts that can be mined for aspects of domain-specific terminology with data-driven learning systems like FLAX. In addition to the open educational practices that can be fostered to remix and distribute EAP resources for uptake across formal and non-formal higher education.

### ***Connecting Study 1 to Study 2***

Study 1 demonstrated the types of open corpora developed in collaboration with stakeholders in response to initiatives in open access publishing and policy, and reforms to UK copyright law that enabled TDM as a limitation and exception for the development of language learning derivatives in FLAX. In Study 2, the focus shifts to the MOOC space to explore the pedagogical implications and issues surrounding the reuse of pedagogic content to develop data-driven support with the learning of domain-specific terms and concepts. Study 2 demonstrates how increased attention to carrying out DDL studies in non-formal and informal online higher education contexts can help scale DDL approaches with online learners to improve the value and applicability of findings in this area.

## ***Introduction to Study 2***

Opening up education and knowledge to the general public has long been a societal mission of higher education. Engagement through public lectures and the dissemination of knowledge via university presses dates back centuries. Current-day public digital scholarship (Weller, 2011) is amplified with affordances from the open access, open data, open-source software, and open educational resources (OER) movements, although tensions exist, and battle lines have been drawn with the growing perception that openness now has a market value in higher education (Weller, 2014).

It is not surprising then that the mainstream MOOC phenomenon and recent poster child of open innovation in higher education, although still expanding throughout the world, has not yet delivered the future of education to the world as espoused in 2012 by founders of Coursera and Udacity (Koller and Ng, and Thrun). More accurately, MOOCs, and the race to platform education irrespective of their underlying learning ideologies and business models (Siemens, 2011; Siemens, 2012), have not only reached millions of learners. They have also facilitated an uneven distribution in educational access to a small minority of learners from developed countries already connected to the Internet who are predominantly young, male, English-speaking, well-educated and employed (Christensen, Steinmetz, Alcorn, Bennett, Woods, & Emanuel, 2013; Stich & Reeves, 2017). Perhaps one of the greatest ironies of the MOOC phenomenon is now happening in plain sight where learners in the global south, who are paying for MOOC credentials and content that is increasingly being placed behind paywalls (Shah, 2018d), are not only funding MOOC content and assessment development, they are also shouldering the costs of providing access to the already educated lifelong learners of the global north who audit the same courses for free.

### ***The datafication of higher education***

Data was dubbed "the new oil" in 2006 by Clive Humber (UK mathematician and architect of the Tesco supermarket *Clubcard*). The value placed on data would be pumped and piped further in subsequent years by the World Economic Forum to become one of the world's most valuable resources, and this value has been extended to include data from the world of higher education. The hyperbole around the datafication of higher education has come to include big data as well as open data (Borgman, 2015). In the MOOC space, for example, data-driven research methods

have included the mining of MOOC discussion threads (Wise, Cui & Vytasek, 2017). Past and present MOOC reporting trends have also been mined from English-medium news reports from around the globe (Kovanović, Joksimović, Gašević, Siemens & Hatala, 2015). One such widely reported trend has been the low retention rates for MOOC completion (Parr, 2013), giving rise to a vested interest in data mining, primarily reserved for the growing area of educational research into learner analytics to track scores and time spent on learning content and task. According to Bainbridge, et. al. (2015), by using simple learning analytics models, educational providers now have the tools to identify, with up to 80% accuracy, which learners are at the greatest risk of failure before courses even begin.

Data-driven in the context of education has become a loaded term, however, where there is a flip side to, for example, MOOC Terms and Conditions that require learners to give away their data, and for universities to sign over the copyright in their teaching and learning content. Digital data about learners is collected and mined, leading to a current-day boon in predictive algorithms and analytics that are sold down the road to third parties who offer derivative educational products and services as part of the well-established retention industry for “at risk” students (Barefoot, 2004).

As with technology in education, the emerging story of big data in education has been projected and promoted in neutral terms. Nonetheless, this presupposed neutrality surrounding data has been called into question. Leading tech ethnographer, Wang (2013), raises questions about the lack of emphasis and research given over to the social and ethical implications of data-mining and data management that reach beyond the technological know-how and capture of data occurring in, for example, data warehouses and data clouds. In a similar vein to Wang, critical pedagogues and sociologists (Selwyn, 2014; McMillan Cottom, 2015) are also calling for research that incorporates thick data in addition to big data to examine the socio-political economy of emerging data applications in higher education.

One of educational technology’s leading critics and bloggers at Hack Education, Audrey Watters, calls for more questioning from within the field on the perceived pedagogic value of generating algorithms in education that are akin to those derived from big data analytics in the music industry for determining consumer preferences and habits to sell more of the same. Stuck in a perpetual loop, Watters likens the Terms of Use for online educational services to those in the music industry, for tracking user data that amounts to little more than “how many times I

rewound the cassette to play Guns & Roses' *Welcome to the Jungle*" in the era before the Internet (Watters, 2016, paragraph 162). Rather than progressing the field, now viewed as increasingly steeped in venture capital funding with the advent of the mainstream MOOC, Watters raises questions about the value of data that feeds predictive learning algorithms and the types of educational results these can produce:

What sorts of classes get recommended? Are students offered something familiar, comfortable? What signals to the algorithm what a student might find familiar? What happens in the face of an algorithmic education to intellectual curiosity? To risk-taking, to exploration, experimentation, play? ... Does the educational system as-is, with or without an algorithm, value these things? And, what happens when classes are devised to perform well according to this algorithm? (Watters, 2016, paragraph 200)

Further questions can be raised about the value of predictive learning algorithms generated from learner data. Algorithms may well be able to identify which learners have not been able to participate successfully in course discussions and written assessments, and without too much difficulty they will be able to identify differences in learners' native languages and the language of instruction. The problem with the Terms of Use of many MOOCs, and the predictive learning algorithms that mine the data that learners are required to give away, is that the design for learning content and learning management systems, MOOC platforms notwithstanding, have become stuck in a perpetual loop to sell more of the same without addressing underlying issues with designing much needed learning support especially with regards to language proficiency.

We present an interdisciplinary project collaboration between education and computer science mindful of the implications of mining big data in education. In a shift in focus away from learner data collection and learner analytics in online education, we discuss and present a prototype of automated open source NLP tools and methods that can be scaled to augment the MOOC platform learning experience. This exploratory study positions the FLAX system as an 'input-based' intervention (Rott, 2004) that supplies and exposes learners to rich and authentic lexicogrammatical data from lecture transcripts and course readings. Course content is also linked to much larger and more powerful databases that can be searched using information discovery strategies to show how the target domain-specific terminology of a particular course is used in a



variety of contexts. Our goal is to help address the challenge of English-medium instruction in higher education, especially in informal online learning and non-formal MOOC contexts where the majority of courses are invariably offered in English. We wish to present a balanced chapter with respects to the issues surrounding open innovation and data-driven research and praxis in education that we hope will be of specific interest to readers from computer assisted language learning and open education, and of general interest to readers from educational technology.

The terms big data and data-driven in the MOOC context are often bandied about but are primarily reserved for the growing area of educational research into learner analytics to track scores and time spent on learning content and task. Data-driven in the MOOC context, however, does not yet refer to a learning support approach with course content that has been automatically analysed, enriched, and transformed into a data-mined resource that learners can browse and query as was put forward by Johns in 1991 for language learning with linguistic corpora (Johns, 1991). 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, 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).

Depending on our goals, data can be viewed as information about the learner or the learning process, and this is where the current interest and business models with proprietary online educational services bifurcate toward learner analytics. Data can also be viewed as learning and teaching content, including the metadata of that content to facilitate the reusability, remixability, and discoverability of digital learning resources, and this is the view we take in this study with data as educational content.

## **Chapter 4: Study 2**

### **Designing and Evaluating an Automated Open Data-Driven Language Learning Support System for MOOCs**

#### **Abstract**

This chapter presents findings from an evaluative study on the design and efficacy of pedagogical English language corpora that have been derived from the content of two MOOCs, (Harvard University with edX, and the University of London with Coursera), and one networked course (Harvard Law School with the Berkman Klein Center for Internet and Society). Automated text and data mining approaches common to natural language processing were applied to these corpora, which were then linked to external open resources (e.g. Wikipedia, the FLAX LC system, WordNet), so that learners could employ the information discovery strategies (e.g. searching and browsing) that they have become accustomed to using through search engines (e.g. Google, Bing) for discovering and learning the domain-specific language features of their interests. Most notably, the non-formal learner participants in this research and development study had registered for courses in law; they had not signed up as language learners. This speaks volumes to the nature of many informal and non-formal higher education offerings, especially MOOCs, the majority of which are offered in English with no or limited support for learning unfamiliar or semi-familiar domain-specific terms and concepts encountered in their courses.

This research triangulates system query data with user studies by way of self-reported learner and teacher perceptions from surveys (N=174) on the interface designs and usability of an automated open source digital library scheme, FLAX (Flexible Language Acquisition [flax.nzdl.org](http://flax.nzdl.org)). Findings indicate a positive user experience with interfaces that include advanced affordances for course content search and retrieval of domain-specific terms and concepts that transcend the MOOC platform and Learning Management System (LMS) standard. Furthermore, survey questions derived from an open education research bank from the Hewlett Foundation are reused in this study and presented against a larger dataset from the Hewlett Foundation (N=1921) on motivations for the uptake of learning support open educational resources that have been designed for learning at scale in online higher education contexts. This study compares respondents' reported experiences of using domain-specific language learning support resources

alongside other learning support techniques for minimally guided instruction in informal and non-formal online learning. Discussion on future research with the development of the F-Lingo Chrome plug-in for FutureLearn MOOCs will also be presented.

**Keywords:** English for specific purposes; higher education; learning support; massive open online courses (MOOCs); natural language processing; open-source software; open educational resources; terminology; user experience

## ***Introduction***

### *The problem with learning support and the business model behind MOOCs*

Presently, with costs in Internet bandwidth, computing memory and processing power declining rapidly, traditional higher education business models, rather than being disrupted and replaced – as predicted by Clayton Christensen of disruptive innovation theory, and Sebastian Thrun, founder of Udacity – have merely been augmented with the phenomenon known as ‘variable cost minimisation’ (VCM) (Kalman, 2014). Where, for example, a university can offer a MOOC to a small or a vast number of learners with the difference in the consumption costs of bandwidth and processing power for each MOOC participant being negligible. Despite the current number of MOOC learners having reached 101 million worldwide, and course numbers having reached 11,400 (Shah, 2018a; Shah, 2018b), the costs in producing learning content for a course remain relatively fixed. With the greater variable costs of providing much-needed support to learners, including academic and digital literacy learning support for those from disadvantaged backgrounds and developing countries, profoundly outstripping the current VCM business model of MOOCs.

To fast-forward to a case in point of the greater variable costs involved with providing dedicated learning support from the present study, all three courses presented in this chapter were originally designed and delivered as MOOCs. However, CopyrightX from Harvard Law School with the Berkman Klein Center for Internet and Society decided to pull their course from the edX platform in 2013. This move was to limit the participation of non-formal learners to 500 places. Combined with formal residential offerings at Harvard of roughly 100 students, and at Harvard affiliated law schools from around the world taught by copyright law professors who follow the online course with approximately 400 students as participants. A total of around 1000 networked

learners including all three types of cohorts. This ambitious blended and networked model was adopted to enable a more rigorous learning and assessment experience. For the non-formal online cohort, this model includes an English language entry exam, lectures that are pitched to “meet the demanding standards of Harvard Law School”, weekly tutorials with Harvard fellows via Adobe Connect, and a final written take-home exam that has “not been ‘dumbed down’ in any way.” (Fisher, 2014, p.8). The CopyrightX MOOC with edX was thus rebranded in 2013 as a networked course, CopyrightX, offered by Harvard Law School with HarvardX outside of the edX consortium. Professor Fisher of CopyrightX concedes, however, that his select pedagogic model of high levels of engagement and support between small student groups and informed teachers, inspired by the research into interactive learning models (e.g. Renkl, 2002; Hake, 1998; Meyers & Jones, 1993) “does not scale easily” (Ibid, p.16). That the costs for successfully networking an open international course online with dedicated web conferencing technologies and tutor provision for small groups of learners would in most cases be prohibitive beyond the auspices of Harvard.

The current study investigates the different types of learning support offered on all three courses by their respective higher education institutions as evaluated in terms of their efficacy by the non-formal learning and teaching participants in this research. For the purposes of this research, we developed an additional layer of domain-specific academic English language learning support for each course that we contend is useful for both native and non-native speakers of English. Identifiable gaps with academic digital literacy training provision in Internet-based learning schemes stack the chances of success against learning with MOOCs for non-traditional learners, especially when learners are unfamiliar with the use of domain-specific terminology in higher education contexts. Access differentiation in higher education has been well documented in the research carried out into learner perceptions of self-efficacy when failure has been experienced with educational systems, making future attempts to engage in educational offerings less likely (Chemers, Hu & Garcia, 2001; Zajacova, Lynch & Espenshade, 2005). Where informal and non-formal learning is concerned the success of reaching learners more than ever before with innovative Internet-based learning solutions – the mainstream MOOC being the latest in a long line of online distance education innovations – is at the same time isolating learners “in a world of text in an unfamiliar or semi-familiar language (usually English)” (Cobb, 2006, p. 628).

*The problem with MOOC language barriers: The case for domain-specific terminology learning support*

Language barriers to learning in MOOCs have been widely reported (Alcorn, Christensen, & Kapur, 2015; Shapiro et al., 2017). Although recent reports show rapid growth in the number of non-English MOOCs with the rise of XuetangX (China) and Miríada X (Latin America), four out of five of the top MOOC providers still offer the majority of their courses in English (Shah, 2016) with Coursera, the biggest player of them all, offering MOOCs in many languages and thus attracting the largest number of learners of different language backgrounds. Toward the monolingual end of the MOOC spectrum, in the final position out of the top five providers, is FutureLearn with an English-only language of instruction policy that calls for all course communications to be conducted in English (Atenas, 2015). In response to reported language barrier problems, the Translation MOOC (TraMOOC) project (Sosoni, 2017; Castilho, Gaspari, Moorkens & Way, 2017), for example, demonstrates research into powerful translation support for the world's major languages or lingua francas. Research has also been carried out, with somewhat limited success so far, into automated essay scoring and calibrated peer review in MOOCs (Balfour, 2013).

In this chapter, we argue that an additional layer to the language problem exists in the English-medium MOOC space with regards to academic literacy where domain-specific “academic English is no one’s first language” (Hyland, 2019, p.19), nor is it “part of the native speaker’s inheritance: it is acquired rather through lengthy formal education and is far from a universal skill” (Ferguson, Pérez-Llantada & Plo, 2011, p. 42). Hyland in particular is concerned with the academic English writing skill for scholarly publishing and has made contributions to research on specificity in EAP, which is a branch of ESP (Hyland, 2002). The future promise with MOOCs is that they are gaining momentum in offering full online degree programmes that would necessitate a focus on the academic English writing skill rather than past and current trends with offering introductory level courses only and micro-credentials based on multiple choice questions in most cases. Despite trending in the direction of online degrees with 47 on offer in 2018 up from 15 in 2017 (Shah, 2018b), this may be the latest in a long line of promises that has been characterised as the second wave of MOOC hype (Shah, 2018c). We contend, however, that many MOOC learners who are new to the subject areas they are studying, irrespective of their first language,

are further isolated by unfamiliar or semi-familiar terms and concepts in the texts of MOOCs (video lecture transcripts and course readings) encountered through reading and listening that reflect domain-specific language features from target academic communities (Stevens, 1988; Hyland, 2002).

In this study, we focus on three non-formal law courses that demonstrate features of legal English. First, a word on legal English. Despite first appearances, legal English is full of sub-technical terms, that is, words which are shared by general and specialised fields. These words are often employed in both contexts, conveying specialised concepts in the legal field while retaining a general meaning in the everyday field, for example, terms such as *case*, *judgment*, *court* etcetera. As D. Mellinkoff states, one of the major characteristics of legal English is the presence of "common words with uncommon meanings" (1963: 11), which certainly adds to the obscure character of this English variety. Examples from the three law courses in this study include: *deadweight loss* (Contract Law MOOC with Harvard Law School and edX), *due process* (English Common Law MOOC with the University of London and Coursera), and *fixation requirement* (CopyrightX networked course with Harvard Law School and the Berkman Klein Centre for Internet and Society).

*The problem with MOOC content browsability and searchability: Design principles for an augmented learning platform experience*

Since 2013, we have embarked on a journey to remix MOOC content (audio-visual lecture segments streamed via YouTube/Vimeo, transcripts, course readings, quiz questions) with the open tools and open data in FLAX to develop MOOC language support collections (Wu, Fitzgerald & Witten, 2014). We have built digital library collections (sets of electronic documents) for each course in the current study (see Table 10). Our work is an extension of the Greenstone digital library system ([www.greenstone.org](http://www.greenstone.org)), 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 (Wu, Franken & Witten, 2009). FLAX works entirely automatically, without any human input, and can be applied to any collection of academic text. We present a prototype of automated open source tools and methods to support the learning of domain-specific terminology that can be scaled to augment the VCM MOOC business

model and help address the challenge of developing necessary academic and digital literacies in non-formal learning.

Transcribed MOOC lectures present an unprecedented opportunity for developing automated domain-specific terminology learning support. English-medium MOOCs, and an increasing number of MOOCs in other major languages, continue to supply a growing tranche of invaluable transcribed linguistic material that could, we contend, be exploited further for language learning purposes to advance the field of computer supported higher education. When data-mined these digital pedagogic corpora can provide learners with the search and browse functionality that they have come to expect when using search engines for information retrieval. In this way, data-mined course content can also result in learners being able to identify and understand specialised terminology and concepts present in domain-specific lectures, instructional videos, readings and so on. We focus on domain-specific terminology because although it has received much attention in the research literature from applied linguistics in formal classroom-based language education (Stubbs & Barth, 2003), the findings have not been exploited in non-formal higher education. Informal learning is the activity of understanding, gaining knowledge or acquiring skills that occurs outside of formal educational institutions. Both non-formal and informal language learning typically occur without teacher or tutor support as a self-regulated learning activity. Similarly, in formal language education and research from applied linguistics there is consensus that most lexicogrammatical language acquisition takes place outside of the formal classroom in the realms of informal learning (Schmitt & Schmitt, 1995; Schmitt, 1997).

Two design principles underpin the Law Collections in FLAX and aim to minimise learning and training efforts for using the system. The first principle is to capitalise on learners' familiarity with online resources (i.e. online dictionaries and Wikipedia); the second is to utilize learners' existing web search and browse skills with search engines. The pedagogical design of the FLAX system is further principled and underpinned by two theories: noticing hypothesis (Robinson, 1995; Schmidt, 2001) and inductive (discovery) learning (Bernardini, 2002). The constructivist data-driven learning metaphors of the language learner as Sherlock Holmes (Johns, 1991) as scientist (Cobb, 1999), and as researcher (Frankenberg-Garcia, 2005) are grounded in a seminal call made by theoretical linguist, J.B. Carroll (1964), for second language vocabulary learning to "mimic the effects of natural, data-driven, contextual learning, except more efficiently" (Cobb, 1999, p. 19) compared with first language acquisition, which occurs over a much greater period

of time. For this acceleration with vocabulary learning to happen, second language learners “need prodigious amounts of information within an artificially short time” (Martin, 1984, p. 130). In this chapter, we propose data-driven learning in the context of MOOCs that mimics typical web search behaviour to support the learning of domain-specific terminology and concepts.

Many who learn a second-language, or specialised terminology specific to a subject domain in their first language, consult search engines using inverted commas "" and asterisks \* to search for keywords and phrases for language use. This activity has been referred to in the literature as GALL or Google Assisted Language Learning (Chinnery, 2008). Although it is difficult to measure such activity, which occurs in the contexts of informal online learning, studies from formal language education contexts indicate that learners face challenges when using search engines to seek reliable language use data in order to understand and use the target language (Boulton, 2015). Following on from this understanding of GALL for how search engines return an overwhelming amount of dross in response to any query, the FLAX system has been designed to mimic typical web search behaviour while tidying up otherwise messy linguistic datasets from e.g. Google and Wikipedia so that they are searchable, browseable, and therefore manageable (Franken, 2014) for the purposes of language awareness and possibly also language acquisition (Boulton, 2009).

Current MOOC platform and LMS designs do not enable browsing of data-mined course content and searching across course document collections. Nor do they enable course content augmentation with auxiliary learning resources external to the MOOC platform or LMS. The FLAX Law Collections have been designed to enable three pathways for learners [1] to browse and search course content, [2] to retrieve domain-specific terms, and [3] to consult relevant powerful auxiliary resources such as Wikipedia and the FLAX LC system of databases of lexicogrammatical patterns with examples of how these are used in wider contexts.

Design principles for the FLAX Law Collections draw on the literature from applied linguistics whereby encountering and interpreting new terms and concepts in multiple contexts is a widely accepted pre-condition for acquiring productive knowledge and competencies with using new language (Mezynski, 1983; Stahl & Fairbanks, 1986). Displaying language input that presents search results in a salient or enhanced way (Bishop, 2004), in manageable units of analysis (McAlpine & Myles, 2003), and with frequency data (Rott, 2004; Zahar, Cobb & Spada, 2001) are also accepted data-driven learning design principles for enabling learners to make



informed selections for language use. Affording open access to authentic content in corpus-based approaches is not always sufficient in and of itself, however (Groot, 2000). Nonetheless, by linking course content collections to larger, and therefore more statistically powerful, linguistic databases boosts provision with quality language input for learners (Widdowson, 2000).

### *Research questions*

In response to the language support collections we have developed for non-formal online learning, the following research questions were devised as a basis to collect data from participants on their perceived experience of using the FLAX system:

1. Are automated domain-specific terminology learning support systems perceived as motivating to use (i.e. user-friendly and efficacious) in non-formal online learning where there is no formal language support provision?
2. Do the affordances of being able to browse and search data-mined course content that has been linked to auxiliary resources positively augment the learning and usability experience of MOOC platforms and Learning Management Systems?

### *Research hypothesis*

We have also tested, at least in part, the following open educational resources (OER) research hypothesis<sup>46</sup> developed in collaboration with the Hewlett Foundation and the Open University in the UK, which has been modified in the current study for our focus on language education research:

OER Learning Support Hypothesis: Non-formal learners adopt a variety of techniques and resources to compensate for the lack of formal learning support, including support with language.

---

<sup>46</sup> [https://docs.google.com/spreadsheets/d/1fL\\_yf-O70ZjvH67Ue8LlfdjEXwtDQ5T0TBe-Z1GYaI/edit#gid=0](https://docs.google.com/spreadsheets/d/1fL_yf-O70ZjvH67Ue8LlfdjEXwtDQ5T0TBe-Z1GYaI/edit#gid=0)

## Methods

The open education movement has emerged as a key player in informal and non-formal online educational research and praxis over the past decade. MOOCs have helped to grow an awareness of the open education movement in higher education; however, evidence of both the benefits and barriers to employing open practices and resources in higher education is currently lacking in each point. In an attempt to help bridge part of this evidence gap, we developed online surveys to collect perception data from respondents learning and teaching in the context of two MOOCs and one networked course. Results from the study are based mainly on the quantitative survey data by using descriptive statistics and analyses, and by using automated content analyses with the qualitative open-ended survey answers. For purposes of data triangulation, we compare the survey data with FLAX user query data for how the system was actually used in addition to comments from learners on their use of the FLAX system from the online course forum discussion areas.

## Materials

Throughout this chapter, we refer to the Law Collections in FLAX, which are derived from openly licensed pedagogic texts and open access publications from law education and research, along with legal code and judicial hearings from case law available in the public domain. Table 10 shows the dedicated online language collections used in this study to support the two law MOOCs and one networked law course, along with larger databases of corpora linked to the collections to boost their performance as domain-specific terminology learning support resources.

Table 10. *FLAX Law Collections*

Period employed	FLAX <i>Law Collections</i> (used in this study)	Source of collection resources
2014 - 2016	English Common Law MOOC (University of	MOOC lecture transcripts and videos (streamed via Vimeo), quizzes licensed under a Creative Commons Attribution-NonCommercial-ShareAlike license (CC-BY-NC-SA) <sup>48</sup> .

<sup>48</sup> <https://creativecommons.org/licenses/by-nc-sa/4.0/>

	London with Coursera) <sup>47</sup>	
2015 - 2016	CopyrightX (Harvard University) <sup>49</sup>	Networked course lecture transcripts and videos (streamed via YouTube) licensed under the Creative Commons Attribution 4.0 License (CC-BY) <sup>50</sup> , and case law that reside in the Public Domain <sup>51</sup> .
2016	ContractsX MOOC (Harvard University with edX) <sup>52</sup>	MOOC lecture transcripts and videos (streamed via YouTube) licensed All Rights Reserved President and Fellows of Harvard College with permissions granted to the FLAX project for the development of non-commercial educational derivatives.
2014 - 2016	British Law Report Corpus (BLaRC) <sup>53</sup> (Used as a further reference resource)	8.85 million-word corpus of judicial hearings derived from free legal sources at the British and Irish Legal Information Institute (BAILII) <sup>54</sup> aggregation website.
2014 - 2016	Legal Terms List <sup>55</sup> (Used as a further reference resource)	A legal English vocabulary derived from the BLaRC using two Automatic Term Recognition Methods (Drouin, 2003; Marín, 2014).
	Linked to the FLAX <i>Law Collections</i>	
	FLAX Wikipedia English corpus	A reformatted version of Wikipedia (English version), providing key terms and concepts as a powerful gloss resource for the Law Collections.

<sup>47</sup> <http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=englishcommonlaw&if=>

<sup>49</sup> <http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=copyrightlaw&if=>

<sup>50</sup> <https://creativecommons.org/licenses/by/4.0/>

<sup>51</sup> <https://creativecommons.org/share-your-work/public-domain/>

<sup>52</sup> <http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=contractlaw&if=>

<sup>53</sup> <http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=BLaRC&if=>

<sup>54</sup> <http://ials.sas.ac.uk/digital/bailii>

<sup>55</sup> <http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=lawwordlists&if=>

	FLAX Learning Collocations <sup>56</sup> system	A re-formatted Wikipedia corpus of contemporary English, consisting of three million Wikipedia articles comprising three billion words for learning collocations as the default database corpus. The FLAX LC system includes the British National Corpus (BNC) of 100 million words, and the British Academic Written English (BAWE) corpus of 2500 pieces of assessed university student writing from across the disciplines.
--	---	--

For all three courses the first author in this chapter had a point of contact in each university for collaborating on the development of the domain-specific terminology learning support collections relevant to each course: a learning technologist and open education practitioner with the English Common Law MOOC at the University of London (hereafter referred to as the ECL MOOC) whom the first author knew from the UK OER community; a Harvard teaching fellow with CopyrightX assigned by Harvard Law School; and a senior manager of program operations assigned by HarvardX for the ContractsX MOOC.

### *Procedures*

First, surveys were developed for the non-formal learners of all three courses and the networked group of CopyrightX teachers working around the world (N=174), to capture respondents' perceptions on the usefulness and usability of the FLAX system for linguistic support in non-formal online learning. Further questions from an open education research bank developed by the Hewlett Foundation were modified and embedded into the surveys to investigate participants' perceptions of the impact of open DDL resources on increased experimentation and motivation with new ways of learning. The surveys mirror one another in content except for those questions related to differences in the design of dedicated non-formal learning support for each course.

Participants in the study were invited via links from the MOOC platform forum discussion areas (in the case of the ECL and ContractsX MOOCs), and the CopyrightX website and LMS, to use the dedicated Law Collections and links to accompany training videos on the FLAX website that corresponded with their courses, and to participate in the surveys. The surveys required participants to interact with the FLAX Law Collections so as to evaluate the user experience of

---

<sup>56</sup> <http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=collocations&if=flax>

the FLAX language system. Log data was also captured in this way as a result of course participants migrating to the FLAX website to use the language learning support collections.

Surveys covered the following areas:

- General and dedicated learning support resources for non-formal online education
  - The extent to which this support motivated learners to study
- Resources used by learners in general to support language i.e. before the present study
- Learner motivations for using FLAX in their non-formal online courses
- User experiences of the FLAX software
  - Evaluation of user interfaces and functionality
  - Evaluation of using FLAX to support non-formal online courses
  - Open-ended questions on the positive and negative evaluations of using FLAX

Second, user queries sent to the FLAX system were automatically recorded and written to log files for each of the three courses in this study and their corresponding corpora in the Law Collections. These log files were analysed to examine how the Law Collections were used over the iterative course period (2014-2016) when the two MOOCs and one networked course were (re)offered – please see Appendices D-F. The log data analyses that we present in this study, similar to traditional analyses of user queries on the Web, provide interesting and revealing insights that could not be gained from small scale focused user studies. To the best of our knowledge, this user query data analysis approach has not been explored in data-driven language learning research. Nonetheless, because we are the systems developers of the FLAX project, we believe it is useful to share this data with our readers in terms of the design and development decisions made by the project team in response to the actual use of the system (see Wu, Fitzgerald, Witten & Yu, 2019 for further work in this area with system query data analysis).

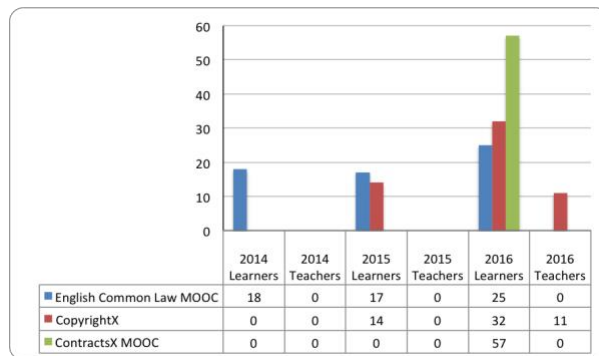
### ***Results and Analyses***

As mentioned in the methods section, the current study has reused and adapted an OER research hypothesis and survey questions that were designed to test the hypothesis on learning support. Reuse of these OER research instruments was done with a view to compare findings from a much larger aggregate survey study into OER uptake by the Hewlett Foundation in collaboration with the OER Research Hub at the UK Open University. Following a short description of the FLAX survey participants' demographic data, and their motivations for using FLAX, we have grouped

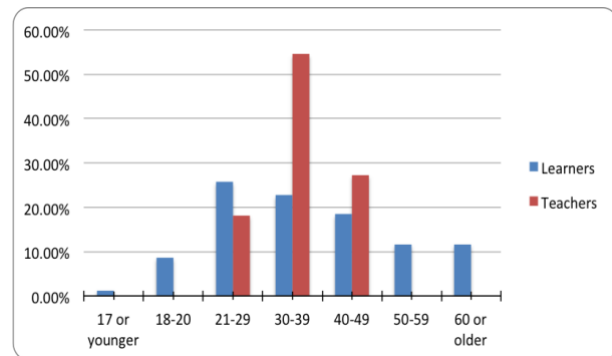
results from across the datasets according to the OER research hypothesis on learning support. The data files from the current study are available at the Open Science Framework<sup>57</sup> online data repository. The data files from the OER Research Hub in collaboration with the Hewlett Foundation are available at Figshare<sup>58</sup>.

### *Demographic data statistics*

The surveys obtained 174 responses from participants who identified as originating from 27 countries and currently residing in 22 countries. The total number of responses collected from across the learning and teaching groups for the three online courses is reflected in Figure 17, with the age of respondents shown in Figure 18.



*Figure 17* Role and courses taken by survey respondents 2014-2016



*Figure 18* Age bands of survey respondents

Of those who responded to the survey, 64.42% were female, 34.97% were male, and 0.61% of respondents preferred not to specify their gender. 66.26% of learner participants were from the United States while 10 out of 11 of the teacher participants were from countries outside of the US. Respondents' education level and employment status are summarised in Figures 19 and 20. It is interesting to note the wide spread of education level and the high numbers in full-time employment amongst the non-formal learner cohort. The data collected closely reflects the demographics of most MOOC takers with the exception being that our dataset shows a far higher percentage of female respondents. Although our learner data shows 21-29-year-olds as being the

<sup>57</sup> <https://osf.io/juakn/>

<sup>58</sup> [https://figshare.com/articles/OERRH\\_Survey\\_Data\\_2013\\_2015/1546584](https://figshare.com/articles/OERRH_Survey_Data_2013_2015/1546584)

largest age band, which is common to most MOOCs, they are closely pursued in percentage points by older age groups.

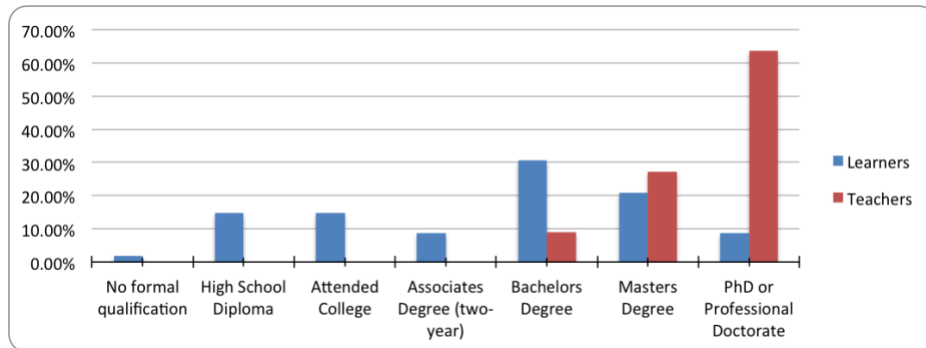


Figure 19 Educational background of survey respondents

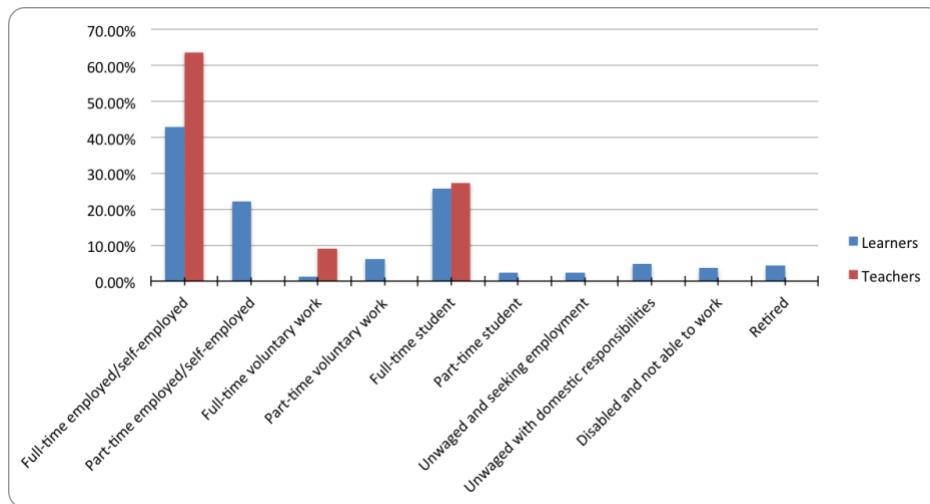


Figure 20 Employment status of survey respondents

Typical of most MOOC taker data, our data also shows a high proportion of non-formal learner participants whose first language is English as being roughly two-thirds (66.87%). It is important to note, however, that the survey was only administered in English. Respondents listed 35 languages in total spoken fluently and 34 languages in which different participants had been schooled and felt they were also able to write fluently. Most learner participants identified as being able to speak English fluently (95.71%), followed by increasingly smaller numbers of participants who identified as being able to speak fluent: Spanish (16.56%), French (12.88%), German (8.59%), Italian (7.98%), Catalan (3.0%), Chinese, Finnish, Gujarati, Swahili (1.84%),

French Creole, Hindi, Japanese, Korean, Luo, Norwegian, Portuguese, Russian, Serbian (1.23%), Arabic, Georgian, Slovak, Thai, Turkish, Ukrainian, Urdu, Vietnamese (0.61%).

### *Learning support resources used and techniques adopted by non-formal learners*

To provide an overview, the surveys asked respondents questions about the types of open learning support they encountered in their courses in addition to FLAX, and any techniques they had adopted to support their learning while taking the courses. The questions were designed to test the OER Learning Support Hypothesis and to compare our data with the larger OER Research Hub dataset (N=1921) asking the same questions of informal online learners. Appendix B shows results for Learning Support (Type A), which is divided into general techniques adopted for supporting non-formal learning shown at the top of the table, and then the different types of dedicated learning support for all three courses.

With respects to the learning support techniques adopted by non-formal online learners, writing study notes ranked similarly between the OER Research Hub cohort and the FLAX cohort at 50.50% and 47.24% respectively. The next sizeable percentages for the adoption of learning support techniques show learners in our study engaged more in social networking with their learning peers at 38.05% compared with the percentage of learners in the OER Research Hub dataset at 26.20%. In addition to using the FLAX Law Collections, it was the dedicated course support identified in our study that shows the real strength in numbers, however, with participants indicating greater uptake with resources designed specifically to support their courses. Only a small fraction of those actually registered on the courses took the FLAX survey, however, with the data reflecting learner behaviour from those who were participating in the courses after several weeks. Other highly used dedicated resources included online course forums, course content and information on the MOOC platforms and networked course website and LMS, and online tutorials conducted in real-time with AdobeConnect for the CopyrightX online cohort.

### *Language resources used by non-formal learners*

As a baseline, we asked respondents about the types of language resources they would have normally resorted to when they wished to express something in English prior to this study. Following the iterative survey studies of Tribble (2015) into the use, or lack thereof, of language



corpora, we decided to ask similar questions of our non-formal online education participants as shown in Table 11.

Table 11. *Survey question: “When you want to find out how to express something in English what resource(s) do you use? You can select more than one.”*

<b>Language Resources</b>	<b>Non-formal Learners (N=163)</b>	<b>CopyrightX Teachers (N=11)</b>
Paper-based dictionaries	18.40%	18.18%
Online dictionaries	76.07%	100.00%
Online reference resources (e.g. Wikipedia)	52.15%	81.82%
Search engines (e.g. Google, using inverted commas "" and asterisks * to search for keywords/phrases for language use)	57.67%	100.00%
Corpora / searchable web-based language collections (e.g. FLAX, WebCorp, Lextutor, COCA)	7.98%	0.00%
Grammar books	11.66%	9.09%
Language course books	1.84%	9.09%
Ask someone	31.90%	27.27%
Need nothing	2.45%	0.00%

When it comes to the language support tools and resources that non-formal online learners are using, there is a clear division between online and offline resources. Although both the learner and smaller teacher datasets show similar results for consulting offline resources for help with expressing something in English, it is the online resources —online dictionaries, online reference resources such as Wikipedia, and search engines— that feature most prominently in the data bar the use of web-based corpora. Respondents from the learner group did, however, report use of corpora (7.98%) but this may well be a reflection of having been exposed to FLAX for the study and may not be an actual reflection of prior exposure to dedicated online language corpora and corpus-based text analysis tools.

### *Motivations for using learning support*

Further survey questions were designed to investigate the different types of learning support and non-formal assessments developed for all three courses that helped motivate learners to study on their courses in addition to the dedicated FLAX language collections we developed, as shown in Appendix C for Learning Support (Type B). The course platforms and websites that contained course content and information came in first place for motivating learners to study at 71%. This was followed by the non-formal assessment of having to do an exam to pass the courses, rated as motivating by 60.74% of learners and 63.64% of CopyrightX teachers.

### *Motivations for using FLAX as learning support*

Results from Table 11 above beg the following question as to why the learners used FLAX to support their non-formal online learning. Table 12 shows results on FLAX user motivations with the following survey question:

Table 12. *Survey question: “What are your/your learners’ main reasons for using the FLAX resources? (Select all that apply)”*

<b>Learner motivations for using FLAX</b>	<b>Perceived by Learners (N=163)</b>	<b>Perceived by CopyrightX Teachers (N=11)</b>
In connection with my/my learners’ non-formal studies	34.36%	81.82%
In connection with my/my learners’ formal studies	NA	9.09%
To improve my/my learners’ legal English	19.02%	72.73%
Personal interest	19.02%	72.73%
Professional purposes	0.61%	36.36%
To assist me/my learners with browsing, searching and retrieving specific subject terms and concepts to help prepare for e.g. note-taking, tutorials, forum discussions, quizzes and exams	60.12%	0.00%
To assist me/my learners with modelling how to contribute to course discussions (e.g. in forums, tutorials)	22.70%	0.00%

To assist me/my learners with modelling how to complete written assessments (e.g. exams)	38.04%	45.45%
To link to further resources (e.g. Wikipedia, FLAX Learning Collocations) to learn how specific legal terms and concepts are defined and used in different contexts other than the course documents	58.90%	63.64%
To save and learn specific legal terms and concepts in the course material through the Cherry-Picking Basket function	41.10%	54.55%

A clear majority of non-formal learners reported in the surveys that the affordances of being able to browse, search and retrieve domain-specific terms and concepts (60.12%) from course content was a key motivating function in supporting learning. However, this same function does not rate with the much smaller CopyrightX teacher cohort. The linking in of external resources such as Wikipedia and the larger FLAX LC system to show how domain-specific terms and concepts are used in wider contexts were also valued highly by learners (58.90%), and most highly by the CopyrightX teachers (63.64%).

### *User query pathway analysis*

We now turn to results from the user query data to determine how users actually employed the FLAX Law Collections in all three courses. The FLAX system records user query entries (user actions or requests for information) in log files while the user is interacting with the system. Appendices D-F show the complete log files for all user query entries for each course. As mentioned in the introduction, three distinct pathways were designed for users to consult the course collections in FLAX:

1. Browse or search course content at the corpus, document, phrase or word level via the collection menu functions
2. Retrieve domain-specific terms and concepts used in the course content

3. Consult auxiliary resources and explore domain-specific terms and concepts in wider contexts (e.g. Wikipedia, FLAX collocations database)

In the following sections, we will discuss three distinct user query pathways that featured prominently in the log data:

- Query pathway A: Browse wikified course documents and consult auxiliary resources
- Query pathway B: Browse, retrieve, consult and save domain-specific terminology located in auxiliary resources
- Query pathway C: Search for keywords at the corpus, document and sentence level

*Query pathway A: Wikification.* Table 13 provides an overview of how users interacted with the Law Collections by looking at user query entries recorded in log files. An example user query pathway for interacting with the system is shown in Table 13 and supported by Figure 21:

Table 13. *Statistics of example query pathway*

<b>Example query pathway: Browse wikified course documents and consult auxiliary resources</b>			
<b>Function type</b>	<b>Query type</b>	<b>Percentage of queries out of total queries (see Appendices D-F)</b>	
1. Click “Browse by title” tab on collection function menu (e.g. lectures, readings, quizzes, extras)	Browse course documents	ECL	25.70
		CopyrightX	25.72
		ContractsX	10.79
2. Click on document menu tabs (e.g. Wordlist, Wikify, Adjective/Noun/Verb) to parse or wikify a course document	Retrieve domain-specific terms and concepts used in the documents	ECL	34.02
		CopyrightX	57.95
		ContractsX	23.74
3. Click on highlighted term or concept in a wikified course document	Consult auxiliary resources e.g. definitions and related topics in Wikipedia	ECL	7.77
		CopyrightX	0.61
		ContractsX	16.33

User query data collected by the FLAX system in this study indicated a far higher use of browsing strategies for full-text course document parsing and wikification. The affordance of being able to reuse full text documents in the Law Collections was made possible by the fact that the majority of the course content used was released as OERs with Creative Commons licences to enable the text and data-mining work carried out by the FLAX project.

As shown in Table 13 browsing full text course documents in the Law Collections can be done at the first level of document querying via the *Browse by Title* (or *Lectures*, *Readings*, *Quizzes*, *Extras*) tabs in the main menu of the collections as shown in Figure 21 where we can also see the submenu tab functions: *wordlist*, *wikify*, and part-of-speech tabs for *adjective*, *noun* and *verb* phrases. Browsing the full text course documents was one of the most significant recorded user activities for the period of the study according to the log data with 25.70% of clicks for the ECL MOOC, 25.72% for CopyrightX, and 10.79% for the ContractsX MOOC respectively. Sub-functions for browsing the full-text course documents in this study can be done at the second level of querying to retrieve domain-specific terms from parsing the documents with wordlists and part-of-speech (POS) syntactic tagging using the OpenNLP<sup>59</sup> toolkit, or from wikifying the documents using the Wikipedia Mining Toolkit. Results show 34.02% of clicks for the ECL MOOC, 57.95% for CopyrightX, and 23.74% for the ContractsX MOOC for this second level of course document querying. Further sub-functions for consulting the external auxiliary resource, Wikipedia, for definitions of domain-specific concepts and for linking to related topic articles in Wikipedia can be carried out at the third level of course document querying with 7.77% of clicks for the ECL MOOC, 0.61% of clicks for CopyrightX, and 16.33% of clicks for ContractsX recorded in the user query data.

FLAX interfaces with the open source Wikipedia Miner toolkit (Milne & Witten, 2013) of machine learned approaches to detect and disambiguate Wikipedia concepts within a course document and to extract key concepts and their definitions from Wikipedia articles as seen in Figure 21 with the *wikify* function. Wikification in FLAX acts as a hyperlinked glossary tool for learners that enables browsing support. It promotes reading and vocabulary acquisition for domain-specific terminology retrieval, and the consultation of auxiliary resources for defining key concepts and linking to related topics in Wikipedia. In the law courses connected to this study, many famous legal cases are mentioned in the lectures and readings. For example, *Carlill v*

---

<sup>59</sup> <http://opennlp.apache.org/>

*Carbolic Smoke Ball Company*, *Butler Tool Machine Company Ltd v Ex-Cell-O Corp Ltd* and *Meeting of the minds* in the lecture document in Figure 21 are identified in FLAX as *Related topics in Wikipedia* for learners to link to and explore further. A definition for a key concept and phrase in contract law, *offer and acceptance*, is also extracted by the Wikipedia Miner, also shown in Figure 21.

The screenshot shows the FLAX interface for the ContractsX MOOC (Harvard University). At the top, there is a navigation bar with links: library, demos, downloads, about, and Login. Below this is a header for the MOOC. The main content area includes a video player titled "Unit 4.0.1 Charitable subscriptions: Law at the margins" with a play button and a progress bar. To the right of the video player is a text box titled "Charitable subscriptions: Law at the margins". The text box contains a definition of "offer and acceptance" and a list of related topics in Wikipedia.

**Charitable subscriptions: Law at the margins**

So far, I've described what might be called the pure geometry of making a contract. There's the idea of a bargain, the idea of *offer and acceptance*, and those are pretty clear. They provide the best way to start thinking about contracts. And they are all. It doesn't always fit into these neat categories.

What I want to talk about now are a number of situations, *charitable acceptance*-- haven't been met. They don't quite fit these tidy has been some major unfairness, something that requires the

Some time ago, in the early decades of the last century, the law where perhaps a college you went to or some other charitable to make a donation. As often happens, the fundraiser calls and have in mind. But then you somehow can't resist the pressure

**Related topics in Wikipedia**

- *Carlill v Carbolic Smoke Ball Company*
- *Butler Machine Tool Co Ltd v Ex-Cell-O Corp Ltd*
- *Meeting of the minds*

Figure 21 Wikification user query pathway for concept definition and related topics in Wikipedia in the CopyrightX MOOC collection

Query pathway B: Collocation. External auxiliary resources such as Wikipedia and the FLAX LC system are linked to language components to give learners opportunities to encounter them in various authentic contexts, and repeatedly (Wu, Franken & Witten, 2009; Wu, Li, Witten & Yu, 2016). Our evaluation of the Law Collections using the various datasets collected in this study does, however, point to limitations with some of the querying functions of the collections menu tabs as being less utilised most likely due to their metalinguistic terminology i.e. collocations. The query pathway identified in the log data for browsing, retrieving, consulting auxiliary resources and saving collocations totalled 5.80% of clicks for the ECL MOOC, 2.05% of clicks for CopyrightX, and 7.71% of clicks for the ContractsX MOOC. Collocations present one of the

most challenging areas of English language learning where there are literally hundreds of thousands of possibilities for combining words. There are many definitions of collocation. We think of collocations in the same way as expressed by Benson et al.:

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. (Benson et al., 1986, p.ix)

Figure 22 shows some of the Top 100 collocations in the *CopyrightX* collection to enable ready identification of useful patterns in the Law Collections by learners. They are grouped under tabs that reflect the syntactic roles of the associated word or words, of which the first seven can be seen here grouped under the “Noun + Noun” tab, along with their contexts. We focus on lexical collocations with noun-based structures verb + noun, adjective + noun, noun + noun, noun + of + noun, and preposition + noun, because they are the most important patterns in academic text. Although only four patterns are offered, more patterns such as verb + adverb, and verb + adjective can easily be added into the system using OpenNLP. The “cherries” icon links to the collocations associated with a particular word, enabling learners to harvest and save collocations to “My Cherry Basket”.

**Top 100 collocations**

Noun+Noun (100) | Adjective+Noun (100) | Noun+of+Noun (100) | Verb+Noun (100) | Verb+Preposition+Noun (100) | Adjective+to+Verb (100) | Adjective+Preposition+Noun (100)

- [copyright protection](#) (295)
- [copyright law](#) (255)
- [district court](#) (175)
- [copyright owners](#) (126)
- [copyright owner](#) (122)
- [copyright infringement](#) (118)
- [summary judgment](#) (95)

■ 21 The relevant facts, drawn primarily from the parties' submissions in connection with their cross-motions for [summary judgment](#)...

■ [...] Ruling on the parties' subsequently-filed cross-motions for [summary judgment](#), the district court (Batts, J.) "impose[d] a way comment on, relate to the historical context of, or critically refer back to the original works" in order to be qualify as fair are transformative only to the extent that they comment on the Photos." [...] The court concluded that "Prince did not intend Photos, or on aspects of popular culture closely associated with Cariou or the Photos when he appropriated the Photos," [...] defendants' fair use defense and granted summary judgment to Cariou.

■ After the completion of discovery, both Mooney and Universal moved for [summary judgment](#), which was granted on August 1, 1979.

■ 22 In a published order, the district court granted in part and denied in part DC's motion for [summary judgment](#), and denied Towle's cross motion for summary judgment.

■ Accordingly, the district court granted [summary judgment](#) on the copyright infringement claim to DC.[...]

■ 43 I dissent, however, because [summary judgment](#) is not appropriate here.

Add collocation to My Cherry Basket

Add Collocation Create a New Category

summary judgment

Choose a category for this collocation:

☒ No category

Figure 22 Collocation user query pathway for top 100 collocations in the CopyrightX collection displaying “summary judgment”

The underlined words in Figure 22, for example *summary judgment*, are hyperlinked to entries for those words in an external collocation database. Clicking on the hyperlinked words displays relevant extracts from a choice of three corpora in the FLAX LC system: the BNC, the BAWE corpus, and the Wikipedia corpus. For example, clicking *judgment* in Figure 22 generates a further popup, shown in Figure 23, that lists *summary judgment*, *default judgment*, *court judgment*, *value judgment*, etcetera, along with their frequencies. Clicking on *summary judgment* in the much larger FLAX LC system with the default Wikipedia corpus selected brings up 271 additional sentences that use this phrase.

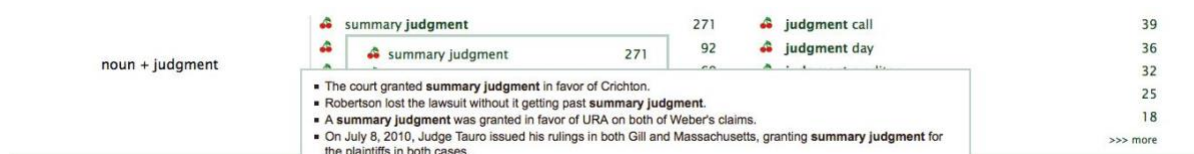


Figure 23 Collocation user query pathway for consulting the term “judgment” in the auxiliary FLAX LC system

Figure 24 below shows learner feedback from the ECL MOOC forum area on the ability to save and organise useful domain-specific collocations with the Cherry Basket feature.

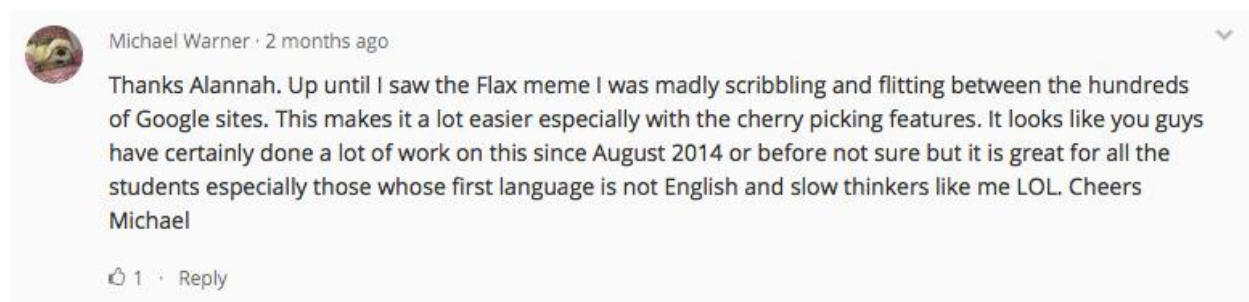


Figure 24 Learner feedback on the Cherry Basket feature in the English Common Law MOOC collection 2016

Query pathway C: Search. By selecting the Search tab in the main menu, learners can perform keyword and phrase searches through the course content at the level of collocations, sentences, paragraphs or full-text documents (e.g. an entire lecture transcript) to locate where and with what



frequency key terms, along with their variants, have been used in the course (e.g. by the course subject academics). We contend that this affordance with search can enhance the user experience with course content beyond the LMS and MOOC platform standard. Language proficiency may, however, also be a factor in participants' ability to use the search menu function of the Law Collections to query the corpora. The total percentage of clicks through the search query pathway (4.45% for the ECL MOOC, 4.6% for CopyrightX, and 16.88% for the ContractsX MOOC) may have been limited to those participants who had prior knowledge of vocabulary items to enable the formulation of search queries, and we will discuss this point further in the following section that presents an automated content analysis of the quantitative variable in the surveys for participants' motivations to search domain-specific terms in the collections.

Figure 25 below shows the first 12 of 151 sentences that utilize the words *common law* in the *English Common Law MOOC* collection; sentences containing the inflected form *commons* are also returned by this search. To recognise inflected forms of a query word, an openly available lemma list<sup>60</sup> containing approximately 15,000 entries is consulted. Clicking the green "arrow" icon at the end of a sentence pops up the paragraph that contains the sentence, to show its context. Search queries can contain more than one word, as is demonstrated here with the two-word noun phrase *common law* in which case sentences are returned that contain all the query terms. For specific phrase searching, a query can be enclosed by quotation marks; for example, "*common law*" returns sentences containing only this phrase.

---

<sup>60</sup> [www.lexically.net/downloads/e\\_lemma.zip](http://www.lexically.net/downloads/e_lemma.zip)

## English Common Law MOOC (University of London with Coursera)

About Collection Search Lectures Quizzes Extras Activities Collocations Wordlist LexicalBundles

My Cherry Basket

### Search Words in Collection

Search for sentences that contain the word common law

You may be interested in commons

### Search Result: 151 sentences found

- We will look at the way in which union law affects the common law.
- In other words, if sovereignty is a construct of the common law and if parliament is behaving in this way limiting judicial review or abolishing courts then it would perhaps be open, perhaps be open to a common law judge to limit the sovereignty of parliament.
- In other words, to use the common law principle that they'd created parliamentary sovereignty to limit parliamentary sovereignty.
- Ah, the question would be whether, in our own common law system, rights are not perfectly well protected anyway, and in fact what the Strasbourg system is doing, the European Convention on Human Rights, is bringing into the law of this country something which simply doesn't belong here.
- I know in common law studies a lot of people don't like studying European law.
- In common law relation to dispute, they compromise of judges magistrates and courts.
- Due Process in the Common Law
- To what extent does the spirit of the common law still need to be thought about, to be engaged with in the on-going struggle for fairness and equality?
- So, if all of these ideas take us back to the common law, they take us back to the, the historical reality, and that is that ideas change. That once we talk about equality before the law in the modern period or today, we mean something quite different from how people understood the term or talked about it in the 1250s, the 1600s, or the late 1700s. Those histories are not obviously completely disconnected, are they? The one history feeds into the other, into that rich understanding that we have of both the reality and the compromises of ideas of fairness in due process, which have characterized its history. I think the question for us now is how these ideas of equality and fairness make sense in our world? To what extent does the spirit of the common law still need to be thought about, to be engaged with in the on-going struggle for fairness and equality?
- ...and indeed these judicial decisions are the principal and most authoritative evidence, that can be given, of the existence of such a custom as shall form a part of the common law.
- Blackstone argues that the authority of law is based on the Church The common law is based on the power of the Crown Judges create the common law through ruling in particular cases.

Figure 25 Keyword search user query pathway for “common law” in the English Common Law MOOC collection

Figure 26 shows learner feedback from the forum area of the ContractsX MOOC with edX, highlighting the ability to navigate through the data-mined course content in FLAX to search for and retrieve domain-specific terms and concepts used by Professor Fried of Harvard Law School.

lvq  
6 months ago

1 Vote +

thanks, Alannah, for this FLAX site for the Contract Law course. i discovered your work a bit late (one month after you posted your announcement of the FLAX site). at the beginning, i was jumping ahead, skipping lectures, and found myself not knowing certain terms (e.g., dead weight loss). i had to go back to the beginning, searching for where Prof. Fried defined this term. if i had known about your FLAX site, it would be of great help. in general, your FLAX site would be great for a review, or for someone searching for certain case. surely, in the future, in case i need to review some concepts of Contract Law, i would hit your FLAX site first. again, great work. thanks.

Figure 26 Learner feedback on the searchability of the FLAX ContractsX MOOC collection 2016

### *Automated Content Analysis of quantitative variables for searching and linking*

By employing the Leximancer software (version 4.5), ACA was performed on the open-ended survey comments in the survey data spreadsheets as text fields (data in text format) and as category fields (data in tabular format) the latter of which performed as variables in the analysis concerning learners' motivations for and experience of using the FLAX system. As previously stated, the software employs semantic and relational algorithms for co-occurrence information extraction (see Smith, 2000a, 2000b, 2003) of concepts and themes from text. Figure 27 depicts a concept map generated by Leximancer with quantitative variables from the surveys for whether or not respondents are native or non-native speakers of English. These variables are examined along with two further variables reflecting the most highly ranked motivations for using FLAX by learner respondents, namely for searching subject-specific terms in course documents, and for linking to further auxiliary open resources (Wikipedia, FLAX LC database) for consulting wider contexts of language and concepts in use. The quantitative tabular data associated with these variables was analysed for the purposes of correlation against the open-ended comments in the qualitative textual data that reflected users' satisfaction with using the FLAX system. ACA renders and quantifies textual data to create concepts and relationships, or words that co-occur, throughout the corpus of text being analysed. Following Bayesian theory, terms are weighted according to how frequently they occur in sentences containing the concept, compared with how frequently they occur elsewhere in the corpus of textual data.

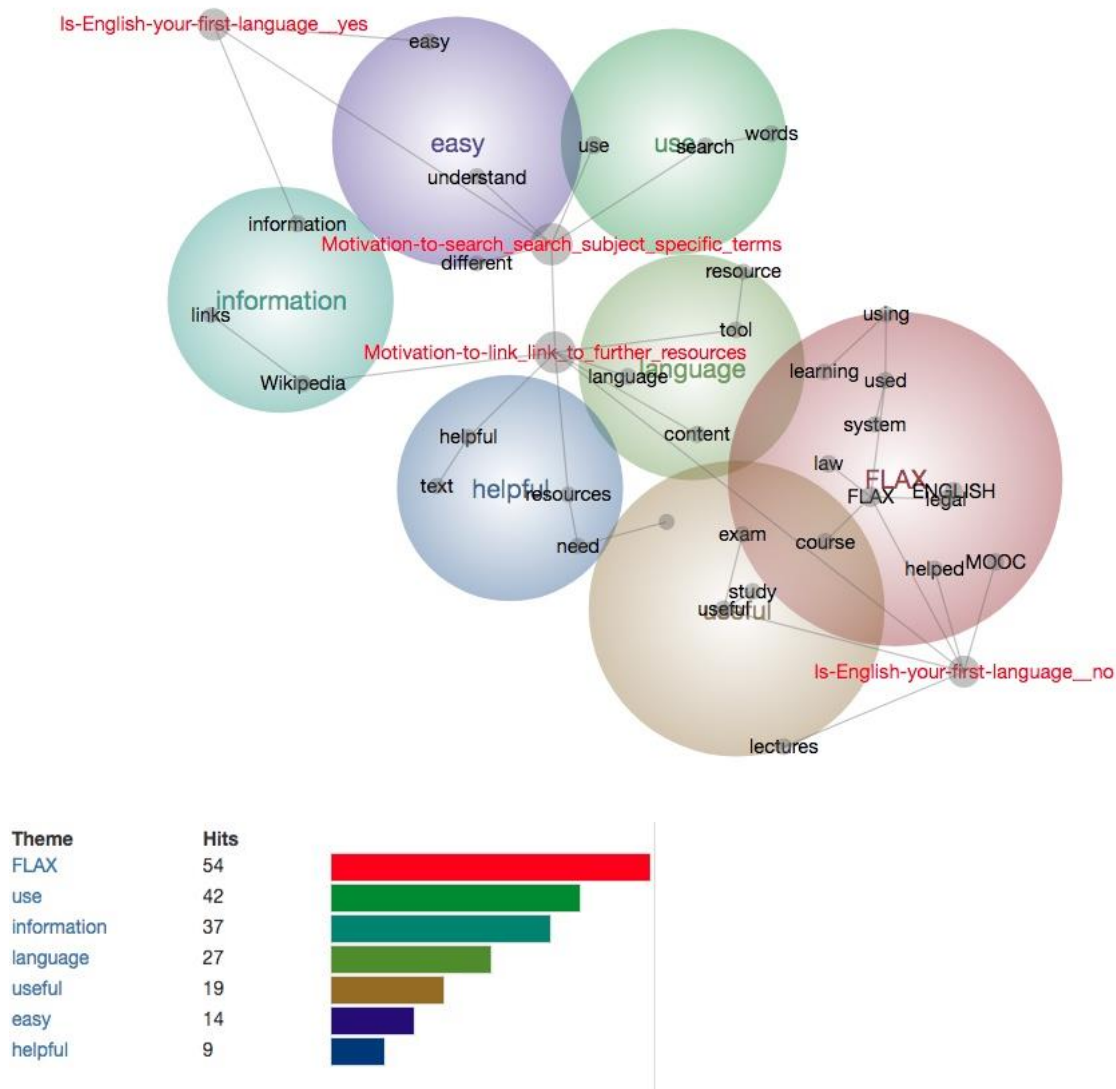


Figure 27 Concept map and key of themes indicating native and non-native English speakers' motivations to search for subject-specific terms and to browse linked OERs

Concepts that are mapped closely to one another indicate a strong semantic relationship (Campbell, Pitt, Parent, & Berthon, 2011; Smith & Humphreys, 2006). These concepts are then clustered into higher-level 'themes' when the map is generated. Figure 27 shows the themes *FLAX*, *use*, *information*, *language*, *useful*, *easy* and *helpful* with the frequency of themes represented in the bar chart to the lower left of the concept map. Leximancer produces a heat map that visually demonstrates the results of ACA with themes that are colour-coded, and where brightness presents the theme's importance (Angus et al., 2013). The 'hottest' or most important

theme appears in red (*FLAX*), with overlapping prominent themes appearing in warm colours, olive (*language*), brown (*useful*), and cobalt (*helpful*) appearing in close proximity to the variables indicating learner participants' motivation to link to and consult further resources for those whose first language is not English. 'Cooler' themes appear at a distance from the warmer coloured themes with variables for English-speaking learner participants and their motivations to search for subject-specific terms. These are represented in cooler colours, turquoise (*information*), with overlapping themes in violet (*easy*) and green (*use*).

The findings from the ACA of these four variables against the open-ended satisfaction textual data indicate that participants' whose first language is English perceived the affordance of being able to search through course document collections as highly motivating. Whereas those whose first language was not English perceived the affordance of being able to link to auxiliary resources to consult wider contexts of use for domain-specific terms and concepts along with their definitions and related topics in Wikipedia as highly motivating. This finding may speak to the observation raised in the previous section where user query data reflected lower levels of searching rather than browsing of the FLAX Law Collection documents with query pathways leading to linked auxiliary resources. This observation may be attributed to whether or not users have the requisite English language proficiency to formulate relevant search queries.

#### *FLAX user experience evaluation statistics*

Granted there are definite limitations in evaluating the FLAX user experience (UX) using learner perception data from surveys as we have done, and without direct contact with learners due to the non-formal nature of the educational contexts; nonetheless, self-reporting data can still shed some light on how learners perceived their user experience of the FLAX system. Satisfaction is perhaps more easily tested in the context of survey-based studies with non-formal learners as they reflect upon their reaction to using FLAX and whether or not they believed it increased their confidence, satisfaction and motivation for the subjects they were studying as shown in the last three statements shaded in grey in Table 14.

Table 14. *Survey question: "Evaluate the following statements about your use of FLAX"*

	Strongly Disagree	Disagree	Neither Disagree	Agree	Strongly Agree

			nor Agree		
Using the FLAX system enabled me to complete English language communication tasks on the course more accurately (reading, writing, speaking, listening)	0.00%	2.55%	24.20%	45.22%	28.03%
Using the Wikify function in FLAX helped me to better understand the full-text course material and related content in Wikipedia	0.00%	4.46%	19.11%	49.04%	27.39%
Using the search function for exploring words and phrases increased my understanding of how these terms were used across the course documents	0.00%	3.77%	19.50%	49.06%	27.67%
Using the collocations and cherry basket functions helped me to understand how important words are combined and used across the course documents and in wider contexts (FLAX collocations database)	0.00%	2.52%	25.16%	45.28%	27.04%
Using the FLAX system increased my independence and confidence in studying the course material	0.00%	2.58%	23.23%	50.97%	23.23%
Using the FLAX system increased my experimentation with new ways of learning	0.00%	3.21%	25.00%	46.15%	25.64%
Using the FLAX system made me more likely to complete the course	0.00%	2.61%	24.18%	49.67%	23.53%

*FLAX user experience.* There is a surprising amount of silence in the literature on text-mining systems for language learning regarding the user experience of interface designs for many of the well-known concordancers and corpus-based systems. The body of research on the design and evaluation of user interfaces for text-mining systems (Shneiderman & Plaisant 2004; Hearst, 2009) has predominantly focused on the internal functionality of systems as a measure of performance rather than evaluating usability performance from the perspective of the users. Our evaluation is largely informed by shifting the focus toward the user experience for determining how far the FLAX system fulfils users' requirements in non-formal online learning. Following user interface evaluation dimensions for text and speech as outlined by King (2007) for determining systems *functionality, reliability, usability, efficiency* and *maintainability*, we devised survey questions that used laymen's terms to map onto each of King's dimensions to determine both positive and negative attributes of the FLAX system according to user-oriented requirements analysis.

UX design considerations are necessitated for how much text will appear on a screen, how this presentation of text can be made more salient and enhanced, and how long this presentation of text will take for diverse computers to process with varying levels of Internet connectivity to load information search and browse queries. These are just some of the considerations for developers who wish to reach out to non-specialist end-users, namely those learners who do not have experience with using corpus-based systems nor the training or exposure to experts in how to utilise them. Figure 28 shows the statements that learner participants were asked to rate on a positive-negative scale regarding their reactions to using the FLAX system.

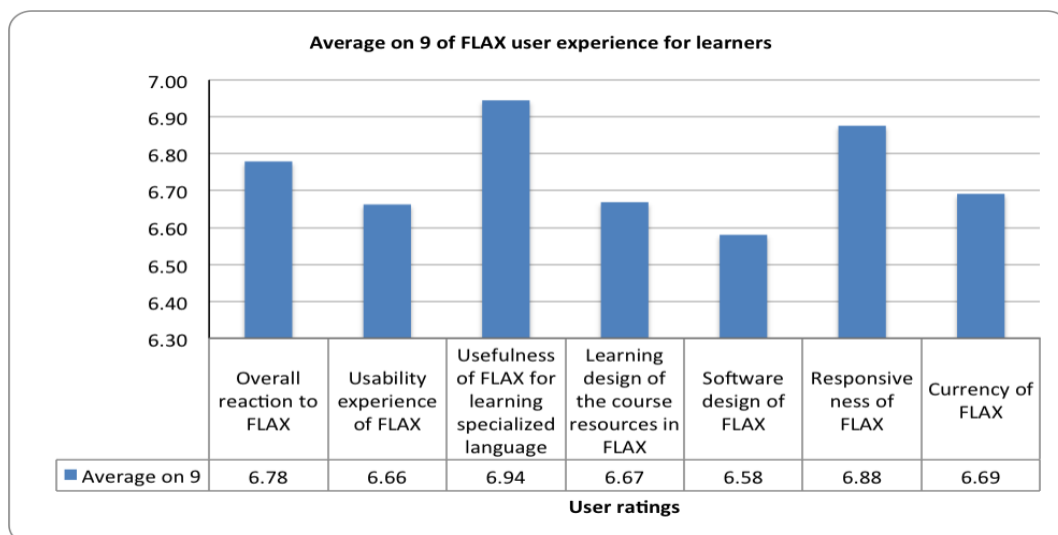


Figure 28 FLAX user experience for non-formal online learners, average on a scale of 9

### *Automated Content Analysis of open-ended survey comments on FLAX user experience*

To dig a little deeper into understanding users' experience of FLAX, we also asked open-ended questions in the surveys concerning what the participants felt to be negative and positive features of the FLAX system and any additional comments that they wished to make.

Once again, we employed the Leximancer software to identify and generate frequently occurring concepts and themes that were repeated across the qualitative textual data categories for survey participants' perceptions of positive and negative features of using FLAX along with further comments from the survey participants on the UX of FLAX. These categories were analysed against the quantitative tabular data category from the surveys for participants' overall reaction to FLAX. Figure 29 shows the concept map generated by Leximancer for the qualitative and quantitative categories described above with *FLAX* represented on the map as the hottest theme. The central *FLAX* theme on the map is overlapped with themes for *use*, *lectures* and *system*, and these themes are clustered around the variables for overall reaction and features of FLAX as being positive according to the survey participants. Specifically, there is a lot of intersection in the positive features comments for the term "FLAX" being qualified with the terms "easy" and "use". Examples from the qualitative data include: "*Easy to navigate*", "*Being able to search through the course content with FLAX made studying for the course so much easier*", "*Studying for the exam was much faster. I feel more confident to use the right words and collocations - a new thing to learn from this FLAX project*".



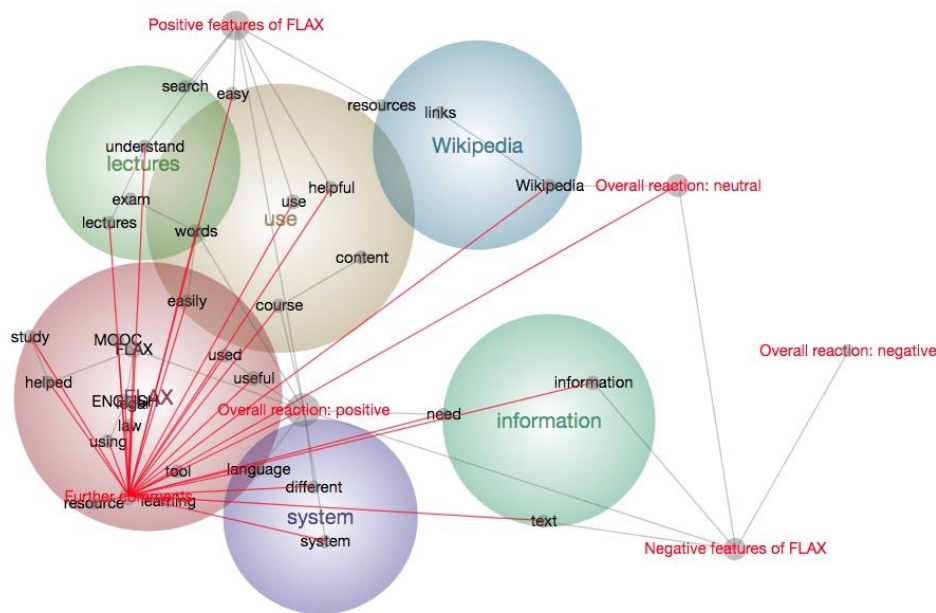


Figure 29 Further comments and overall reactions to the positive and negative features of FLAX

The themes *information* and *Wikipedia* are shown in the concept map as appearing in the middle ground of neutral comments with connections to both positive comments and to a lesser extent to some negative comments. To drill deeper into these themes and their composite concepts, for example, “*information*” in the positive comments was used just as frequently as in the negative comments. In the positive comments, the *information* theme underscores the benefits of having more information choices via “search” and “links” to “resources” that are “helpful” in “understand(ing)” the (“legal”) course (“MOOC”) “content”, e.g. “lectures” and “words” in preparation for the “exam”. Within the neutral and negative comments, these included the participants’ reactions to wordiness: “*information*” overload and the design or look of “text” with the FLAX user interface experience. Examples of negative features from the qualitative data include: “A lot of words on one screen”, “the amount of information on one page is overwhelming”, “hope the text would be in larger letters because you could have someone with a sight impairment”.

Although the use of Wikipedia was rated more positively in the survey responses as shown by the overlapping *Wikipedia* and *use* themes in the concept map, its use in higher education is a contentious one, and also features to a lesser degree in the negative comments in relation to

concerns about the accuracy of the related Wikipedia articles that FLAX links into the user interface experience. One survey respondent managed to distil both the negative and positive views that Wikipedia manages to provoke with the following comment: “*relying on Wikipedia is iffy. (And I write regularly for Wikipedia myself.)*”

## ***Discussion***

One of the rationales for this study has been to gain a fuller insight into the resources and approaches that learners, and their teachers (in the case of CopyrightX), have found to be most motivating and useful in supporting self-directed online learning. A further rationale is to demonstrate the types of open tools and resources available, including tools and resources for TDM and analysis, to support domain-specific terminology learning in mainstream MOOC provision, which has in most part followed the LMS platform-based approach to learning with minimal guidance.

### *Diverse motivations for adopting learning support in the MOOC space*

Diversity in learner motivations and expectations is a result of the open and global nature of MOOCs (Kizilcec et al., 2013), and networked courses, like the ones presented in this study. To answer our first research question on whether automated domain-specific terminology learning support resources are perceived as motivating to use (i.e. user-friendly and efficacious) in non-formal learning where there is no formal language support provision, our findings show that diverse motivations existed among the learner participants for the types of learning support they adopted in all three courses. Perhaps one of the most surprising findings was that both native and non-native English speakers valued and were motivated to use the data-driven domain-specific terminology learning support in the FLAX system with many reporting an overall positive user experience of the system. Indeed, our perception data collected from surveys and online forum discussion areas points to the motivational value participants placed on dedicated course learning support from not only the FLAX project but from the universities offering the courses who devised tailored support resources.

Research into the interdependent processes of motivation and learning identifies individuals who exhibit self-regulated learning traits as being more motivated in their approach to learning, both offline (Zimmerman, 1990) and online in MOOCs (Littlejohn et al, 2016). Our findings

support the Hewlett Foundation's OER learning support hypothesis that non-formal learners adopt a variety of techniques and resources to compensate for the lack of formal learning support, including support with language. The diverse motivations of non-formal learners, and the types of learning that MOOCs and other instances of non-formal learning can support, remains under-researched, however, and calls for further research in this area (Littlejohn et al, 2016, Gillani & Eynon, 2014; Milligan, Littlejohn, & Margaryan, 2013). Indeed, research into learner motivations and learning support in the MOOC space pales in comparison with the far greater number of studies that can easily be carried out at scale and which are driven by large datasets for identifying percentages in, for example, learner progression, retention and completion rates (Breslow et al., 2013; Guo & Reinecke, 2014; Kizilcec, Piech, & Schneider, 2013; Liyanagunawardena, Adams, & Williams, 2013).

*Designing and evaluating augmented learning support systems for domain-specific terminology*

To answer the second research question, perception data gathered from surveys and online course forum discussion areas show that participants viewed the affordances of being able to search and browse through course content that has been linked to external open resources (e.g. Wikipedia, case law in the public domain, the large FLAX collocations database etcetera) can positively augment the LMS experience of MOOC platforms. Log data of user query entries from the FLAX system confirms that a number of participants were clearly able to make use of the Law Collections not just in a limited or restricted way but also in a way where they could move beyond the pre-determined MOOC or networked learning spaces to consult relevant and authentic auxiliary open educational resources. They were not confined to educational platforms with limited pedagogical content where they were in effect being managed (Groom & Lamb, 2014).

The pre-condition with language learning needing to “mimic the effects of natural, data-driven, contextual learning” (Cobb, 1999, p. 19) as put forward by theoretical linguist J.B. Carroll (1964), and as applied to Cobb's metaphor of the language learner as scientist, occurs most prolifically and efficiently in informal learning beyond the parameters of the language classroom (Schmitt, 1997). This activity also raises important questions about the rigidity of closed online learning environments, the LMS and MOOC platforms notwithstanding. A well-known contributor to the *Language Learning & Technology* journal's emerging technologies column,

Robert Godwin-Jones, denotes how the “LMS contributes little to the kind of technology literacy [learners] will need for their personal and work lives” (Godwin-Jones, 2012, p. 6). With the findings from this study with both native and non-native speakers of English, it would, therefore, behove MOOC providers, and educational technologists in general, to think outside of the LMS box, which has become the standard bearer in educational technology research, development and sales. Critical pedagogue, Neil Selwyn (2015), presents an analysis of the inflated hype in rhetoric surrounding vendor sales of LMS technology, while educational technology’s Cassandra, Audrey Watters, points to how the ongoing fixation with the LMS has effectively eclipsed the imagination of the field of educational technology:

Over the course of the past twenty years, the learning management system has become a cornerstone of education technology - how it's engineered, how it's purchased, how it's implemented...It has, perhaps most damagingly I'd contend, become the cornerstone of our imagination - shaping our expectations of what education technology "looks like", how it functions, to what end, and to whose benefit. The learning management system has become a behemoth, an industry unto itself, part of a larger behemoth of an increasingly technologized university. (Watters, 2016, paragraph 136)

The design of any technology user interface for uses in education has a better chance of success if it follows the design principles of simplicity, accessibility and functionality. Downes (2004) defines simplicity in educational technology design as those tools which are not only easy to use but those which have been designed to perform necessary functions only. From its earliest inception, the FLAX system has been envisioned and advanced with the language learner in mind rather than the corpus linguist. In a move away from traditional concordancers for presenting KWIC language output, which stem from research into corpus linguistics for querying and analysing text, we have moved toward the development of an open source data-driven language learning system that mimics typical online search and browse behaviour; wherein course content is linked to authentic web-based content that has been cleaned up and data-mined for language learning purposes (Wu, Franken & Witten, 2009). In this way, for the evaluative purposes of this study, we have developed simpler user interfaces to enable novice users to successfully interact with complex linguistic datasets without any prior linguistic or metalinguistic knowledge for querying corpora.

## ***Limitations***

As with any study, there are a number of limitations influencing the findings and conclusions that can be drawn. Without direct access to the delivery side of the MOOC and networked course spaces presented in this chapter, we have instead relied on the willing collaboration of individuals at the participating institutions. This has resulted in limited attempts to inform learners of the existence of the FLAX Law Collections and to provide the necessary online training and support with how to implement the resources to effectively support learning with domain-specific terms and concepts. As previously stated, this was an exploratory study and the data were generated in the context of participants' limited experience with the system. While this is acknowledged as a limitation, nonetheless the results are encouraging, especially with respects to the log data that reflect a high percentage of user query pathway entries for browsing full-text course documents for the retrieval of domain-specific terminology from wikified or syntactically parsed course documents along with high instances of consulting auxiliary resources such as Wikipedia.

User query data reveals far lower instances for use of the other main menu function tabs in the Law Collections for *Search*, *Collocations*, *Wordlist* and *Lexical Bundles*. As discussed earlier, the use of the search function may be limited by whether or not the user is a native English speaker and has the requisite vocabulary knowledge to be able to recall and enter relevant search terms. The other menu function tabs are named according to metalinguistic categories, which, once again, without any explicit training in how to exploit these functions would inevitably result in their underuse by non-expert users of the system. The findings from the user query data are perhaps the most useful evaluation of the systems' efficacy in non-formal online learning contexts. Current research and development work for the use of the FLAX LC system and the Wikipedia Miner toolkit have resulted in an even more radical departure for data-driven language learning systems design for the non-formal online learning context with the development of the F-Lingo Chrome extension for the FutureLearn MOOC platform. This work headed by Jemma König at the University of Waikato has eliminated all metalinguistic terms from the F-Lingo software. Its primary function is to automatically traverse and retrieve domain-specific terms and concepts for browsing course documents, and by linking to external auxiliary resources using menu function tabs that are in plain English: *Words*, *Phrases* and *Concepts*.

The findings from this research are imperfect because of the small number of actual learners who participated in our study are only a fraction of the number of those who registered in all

three courses over the period 2014-2016. However, bearing in mind that MOOC completion rates are 7% on average (Parr, 2013), our findings do nevertheless offer an insight into the non-formal online education community and are worth bearing in mind when considering future provision in this area. Furthermore, the study only sampled participants who participated in the MOOCs and networked course leading up to and inclusive of the final course assessments. MOOCs suffer from high attrition rates, particularly in the first few weeks. As with many MOOC and non-formal learning studies, capturing trace data earlier from learners who do not complete MOOCs could provide valuable insights into the reasons for their dropping out as this is a phenomenon which confounds many studies in MOOC and non-formal learning spaces.

### ***Conclusion and future research***

This study made use of a tracking system that wrote users' queries of the FLAX system to log files to investigate participants' use of different strategies to browse, search and retrieve domain-specific terms from course documents, and to consult relevant and authentic auxiliary resources for term usage in wider contexts (FLAX LC), and for concepts and related topics (Wikipedia). This tracking system data was presented alongside participant perception data from surveys both of which served to investigate how participants functioned and perceived themselves (and their learners in the case of the CopyrightX teachers) as 'learner scientists', to use Cobb's (1999) metaphor. Given the increasing availability of online data-driven language learning systems, studies that trace user query entries as a means of documenting strategies and learning pathways for employing such systems would appear to be important in evaluating how they can be improved to better mediate and support learners' information retrieval strategies.

Future research could make use of trace data written to log files in this way in addition to being supplemented with learners' explanations of the strategies they choose to employ and the learning pathways through online learning systems they choose to follow. Richer and more specific data may be gained from employing think-aloud protocols and techniques (Ericsson & Simon, 1987), and cognitive walkthroughs for evaluating the usability and learnability of the data-mined MOOC content. In this way, insights from probing learners' choice of strategies for browsing and searching course documents and consulting external auxiliary resources to retrieve subject domain information may provide useful accounts of the various factors that can affect strategy use in real-time.

Furthermore, with all the emphasis given over to data-driven applications in the mainstream MOOC space, and the considerable amount of start-up funding from the private sector in the race to innovate higher education online learning platforms, it would be reasonable to imagine the future of MOOC content —an ever-amassing online pedagogical corpus — as being enhanced with the text and data-mining capabilities of search and browse; with links to further open educational resources; and with domain-specific terminology analysis tools for dedicated learning support. The research presented here provides proof of concept for taking the affordances of open data-driven learning to the non-formal online education context with perception and log data that show non-formal learners value and find motivating dedicated learning support in domain-specific subject areas. However, this approach is not currently scalable in the non-formal online learning context. This is apparent in current mainstream MOOC provision where existing business models have not yet anticipated the need for enriched course content using text and data-mining approaches to augment the closed LMS-style MOOC platform learning experience. Nor has there been the shared understanding that openly licensing course content is a priority for reuse by the wider education and research community for developing domain-specific terminology learning support being one useful example of reuse.

The findings presented here raise issues for further consideration in higher education policy for innovating the design and development of minimally guided non-formal online learning with particular importance placed on the challenge of acquiring domain-specific terminology for academic and professional purposes. Current research with the FLAX project for supporting non-formal online learning of domain-specific terminology is centred on iterative design and development work with the FutureLearn MOOC platform by way of development of a Chrome browser extension, F-Lingo. This work will bring the affordances of the FLAX project into the MOOC platform interface, which will be discussed in more detail in Chapter 6. In response to the positive outcomes with participants in the current study for developing and providing data-driven language learning support, it is our intention to scale the research with participating FutureLearn MOOC host institutions and registered learners across various domain-specific subject disciplines.

## Notes

Setbacks were encountered in Study 2 concerning timely access to and permissions to reuse the MOOC and networked course content for creating the Law Collections in FLAX. Delays occurred with course content being developed up to the last minute by the universities delivering the courses for the first time, and with content being updated for successive course reruns. Further delays were encountered with copyright restrictions with some of the third-party reading material selected for CopyrightX, which we were not able to gain clearance to reuse. Nonetheless, both the ECL MOOC and the CopyrightX networked course published their lecture material using Creative Commons licenses (see Figure 30).

Source
<ul style="list-style-type: none"><li>■ Copyright 2016 William Fisher. This video is licensed under the Creative Commons Attribution 4.0 License (CC-BY). This lecture from Harvard University has been streamed into the FLAX language system from the third party video service provider, YouTube. The following licensing information for the lecture has been duplicated here in this source information area of the CopyrightX FLAX collection:</li><li>■ The portions of the CopyrightX lectures consisting of original material are licensed under the Creative Commons Attribution 4.0 License. The lectures also contain excerpts of non-original material, some of which are subject to copyrights held by other parties. The use of those excerpts in CopyrightX is privileged under the fair use doctrine. However, Prof. Fisher and Harvard University lack authority to license others to use the excerpts. Thus, persons who reuse a CopyrightX lecture must either remove the pertinent excerpts or ensure that they too enjoy legal authority to reproduce or perform them.</li><li>■ William Fisher, CopyrightX: Lecture 1.1, The Foundations of Copyright Law: Introduction. (2016). Retrieved from <a href="https://www.youtube.com/watch?v=CqkonSY__ic">https://www.youtube.com/watch?v=CqkonSY__ic</a></li></ul>

Figure 30 Licensing information for source material in the FLAX CopyrightX collection

Although part of the HarvardX and edX consortiums, the ContractsX MOOC lecture material is licensed as All Rights Reserved (see Figure 31). Harvard University entered into a legal copyright agreement with the FLAX project at the University of Waikato, drawn up by Harvard lawyers, for the non-commercial educational reuse of the ContractsX material on the FLAX website. This legal process created a further delay with the live version of ContractsX in early 2016.

Source
<ul style="list-style-type: none"><li>■ Created by HarvardX and delivered as an edX MOOC. Copyright President and Fellows of Harvard College. All Rights Reserved with permissions to the FLAX project for the development of non-commercial educational derivatives.</li><li>■ Illustrations by Benjamin Maurer Visual Art LLC</li><li>■ What is a contract? Introduction (2015). Retrieved from <a href="https://www.youtube.com/watch?v=5fUcC1perfY">https://www.youtube.com/watch?v=5fUcC1perfY</a> This video content can also be accessed at the edX MOOC, Contract Law: From Trust to Promise to Contract <a href="https://www.edx.org/course/contract-law-trust-promise-contract-harvardx-hls2x">https://www.edx.org/course/contract-law-trust-promise-contract-harvardx-hls2x</a></li></ul>

Figure 31 Licensing information for source material in the FLAX ContractsX MOOC collection



### ***Connecting Study 1 and Study 2 to Study 3***

So far in this thesis, we have looked at the affordances of open data, open access publications, and open educational resources in the co-design and co-creation of corpus-based systems in FLAX with relevant stakeholders. Whereas Study 1 provided reflections on the remixability of open linguistic data and further open resources for DDL systems design, and the resulting academic corpora that have been developed for this doctoral research, Study 2 demonstrated proof of concept for how this open corpus-based development approach could be applied to the MOOC space and non-formal online learning in higher education. Study 1 provided insights through qualitative data in the form of interviews, focus-group discussions, observations and meetings with stakeholders that were then analysed using the Leximancer automated content analysis software that generated concept maps with dominant themes arising out of the datasets. Study 2 employed mixed methods combining surveys to collect MOOC learner and teacher perception data on the efficacy of the FLAX system and this data was discussed in relation to the FLAX log data. Now with Study 3 we turn our focus back to the traditional language learning classroom in formal higher education where OERs from the English Common Law MOOC with the University of London and Coursera were reused and developed into the ECL MOOC Law Collection in FLAX. This openly licensed MOOC content was reused in an experiment with legal English corpus linguistics and translation researchers at the University of Murcia. Here, we investigated the learning of legal English terminology with performance data from the analysis of student writing to evaluate to what extent the ECL MOOC Law collection had attributed to the use of specific legal English terms in writing.

### ***Introduction to Study 3***

#### *The reuse value of domain-specific OERs in higher education teaching and learning*

A great deal of the value placed on OERs, in higher and further education especially, is their cost-saving value. Hilton III and Wiley have carried out extensive research into the cost-effectiveness of open textbooks measured against those from the big brand commercial publishers (Hilton III & Wiley, 2011; Hilton III, 2016; West, 2018). This research has been used to lobby governments and philanthropic foundations such as the Hewlett Foundation to invest in educators for the development of OERs and open textbooks that can be reused, remixed and redistributed by the education community for non-commercial purposes. One of the driving principles behind this investment is that it will free up educators to reuse and develop OERs, and to write open textbooks with their peers rather than having to follow the marketing whims and directives of commercial publishers. It will also save students millions in a move to divert their learning content needs away from the commercial education publishing industry as has been evidenced in recent substantial budgetary allocations from the US Department of Education for teachers to develop OERs and open textbooks (see SPARC, 2018).

The following chapter presents a study in the area of DDL that has been driven by scarcity of authentic reusable resources in specialised English varieties for supporting English for Specific Academic Purposes (ESAP). By adopting TDM and NLP approaches with open access content and OERs, the FLAX project offers a solution to the problem of access to authentic resources that represent the language and genre features of specialised English varieties in supporting domain-specific language learning for academic and professional purposes. A further driver to this study was our interest in investigating what, if any, value could be derived from reusing authentic data-driven pedagogic resources for learning specialised terminology and whether or not this approach could positively influence the usage of specialised terminology in student writing. Study 3 will present figures from analyses carried out into domain-specific term usage that indicate that the experimental group in our study employed specialised terminology better in their essay writing than did the control group in our study.

Many OER studies have focused exclusively on the cost reduction aspect of using and developing OERs but very few studies in open education have looked at whether OERs can improve learning performance. The following study provides a window onto the potential of

reusing domain-specific OERs in higher education learning and teaching with respects to their value in writing instruction.

The experimental group in Study 3 had access only to the data-mined and enriched English Common Law MOOC corpus on the FLAX system and were asked to write an essay on a topic about the same legal system. The control group had access to any information source on the Web and were asked to complete an essay on the same topic. Findings from the study indicate better usage of the specialised legal terminology in student writing from the experimental group, which gives pause for concern when we consider the ‘Google effect’ on the (re)search and retrieval capabilities, and the reading comprehension and writing performance of students who are ‘directed’ to use the Internet as their primary information source for learning. Brabazon (2006; 2013) characterises the ‘Google effect’ as the equal rendering of all data by the search engine where no distinction is made between “important” as opposed to “popular”, “banal” and “repetitive” information. A paucity of useful information contrasts with a plethora of more popular, yet less tested information that has been pushed higher in Google rankings. Without strong information literacy skills, the characterisation of the ‘Google effect’ continues that often less robust information that has been retrieved from employing Google as a source for research can be evidenced in a negative transfer to student writing:

To translate [McLuhan's] statement [on living in a time of speeded up information] for the purposes of this book, I investigate the impact of an information glut that is not only rich and complex, but repetitive and banal. When there is too much information in the present, how is it judged, sorted and sifted, to separate the basic and simple from the important and complex? Such a process is rendered more complex because of 'The Google Effect' [Brabazon, 2006]. At its most basic, this phrase describes a culture of equivalence that renders all data equally ranked before a search engine, creating confusion between the popular and the important. The impact of this confusion is problematic for many institutions but is most serious for schools and universities. [...] I experienced the consequences of [Arum and Roksa's, 2011, p.98] research first hand when assessing assignments for a first-year course in North America that was taken as an 'easy elective' for third- and fourth-year students. When marking their papers, I could not tell the difference between the quality and standards of first and fourth years. The level was indistinguishable. There was no distinction in analysis, investigative depth, or interpretative complexity. (Brabazon, 2013, p. 2)

## Chapter 5: Study 3

### Evaluating the Efficacy of the Digital Commons for Scaling Data-Driven Learning

#### *Abstract*

Open principles for Computer Assisted Language Learning (CALL) design and practice will be addressed in this chapter. Open educational practices for designing and developing domain-specific language corpora with the open-source FLAX language project will be demonstrated and discussed with respects to the re-mix of openly licensed pedagogic, research and professional texts from the digital commons. The design of the open Law Collections in FLAX will be used as a running example throughout this chapter in response to the scarcity of reliable and specific resources for learning legal English. A loop-input discussion will also be presented on the legal development of the Creative Commons suite of licenses, which have enabled this novel approach to English language materials development practices with open educational resources and open access publications for data-driven learning in the area of English for Specific Academic Purposes (ESAP).

This chapter presents a data-driven experiment in the legal English field to measure quantitatively the usefulness and effectiveness of employing a corpus-based online learning platform, FLAX, in the teaching of legal English. Participants in the study included 52 students in the fourth year of the Translation Degree program at the University of Murcia in Spain who were selected as informants over two semesters. All of the students' linguistic competence level complied with the Common European Framework of Reference for Languages requirements for the B2 level. The informants were asked to write an essay on a given set of legal English topics, defined by the subject instructor as part of their final assessment. They were then divided into two groups: an experimental group who consulted the FLAX English Common Law MOOC collection as the single source of information to draft their essays, and a control group who used any information source available from the Internet in the traditional method for the design and drafting of essays before this experiment was carried out. The students' essays provided the database for two small learner corpora. Findings from the study indicate that members of the experimental group appear to have acquired the specialized terminology of the area better than

those in the control group, as attested by the higher term average obtained by the texts in the FLAX-based corpus (56.5) as opposed to the non-FLAX-based text collection, at 13.73 points below.

**Keywords:** digital commons; English for specific academic purposes (ESAP); data-driven learning (DDL); corpus linguistics, learner corpora; massive open online courses (MOOCs); open educational resources (OER)

## ***Introduction***

### *The growing digital commons and open educational resources*

This chapter presents the open source FLAX project (Flexible Language Acquisition, [flax.nzdl.org](http://flax.nzdl.org)), an automated digital library scheme, which has developed and tested an extraction method that identifies typical lexicogrammatical features of any word or phrase in a corpus for data-driven learning. Here in this study, FLAX will be described and discussed in relation to the reuse of openly licensed content available in the digital commons. Typically, the digital commons involve the creation and distribution of informational resources and technologies that have been designed to stay in the digital commons using various open licenses, including the GNU Public License and the Creative Commons suite of licenses (Wikipedia, 2016; see also the chapter by Stranger-Johannessen, this volume). One of the most widely used informational resources developed by and for the digital commons is Wikipedia. In response to the growing digital commons, we will provide insights into design considerations for the reuse of transcribed video lectures from MOOCs that have been licensed with Creative Commons as Open Educational Resources (OERs). We will demonstrate how OERs can be remixed with open corpora and tools in the FLAX system to support English for Specific Academic Purposes (ESAP) in classroom-based language education contexts.

This research arose largely in response to the open education movement having recently gained traction in formal higher education and in the popular press with the advent of the MOOC phenomenon. The OpenCourseWare movement, which began in the late 1990s, preceded MOOCs with the release of free teaching and learning content onto the Internet by well-known universities, most notably the Massachusetts Institute of Technology. Indeed, MOOCs are the latest in a long line of innovations in open and distance education.

This chapter also draws attention to the OER movement, where the emphasis on ‘open’ signifies more than freely available teaching and learning resources for philanthropic purposes (open gratis). Here, we focus on the truly open affordance of flexible and customisable resources that can be retained, revised, repurposed, remixed, and redistributed by multiple stakeholders for educational purposes (open libre). In the present research with the FLAX project, open resources are specifically employed in the design and development of domain-specific language corpora for scaling DDL approaches (discussed below) across informal MOOCs and formal language learning classrooms.

The mainstreaming of open content, including OERs and open access publications, came swiftly on the back of the development of the Creative Commons suite of licenses by copyright lawyer, Larry Lessig, in collaboration with Internet activist and open education advocate, Aaron Swartz. Their collaboration resulted in six Creative Commons licenses that were released in 2002 to retain the copyright of authors for enabling ‘Some Rights Reserved’ in a movement away from the default ‘All Rights Reserved’ restrictions of licensed creations. An estimated one billion Creative Commons-licensed works now reside in the digital commons (Creative Commons, 2015). This growing movement provides evidence that the read-only culture of analogue content developed by commercial publishers and broadcasters for passive consumers is being eclipsed by the read-write digital culture of remix, with an increasing number of active creators electing to share content online with free culture licenses (Lessig, 2004; 2008). According to Wiley (n.d.), Creative Commons licenses enable the following permissions to the education community by means of defining the affordances of OERs:

1. Retain: the right to make, own, and control copies of the content (e.g., download, duplicate, store, and manage).
2. Reuse: the right to use the content in a wide range of ways (e.g., in a class, in a study group, on a website, in a video).
3. Revise: the right to adapt, adjust, modify, or alter the content itself (e.g., translate the content into another language).
4. Remix: the right to combine the original or revised content with other open content to create something new (e.g., incorporate the content into a mash-up).

5. Redistribute: the right to share copies of the original content, your revisions, or your remixes with others (e.g., give a copy of the content to a friend). (Wiley, n.d.)

### *Open data-driven learning systems in specialised language education*

Concerning the use of corpus-based language teaching materials in language instruction, Tim Johns is often regarded as the pioneer in the field, coining the term DDL to refer to the method of inferring the rules of language by directly observing them in corpora using text analysis tools. He affirmed that by discovering the rules of language underlying real samples extracted from corpora, learners become “language detectives” (Johns, 1997, p. 101). The term DDL was revisited by Boulton (2011), who considers Johns’ definition of DDL as too broad to be systematized. Boulton also offers some of the most comprehensive overviews of research carried out in DDL and identifies the number of experiments in the field of legal English as quite reduced (Boulton, 2011).

### *Research questions*

This identifiable lack was a motivating factor for conducting the experiment described below in response to the following research questions. They arose from the planning, implementation, and analysis of the data obtained from our experiment:

1. To what extent can the digital commons of open and authentic content enrich data-driven learning across formal and informal language learning?
2. What effect does the application of DDL methods for querying open and authentic content have on the acquisition of specialized terminology, as opposed to accessing non-DDL-based online resources?

Throughout this chapter, we will refer to the Law Collections in FLAX, which are derived from openly licensed pedagogic texts and open access publications from law education and research, along with legal code and judicial hearings from case law available in the public domain. In the area of legal English, as with many areas of ESAP, corpora and published language learning resources are too scarce, too small, too generic, and in most cases inaccessible due to licensing restrictions or cost. The Law Collections in FLAX demonstrate the potential for

engagement with diverse higher education audiences by drawing attention to the growing digital commons of openly available and high-quality authentic texts, which can be mined by DDL approaches to render them linguistically accessible, discoverable, and adaptable for further remixing in ESAP education.

This inquiry is directly concerned with the scalability of DDL applications and their potential to take root across both informal online learning and formal classroom-based language learning (see the de Groot chapter from this book). We also contend that our open research and development methodology enables critique by relevant stakeholders within the fields of language education, applied corpus linguistics, and now open and distance education.

### ***Tools in this study***

#### *Transcending concordance: Augmenting academic text with FLAX*

Many language learners consult concordancers. Although successful outcomes are widely reported, learners face challenges when using such tools to seek lexicogrammatical patterns. Concordancers are popular tools for supporting language learning. They allow learners to access, analyze, and discover linguistic patterns in a particular corpus, which can be chosen to match the task at hand. Researchers report positive responses from students using concordance data for checking grammatical errors, seeking vocabulary usage, and retrieving collocations (Gaskell & Cobb, 2004; O'Sullivan & Chambers, 2006; Varley, 2009; Yoon & Hirvela, 2004).

However, these tools were originally designed for linguistic analysis by professionals, and not all their facilities can be easily navigated and investigated by language learners. Learners are often overwhelmed by the vast amount of data returned. The presentation of concordance lines appears random, with no discernable ordering. It is challenging and time-consuming to go through lines of text to identify patterns. Learners may pick up a rare exceptional case for a rule and over-generalize it. Advanced search options, for example, seeking the verb collocates of a word, are sometimes provided but expressed in a syntax that requires specialized knowledge and varies among concordance providers.

Some researchers suggest that concordance data be screened before being presented to students (Varley, 2009). Others ask for commonly used linguistic patterns to be made more accessible (Coxhead & Byrd, 2007), perhaps through a simple interface for retrieving collocations (Chen, 2011). Consequently, the tool described in this chapter was conceived as a



solution to these shortcomings, making it easier for language learners to seek language patterns by going far beyond simply returning concordance lines. The FLAX system supports the following functions and presents a design departure from traditional concordancer interfaces for (1) checking vocabulary usage, (2) seeking grammatical patterns, (3) looking up collocations and lexical bundles, and (4) glossing and augmenting full-text documents with additional open and multi-media resources.

By way of introduction, FLAX is an automated scheme that extracts salient linguistic features from text and presents them in an interface designed specifically for language learners. An extraction method was developed to build the Law Collections, which identifies typical lexicogrammatical features of any word or phrase in the corpora. For example, as shown in Figure 32, learners can search at the article, paragraph, sentential, or collocational level, highlighting search terms in colour. Clicking on the colour arrows at the end of the sentences enables learners to move up a resource granularity level, for example, to the paragraph level, to enable the inspection of search terms along with their contextual information.

The screenshot displays the CopyrightX (Harvard University) website. The header includes navigation links: About Collection, Search, Lectures, Readings, Activities, Collocations, Wordlist, and LexicalBundles. A search bar is present with the text "Search for sentences that contain the word creative". Below the search bar, it states "Search Result: 142 sentences found". The results are grouped by patterns, showing several sentences with the word "creative" highlighted in yellow. Each sentence has a small green arrow icon at the end, indicating a link to more context. The sentences include:

- There remains a narrow category of works in which the creative spark is utterly lacking or so trivial as to be virtually nonexistent.
- By now, you should be familiar with the rules governing what types of creative works are subject to copyright protection, who owns the copyrights on those things, and how one secures and transfers a copyright.
- Surprisingly often, evidence survives of a defendant's conscious, intentional use of the plaintiff's work during the course of his own creative processes, and that evidence turns up in discovery.
- As you might expect, when expanding or revising copyright law, lawmakers frequently deploy arguments concerning how much legal protection creative works ought to get.
- So now let's turn to the topic for today, which, as I've indicated, is copyright theory. As I just mentioned, by theory, I mean nothing more than arguments concerning when and why copyrights should be created and what the scope or limits of those rights should be. Before zeroing in on the specific theories that will occupy us during this lecture, I should say a bit more about what these theories do and why they matter. Here's the first and most obvious role of theory. As you might expect, when expanding or revising copyright law, lawmakers frequently deploy arguments concerning how much legal protection creative works ought to get. When engaged in debates of these sorts, lawmakers or their advisers sometimes advert directly to formal theories developed by economists, political theorists, and philosophers that address that issue, and even more often, to less formal variants of those theories in general circulation in popular discourse. The result is that, in order to understand how copyright law has assumed over time its current form and how it's likely to evolve in the future, you need to know, among other things, the content of the theories that the lawmakers in practice attend to.
- 183 [14] Amici McNealy and Sutphin explain that "a quick examination of other programming environments shows that creators of other development platforms provide the same functions with wholly different creative choices."

Figure 32 Keyword search for “creative” in the CopyrightX collection

FLAX first facilitates the retrieval of typical words or phrases by grouping concordance data and sorting search results to show the most common patterns first. Second, it incorporates

grammar rules involving prepositions, word inflection, and articles, and it makes common patterns stand out. Third, it retrieves collocations and lexical bundles according to part-of-speech tags—for example, all adjectives associated with a particular noun—without using any special syntax. Fourth, it links texts to larger corpora, such as the FLAX LC and Wikipedia to provide further examples of collocates and to gloss key terms. FLAX is available on the web for anyone to use. Its design, with regard to the Law Collections in FLAX, is illustrated in this chapter. However, this method can be applied to any specialized corpus, including samples of writing collected by an individual teacher (provided they are available electronically for reuse) or writing completed by students.

### *Research on academic text*

Academic text has considerable value for supporting ESAP, and many pedagogical implications have arisen from studies of academic corpora. Although specificity in academic text has received much attention in the research literature, the findings have not been fully exploited in language teaching and learning practice. Suggestions from the research literature, for example, for bridging the gap between expert and novice academic English language proficiency include helping students appreciate the importance of common collocates and recurring lexical and grammatical patterns in different contexts (Coxhead & Byrd, 2007), making commonly used lexical bundles more accessible (Hafner & Candlin, 2007), and providing more realistic models for students (Hyland, 2008a). Emphasis in this study has therefore been placed on supporting the acquisition of specialized terminology from academic text. Also highlighted in this research, are the affordances of open and authentic texts for increased uptake by practitioners in the design and application of DDL methods in teaching and language materials development, for imparting the learning of specialized terminology in ESAP.

*Words and wordlists.* Great emphasis has been placed on identifying the language features of academic texts. Coxhead (2000) developed the Academic Word List (AWL), a list of 570 academic word families from a 3.5 million-word corpus of academic writing, which has become a widely used resource for teachers and students. Computer tools, such as the Vocab profiler available at the Compleat Lexical Tutor website, help teachers and learners analyze the vocabulary in a text with reference to the AWL and other wordlists. Certainly, learning

vocabulary involves far more than simply memorizing words in lists or looking them up in dictionaries. Users can explore the most frequent one to two thousand words from the general service list, and academic words from the AWL. Clicking the Wordlist tab in the CopyrightX collection menu, as shown in Figure 33, yields the different wordlist options.

CopyrightX (Harvard University)

About Collection

Search

Lectures

Readings

Activities

Collocations

Wordlist

LexicalBundles

My Cherry Basket

academic Words

sort by frequency

section	887	author	414	issue	305	create	279	license	211
theory	195	require	191	grant	187	lecture	180	legal	176
factor	171	design	164	evidence	155	element	149	conclude	149
creative	146	exclusive	146	derivative	146	access	141	seek	139
publish	135	principle	132	specific	129	code	128	approach	124
requirement	124	constitute	120	involve	119	similarity	119	summary	118
available	118	version	117	individual	114	benefit	113	identify	113
computer	110	display	109	distribute	109	potential	104	obtain	103
series	103	image	102	portion	100	definition	99	file	98
creation	98	aspect	97	feature	96	establish	96	distribution	96

Figure 33 Most frequent Academic Word List items in the CopyrightX collection

*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. Topic-specific corpora are therefore valuable resources that help learners build up collocation knowledge within the areas that concern them.

With FLAX, learners can browse as well as search collocations. Figure 31 shows some of the Top 100 collocations in the British Law Reports Corpus (BLaRC) to enable ready identification of useful patterns in the corpus by users. They are grouped under tabs that reflect the syntactic roles of the associated word or words, of which the first four can be seen here grouped under the “Adjective + Preposition + Noun” tab, along with their contexts. The “cherries” icon links to the collocations associated with particular a word, enabling learners to harvest and save collocations to “My Cherry Basket”.

The underlined words in Figure 34, for example relevant to the question, are also hyperlinked to entries for those words in an external collocation database built from all the written texts in the BNC. For example, clicking relevant in Figure 34 generates a further popup, shown in Figure 35, that lists *relevant to the case*, *relevant to the needs*, *relevant to the study*, etc., along with their frequency in that corpus. Furthermore, samples of these collocations in context can be seen by

clicking on them in Figure 34, which displays relevant extracts from a choice of three corpora in the FLAX LC system: the BNC, the BAWE corpus, and a Wikipedia corpus. For example, clicking *relevant to the study* brings up 22 sentences that use this phrase.

**British Law Report Corpus (BLaRC)**

About Collection Search Browse Collocations Wordlist LexicalBundles My Cherry Basket

**Browse Collocations in Collection**

a b c d e f g h i j k l m n o p q r s t u v w x y z Top 100

**Top 100 collocations**

Noun+Noun (100) Adjective+Noun (100) Noun+of+Noun (100) Verb+Noun (100) Verb+Preposition+Noun (100) Adjective+to+Verb (100)

Adjective+Preposition+Noun (100)

- pursuant to section (183)
- wrong in law (107)
- such as the present (92)
- incapable of work (62)
- relevant to the question (62)
  - "It is not possible or desirable to produce an exhaustive list of all circumstances that are or may be relevant to the question of whether the Secretary of State to detain a person pending deportation pursuant to paragraph 2(3) of Schedule 3 to the Immigration Act 1971."
  - (2) A State to which this Part applies shall be treated, in so far as relevant to the question mentioned in sub-paragraph (1), as if it were a State to which this Part applies."
  - 10. Regulation 7(9)(b) of the 1999 Regulations – to which regulation 7(2)(a) must be regarded as subject if it would otherwise be the case – paragraph (9)(b) makes more specific provision as to the effective date of supersessions in attendance allowance and disability living allowance. A change of circumstances is relevant to the question of entitlement to a particular rate of benefit – ensures that a supersession decision awarding a higher rate of benefit is effective from the end of the three-month qualifying period, provided the relevant change of circumstances leading to the award of the highest rate is reported before a month after that date (i.e., within four months of the change of circumstances itself).

Add collocation to My Cherry Basket

Add Collocation Create a New Category

relevant to the question

Choose a category for this collocation:

No category

Figure 34 Preview of some of the top 100 collocations in the British Law Report Corpus (BLaRC) displaying “relevant to the question”

**Learning Collocations**

relevant in contemporary English (Wikipedia) go

Family words: irrelevance irrelevant relevance Synonyms Antonyms

**used as an adjective**

Category	Collocation	Count	Collocation	Count
relevant + noun	relevant information	581	relevant data	116
	relevant English-language forum	443	relevant authorities	113
	certain relevant categories	186	relevant section	108
	relevant issues	159	relevant legislation	100
	relevant documents	130	relevant facts	94
adverb + relevant	particularly relevant	193	clinically relevant	89
	socially relevant	127	directly relevant	68
	highly relevant	98	culturally relevant	61
	less relevant	95	biologically relevant	31
	especially relevant	94	morally relevant	29
relevant + preposition + noun	relevant to the case	31	relevant to the development	18
	relevant to the case	26	relevant to the topic	18
				17
				17
				16
adjective + relevant				9
				8
				8

Domestic surveillance by the Army may be relevant to the case against Hiss.

If they react strongly to the guilty information, then proponents of the test believe that it is likely that they know facts relevant to the case.

In controversial cases, it may be written into a settlement that both sides keep its contents and all other information relevant to the case confidential.

They require the trial judge to act as a gatekeeper before admitting the evidence, determining that the evidence is scientifically valid and relevant to the case at hand.

However, federal courts will refuse to force journalists to reveal sources, unless the information the court seeks is highly relevant to the case, and there's no other way to get it.

Figure 35 Related collocations for the word “relevant” linked in from the FLAX LC system

*Lexical bundles.* To become proficient in ESAP, learners need to develop a repertoire of discipline-specific phrases. Recently, Biber and his colleagues developed the notion of “lexical bundles,” which are multi-word sequences with distinctive syntactic patterns and discourse functions commonly used in academic prose (Biber & Barbieri, 2007; Biber, Conrad, & Cortes, 2004). Typical patterns include noun phrase + of (*the end of the, the idea of the*, as shown in Figure 36), 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 seen 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, inference, 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 disciplines. This variation suggests the need for learners to understand relevant discourse features in their subject domain.

The screenshot shows the interface of the 'English Common Law MOOC (University of London with Coursera)'. At the top, there is a navigation bar with links: 'About Collection', 'Search', 'Lectures', 'Quizzes', 'Extras', 'Activities', 'Collocations', 'Wordlist', and 'LexicalBundles'. A 'My Cherry Basket' icon is also present. Below the navigation bar, the 'LexicalBundles' section is active, displaying a list of bundles under the heading 'At the beginning' and 'In the middle'. The bundles listed are:

- of the common law (32)
- the way in which (29)
- of the common law. (21)
- the law of the (21)
- of the House of (19)
- the idea of the (15)

The bundle 'the idea of the' is selected, and its function is detailed in the text below. The text explains that this bundle is used to introduce a topic or to frame an argument. It provides examples from the text, such as 'the idea of the rule of recognition' and 'the idea of the stability of government'. The text also notes that this bundle is used to introduce a topic or to frame an argument.

Figure 36 Lexical bundles function in the English Common Law MOOC collection



*Augmenting text with Wikification.* FLAX also interfaces with the Wikipedia Miner tool (Milne & Witten, 2013) to extract key concepts and their definitions from Wikipedia articles. Wikification in FLAX acts as a glossary tool for learners, promoting reading and vocabulary acquisition in domain-specific areas, as seen in Figure 37 with the wikify function.

The wikification process goes as follows. 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. Third, the most salient linked (and disambiguated) concepts are selected to include in the output. For example, *Stare decisis*, *Qiyas*, *Common law*, *Certiorari*, and *Lower court* in the lecture document in Figure 37 are identified in FLAX as Wikipedia concepts. A definition for *precedent* is also extracted by the Wikipedia Miner, as shown in Figure 37, within the *English Common Law MOOC* collection.

The screenshot displays the 'English Common Law MOOC (University of London with Coursera)' interface. At the top, there is a navigation bar with links: 'About Collection', 'Search', 'Lectures', 'Quizzes', 'Extras', 'Activities', 'Collocations', 'Wordlist', 'LexicalBundles', and a 'My Cherry Basket' icon. Below the navigation bar, a link '<=Back to document list' is visible. The main content area is titled 'Lecture 4.4: When precedent does not bind'. It features a video player with a thumbnail of a man speaking and the text 'When precedent doesn't bind'. Below the video player, there is a 'wikify' button and a list of related topics: 'Stare decisis', 'Qiyas', 'Common law', 'Certiorari', and 'Lower court'. The text below the video player is a snippet from a lecture, discussing the concept of precedent and its application in the Court of Appeal. The text is highlighted in green, indicating it has been wikified.

Figure 37 Wikify glossary function in the English Common Law MOOC collection

## ***Methods***

### ***Participants***

The experiment described herein was conceived as a method to measure quantitatively the usefulness and effectiveness of employing a corpus-based online learning platform, FLAX, in the teaching of legal English. To that end, a group of 52 students in the fourth year of the Translation Degree program at the University of Murcia (Spain) were selected as informants. All the students' linguistic competence level complied with the Common European Framework of Reference for Languages requirements for the B2 level. Our initial intention was to incorporate FLAX as part of the course methodology itself, trying not to alter the original syllabus of the subject in its essence.

### ***Procedure***

The informants were asked to write an essay on a given set of legal English topics (see Appendix G), defined by the subject instructor as part of their final assessment. They were then divided into two groups. The experimental group (16 informants organized into four sub-groups) were requested to only consult the FLAX English Common Law MOOC collection as the single source of information to draft their essays. The remaining 36 students (divided into nine different sub-groups) would act as the control group, following the traditional method for the design and drafting of essays before this experiment was carried out, that is, using any information source available.

The students' essays provided the database for two small learner corpora. The difference in the number of students in the control and experimental groups resulted from the fact that only two-thirds of the essay topics suggested by the subject instructor prior to the experiment were covered by the content of the English Common Law MOOC collection in FLAX.

## ***Results***

The quantitative analysis of the two corpora yielded results which reinforced our belief that the use of a corpus-based learning platform like FLAX may be a good methodological choice for the legal English instructor to complement more traditional teaching methods employed in the ESAP classroom.

### *Corpora description and methods of analysis*

Once the essays were completed, they were divided into two small learner corpora whose size differed considerably. The FLAX-based corpus contained 16,939 tokens, while those texts not based on consulting FLAX amounted to 55,030. The term “type” refers to every different word in a corpus, whereas “token” stands for the number of repetitions of the same word within it. The former corpus was articulated into four texts, whereas the latter comprised nine. (Each of these texts corresponds with the essays assigned to the experimental and control groups respectively.) Both corpora were processed automatically using Scott’s (2008) Wordsmith Tools software, with the aim of extracting information that could allow us to measure the degree of effectiveness in the use of FLAX as an experimental learning method. The texts were analysed quantitatively by applying corpus linguistics techniques for the exploration of the lexical level of the language, focusing on specialized term usage.

### ***Analysis and Discussion***

#### *Specialised term usage*

On a lexical level, the parameter that was measured as part of the quantitative analysis was term usage. To that end, both corpora were analysed using Drouin’s (2003) TermoStat, an online Automatic Term Recognition method (ATR henceforth). According to Marín (2014), this method turned out to be the most efficient method in the extraction of legal terms from an 8.85 million-word legal corpus, the BLaRC, reaching a peak precision rate of 88% for the top 200 candidate terms. Automatic identification of terms from the BLaRC employing the ATR method confirmed them as true terms after comparing them with a legal English glossary.

TermoStat mined 226 specialised terms from the learner corpus based in FLAX and 405 from those texts not using FLAX as reference. The difference in size between the two corpora, and the fact that the number of topics covered by the non-FLAX based corpus was twice as big as the other corpus, led to a size reduction of the former corpus (non-FLAX) with the aim of making the results comparable. Applying a normalization procedure such as dividing the number of terms by the number of tokens in each corpus would have sufficed for the comparison. However, the greater number of topics in the non-FLAX corpus would have caused the results to be skewed. The higher the number of different topics in a specialised corpus (as illustrated by Table 15), the higher the number of technical terms employed in it (there are more areas and sub-areas to be



covered). Therefore, this variable also had to be taken into consideration in the calculations applied in each case. In order to try to compensate for that fact, the results were divided by the number of topics, four for the FLAX texts and nine for the non-FLAX ones.

As Table 15 shows, the term average obtained for those essays written using FLAX as a resource was 13.73 points higher than the same parameter for the non-FLAX-based corpus. It could therefore be argued that those students resorting to the FLAX English Common Law MOOC collection as an information source for the drafting of their essays displayed a greater command in the use of legal terms than those who did not. The different possibilities offered by the platform, such as the “wikify” option (allowing search for definitions or related topics to a given term) or the activities aimed at fostering the acquisition of specialised terminology, may have contributed to the greater command of employing legal terms by the experimental group.

Table 15. *Term average in each legal English learner corpus*

	FLAX Corpus	Non-FLAX Corpus
Terms Identified by TermoStat (A) (Drouin, 2003)	226	385
Corpus Size After Reduction	16,939	16,264
Number of Topics (B)	4	9
Term Average (A/B)	56.5	42.77
Standardised type/token ratio	35.3	38.63

Furthermore, Drouin’s (2003) ATR method allows for the ranking of terms according to their level of specialization, which is calculated using such values as term frequency or distribution in the general and specialised fields. The average value of this parameter also turned out to be higher for the FLAX-based corpus, reaching 14.68 against 13.37 for the non-FLAX text collection. This difference could be interpreted as a greater capacity on the part of the experimental group to express themselves more accurately through more specific terms than those in the control group. However, the difference is not substantial enough for us to be able to state this conclusion with absolute certainty. Therefore, a larger sample would thus be required to confirm our observations. Furthermore, a qualitative study of a corpus sample (instead of an

automatic analysis of the whole text collection) — examining text excerpts with regard to term usage — would also be helpful to reinforce this perception.

According to the data, the members of the experimental group appear to have acquired the specialised terminology of the area better than those in the control group, as attested by the higher term average obtained by the texts in the FLAX-based corpus (56.5) as opposed to the non-FLAX-based text collection, at 13.73 points below (see Table 15). This result goes some way toward answering our second research question on the effect of DDL methods using open and authentic content on the acquisition of specialised terminology, as opposed to using non-DDL-based online resources. Employing Drouin's (2003) TermoStat ATR method as a reference, the terms identified in the former corpus are more specialised than those in the latter; that is, they are assigned a higher specificity average value based on such data as their frequency or distribution.

However, the standardized type/token ratio assigned to each set of texts, which is often indicative of the richness of the vocabulary (the higher, the richer), is lower for the FLAX-based texts, standing at 3 points below the texts written by the control group (as shown in Table 15). Although the difference is not substantial, the proportion of different types is greater in the latter corpus and hence the greater diversity of its lexicon.

### ***Policy Implications***

Formal language teacher qualifications are still predominantly concerned with training teachers in how to adapt authentic linguistic content for classroom use with minimal attention to copyright and licensing. This training extends to the adaptation of All Rights Reserved proprietary language course books and their free supplementary resources, also intended for classroom use. A notable gap in formal language teacher education arises, however, when teachers wish to share their teaching materials, which they have developed using third-party content, on the Internet beyond the secret garden of the classroom. This gap in formal teacher education also extends to developing and sharing language corpora on the Internet where issues around copyright infringement and enforcement are more likely to arise than in schools.

Policy implications for language teacher education include the need for increased awareness of the digital commons and open licensing for developing digital literacies in online language materials development and distribution. Imparting understanding of the difference between free proprietary resources and OERs licensed with Creative Commons that afford reuse and remix is

also essential for redressing the current shortfall in formal language teacher training where understanding copyright is concerned. Indeed, we are already witnessing a growing awareness of OERs among educators outside of formal teacher training channels, and the advent of Amazon Inspire—a free service for the search, discovery, and sharing of digital OERs—will further increase this awareness especially in the K-12 sector. We are also witnessing changes in, for example, university policy on open education and in government regulation where publicly funded education initiatives for developing learning resources require open licensing with Creative Commons.

In this chapter, we have also illustrated a novel corpus-based tool, FLAX, that identifies useful lexicogrammatical patterns and extracts academic words, collocations, and lexical bundles in academic text. All these features are made easily accessible through a unified searching and browsing interface. Our goal is to make current corpus technology suitable for L2 learners, helping them seek salient language samples in academic texts during writing and editing. The design was guided by outcomes and findings recorded in the research literature, and the process is entirely automatic. It should be emphasized again that although for illustrative purposes our description has focused on particular corpora, the Law Collections in FLAX, it is certainly not restricted to those ESAP resources.

The versatility of the approach we have presented here also has wide-ranging implications regarding the adoption of open education policy across formal and informal learning contexts. The implementation of policy to encourage the practice of licensing pedagogic and academic texts with Creative Commons will ensure that high quality authentic texts are openly accessible to language teaching and research professionals for educational and research purposes. It is widely understood that English is the academic lingua franca of research and teaching. Open licensing will, therefore, have the positive effect of rendering pedagogic and academic texts as remixable for the development of authentic ESAP materials to support specialised language learning, both online and offline.

### ***Further research***

The corpus-based research presented in Study 3 with a focus on specialised term usage has been extended in collaboration with colleagues at the University of Murcia, Maria Jose Marín and María Angeles Orts, using the same learner corpora to see if further corpus linguistics methods

could throw light on the decisions made by second language learners at a pragmatic level in the deployment of metadiscourse markers (Marín, Orts & Fitzgerald, 2017). Specifically, an analysis was carried out using Hyland's (2005) taxonomy of conceptual analysis for metadiscourse markers in academic English writing.

## **Chapter 6: Conclusion**

### ***Introduction***

The individual studies presented in this thesis in collaboration with the FLAX project and relevant stakeholders are all interlinked insofar as they attempt to demonstrate the efficacy of open resources and practices in the development of DDL systems for uptake in higher education. While corpus-related studies in linguistics have increased in popularity over the last several decades with a significant increase in corpus-related publications since 2000 (Liao & Lei, 2017), it is important that the knowledge generated from this research is mobilized into the development of accessible tools and corpora for knowledge users to be able to successfully carry out data-driven applications.

The three studies presented here contribute to this goal by demonstrating how advances in TDM approaches coupled with advances in open policy in research and higher education can facilitate the development of new types of DDL systems for uptake in formal and non-formal higher education. Each study has brought DDL to the attention of new stakeholders with the specific objective of pushing at the parameters of policy to see how far the collaborations could go in this design-based research for the reuse of open access content and open educational resources in the development of language learning derivatives. Each study has employed specific methods in different research settings, resulting in unique findings. The studies presented herein have also been designed to work together as a cohesive whole. With this final chapter, I start with a brief summary of key findings and linkages within and between the three studies. Next, I present conclusions from this set of studies and discuss their original contributions to knowledge with respects to the new paradigm for open data-driven systems design in higher education that I am proposing; where the reality of much needed support with learning domain-specific terminology will only increase in demand with the growing numbers of online learners worldwide seeking and entering higher education, and where English is the academic lingua franca of much research and teaching. Following this, pedagogical implications, as well as limitations and planned future research with the FLAX project are provided.

### ***Overview of Key Findings***

Study 1 was primarily concerned with scoping out and engaging potential knowledge

organisations that produce, manage, curate or aggregate authentic open access content that was deemed to be of value for learning features of specialised varieties of English academic text. These artefacts of the academy in the form of research and pedagogic texts were demonstrated in Study 1 as having positive reuse value for applying TDM methods in the development of open data-driven language learning systems for uptake in formal and non-formal higher education.

The methodological focus of Study 1, which engaged a range of different stakeholders using design ethnography and design-based research, provided proof of concept for the different types of open tools and corpora that can be developed with TDM approaches. Two applications from the work discussed in Study 1 were carried over into Studies 2 and 3 where open corpora were implemented for evaluative purposes into non-formal and formal higher education contexts respectively. Together, all three studies highlight the importance of open educational practices for iteratively designing, developing, evaluating and continuously improving corpus-based tools and resources in collaboration with key stakeholders to increase their usability and uptake with supporting domain-specific language learning in higher education.

### ***Conclusions from the three studies***

Findings from the different studies in this thesis all point to the added value that TDM methods with authentic open access content and open educational practices afford in the design, development of data-driven language learning derivative resources for uptake in a variety of higher education contexts. The affordances offered by open policy, open licensing and reforms to copyright law for exceptions and limitations with TDM are supported by advances in NLP technologies and machine learning approaches such as those presented and evaluated in this doctoral research with the FLAX project. What is more, advances with the open infrastructure of open access, open data, open educational resources, and open-source software are identified as enabling one of the central aims of this doctoral thesis research: the mobilisation of knowledge from corpus linguistics and computer science research in the collaborative design, development and evaluation of user-friendly open tools and collections for data-driven language learning with key stakeholders for useful and scalable applications in higher education. What is clear from the various scoping and monitoring reports from UNESCO is the rapidly rising number of learners worldwide who will be looking to access higher education in all of its modalities: formal, non-formal and informal. Higher education needs bolder and more open infrastructure, therefore, to meet the needs of a growing cohort

of learners worldwide who are increasingly coming online and who will invariably require facilities for accessing and utilising domain-specific research and pedagogic content in English. The open NLP and TDM approaches presented and discussed in this research in the context of the wider open infrastructure as distilled in the FLAX project offer a departure point for DDL research to consider applications for scaling automated language learning support within the learning architecture of online learning platforms that dominate and will continue to dominate the educational landscape.

In one sense, digital library collections like the ones presented in FLAX from this research are simply web resources that are accessed through hyperlinks just like any other resource—and LMS and MOOC platforms certainly accommodate hyperlinks. However, learners must leave the LMS or MOOC platform to visit and consult external resources via hyperlinks that their course tutors have included. More importantly, this practice does not encourage course tutors to collate digital resources e.g. transcribed video lecture and reading material related to the course and present them in searchable and browsable form. Rendering documents searchable and browsable relates directly to the TDM and NLP affordances of digital libraries presented in this research that are able to capitalise on the electronic nature of documents to allow them to be reused in novel educational ways, such as the raw content material for data-driven language learning that has been presented in this thesis to assist with the acquisition of specialised English varieties.

In Study 1, automated content analysis was carried out with a unique qualitative dataset of design ethnography logs and principles generated over a number of years with key stakeholders engaged with the FLAX project. The decision to include different types of stakeholders in this ethnographic design-based research – knowledge organisations, researchers, and knowledge users – has provided unique insights into the motivations as well as the concerns faced by different participants in the research regarding open initiatives for the reuse of content in the development of derivative educational resources for DDL. The cultural norms and business models of the participating knowledge organisations and their attention to or lack thereof for developing open policies for content reuse were offset against those of participating knowledge users working within formal university EAP programs where there are notable barriers to developing and sharing corpus-informed teaching and learning materials as open educational resources. An original contribution to knowledge has been made in this overarching study with respects to working at the parameters of open policy to better understand what was possible in terms of pushing forward with the reuse and remix of open access content in data-driven language learning systems development. Many

openings were revealed through directly engaging with knowledge organisations, and those individuals working therein, to devise means for widening participation with the reuse of digital content and collections for applications in higher education.

The second group of conclusions that can be drawn from this body of research are made in reference to Study 2 with the value placed on data-driven domain-specific terminology learning support by both native and non-native English speakers in non-formal higher education (MOOCs). Study 2 makes an original contribution to knowledge by conducting research into user experiences with novel interface designs that deliver automated stand-alone language support in non-formal online learning. Reporting on research into user experiences with web-based DDL systems for those users who have not received any training in how to use the systems is an under-represented area of investigation in the DDL literature. Designing DDL systems that mimic web search engine behaviour has been one of the affordances of the Greenstone digital library software that the FLAX system is based on that transcends not only the user experience with traditional concordancers for language learning but also the user experience with LMS-type MOOC platforms for online learning.

The research presented here into user experience design differs from the existing DDL research which has thus far been limited to think-aloud protocols employed in conjunction with first-hand training in how to exploit more traditional concordance interfaces for querying corpora. In addition, user perception data collected from the surveys in Study 2 were triangulated with a user query analysis—based on an observable artefact of how non-formal online learners actually used the FLAX system over the three-year period that the three online courses were run and re-run—to examine how the system is used to search and browse course documents, and to retrieve keywords, phrases and related concepts via Wikipedia, collocations, extended collocation chunks, related collocations, and sample sentences of collocations in authentic contexts via the FLAX collocations database. Non-formal as well as informal online language learning is an under-researched area in the literature due to constraints faced with data collection. The log data analyses that are presented in Study 2, similar to traditional analyses of user queries on the Web, provide interesting and revealing insights that could not have been gained from small scale focused user studies in formal lab-based language education. To the best of my knowledge, this user query data analysis approach has not been explored in DDL research.



Study 2 also revealed that diverse motivations existed among participants for the types of non-formal learning support adopted. With specific reference to the FLAX project, externally linked open resources (Wikipedia, WordNet, the FLAX LC system etcetera) were valued highly by participants; in addition to the reported affordance of being able to search and browse through full-text course documents, which supplemented the LMS user experience with MOOCs. This last point about the limited functionality of most LMS platforms, MOOC platforms notwithstanding, gives rise to important user experience design considerations for what learners can and cannot do with existing LMS platforms.

Studies 2 and 3 demonstrate how openly licensed MOOC content expedited the development of open source learning support derivatives for non-formal online learning that could then be reused in formal language learning and translation studies. Study 3 extends this doctoral research by demonstrating the value of reusable pedagogic data and associated automated forms of corpus linguistics analyses for comparing the effects of usage of specialised legal terminology in two learner corpora: one from an experimental group of learners employing only the ECL MOOC corpus in FLAX and the other from a control group of learners employing any information source from the Internet. One of the key factors which motivated Study 3 was the fact that DDL resources and experiments in the area of legal English are scarce, indicating that this specialised English variety along with many other varieties of ESP remain underexplored in the literature.

A further conclusion from the research in Study 3 is the efficacy of using authentic data-driven pedagogic resources from the digital commons for learning the terminology of specialised English varieties in different subject domains. The open FLAX corpus used in this study positively influenced the usage of specialised legal terminology with figures from the analyses carried out indicating that the experimental group employed the specialised terminology better in their essay writing than did the control group. Many OER studies have focused exclusively on the cost reduction aspect of using and developing OERs but very few studies in open education have looked at whether OERs can improve learning performance. Although Study 3 is quite a fledgling study in many ways due to the learner corpora size, it shows great promise for the efficacy of employing open DDL approaches in specialised language learning and teaching for making an original contribution to knowledge in the area of open data-driven language learning in higher education.

### ***Implications and limitations***

This research carries several important pedagogical and policy implications for enabling the research and development of language learning derivative resources from an increasingly available tranche of open access content in higher education. Because of the potential accessibility of much of this content, and the data and metadata that supports it, for non-commercial reuse in research and education, an important and unique opportunity presents itself to those responsible for teaching and learning within formal higher education institutions, and those responsible for delivering non-formal higher education offerings online. This same opportunity for content reuse is harder to reach by commercial education publishers due to much of this content being off-limits for commercial reuse. Nonetheless, proponents lobbying within the open education and open data movements realise that the responsibility lies with knowledge organisations putting open policies into place to steward the non-commercial reuse of their valuable content by and for the education community.

Limitations specific to each study have already been discussed in the preceding chapters, which also have a bearing on broader limitations that apply to this doctoral thesis research as a whole. Findings from Studies 1 and 2 reveal that the provenance for content reuse is mixed depending on the dominant business models and organisational cultures that exist within higher education institutions and other knowledge organisations such as libraries, archives, and research aggregation services. Particularly within higher education institutions where there is more attention paid to open access research policy over and above open education policy. This discrepancy in terms of open policy has resulted in a lack of awareness by the majority of academics working in higher education for developing the necessary facility with practices in open research as well as in open education. The current emphasis on data mining with learner data for developing learner analytics, and the commercial interests in selling this data down the road to third party educational services is currently eclipsing the wider debate on data reuse and stewardship in higher education. In a similar vein, the current LMS capabilities, which have become the standard bearer in educational technology applications for higher education, dominate and limit the vision for what could be the TDM enriched and enhanced solutions for learning content management presented in this research.

### *Current and future research*

I am currently an honorary research fellow with the Department of Computer Science at the University of Waikato in Aotearoa/New Zealand working under the supervision of Emeritus Professor Ian Witten with postdoctoral funding from the Fonds de recherche du Québec – Société et culture (FRQSC). I have developed clearly delineated plans for the next few years to build on my doctoral research in designing open data-driven systems for learning domain-specific terminology in higher education. These plans are strengthened through my on-going collaboration with the FLAX team at the University of Waikato. With members of the FLAX project research group, we have begun further research into data-driven learning systems design and development with high-profile collaborators, including: FutureLearn, the British Library, the CORE open access aggregation service at the Knowledge Media Institute at the UK Open University, and leading corpus linguistics research groups at Université Paris Diderot in France and Universidad de Murcia in Spain. I also intend to continue my collaboration with key players in the open education community from around the world, including the Hewlett Foundation-funded Global OER Graduate Network (GO-GN) of which I am an alumnus.

### *F-Lingo: Scaling automated domain-specific terminology learning support in MOOC platforms*

My current research and development work with the FLAX research group is in providing powerful tools and robust corpora for informal online learning, including non-formal MOOC learning. PhD candidate, Jemma König, also working under the supervision of Professor Ian Witten, has developed F-Lingo<sup>61</sup>. Implemented as a Chrome extension, F-Lingo works on top of the FutureLearn MOOC platform to help learners with the selected words, phrases, and concepts in the texts they are reading, for example, video transcripts, course information, and course readings. Jemma König's work with F-Lingo furthers our team's research into MOOC language support, and previous research into MOODLE LMS language support (Witten, Wu & Yu, 2011), with the development of a new experimental system that draws on FLAX collections using NLP and machine learning approaches, but which has also made a significant departure from the

---

<sup>61</sup> To trial, download [F-Lingo from the Chrome store](https://chrome.google.com/webstore/search/flingo) (<https://chrome.google.com/webstore/search/flingo>) and install it. Restart your browser and visit any page of the [Data-mining with Weka MOOC](https://www.futurelearn.com/programs/data-mining) (<https://www.futurelearn.com/programs/data-mining>) from the University of Waikato; the rest happens automatically. If you want to see what F-Lingo does without installing it, this [3-minute video](https://www.youtube.com/watch?v=FRGwuexvkus) (<https://www.youtube.com/watch?v=FRGwuexvkus>) illustrates its facilities.

Greenstone digital library software. For her PhD, Jemma König is currently collecting experimental usage data and usability survey data for implementing F-Lingo in conjunction with Professor Witten's Practical Data Mining<sup>62</sup> courses with FutureLearn. In the following paragraphs, I will outline the basic functions of F-Lingo with a section on my planned contributions to the educational arm of the F-Lingo research in collaboration with the FLAX team.

Once it has been installed, if a FutureLearn course has been added to F-Lingo, it will traverse the content on its pages to highlight keywords, phrases, and concepts in the text as shown in Figure 38 with the F-Lingo menu on the right of the screen and with the phrases tab activated to highlight phrases within a FutureLearn MOOC video transcript. F-Lingo provides data-enriched browsability of course documents. It also provides an interactive interface for gaining further information about each highlighted feature, such as definitions, example sentences, and related collocations. The interactive features of highlighting and look-up are done in real time by the Chrome extension.

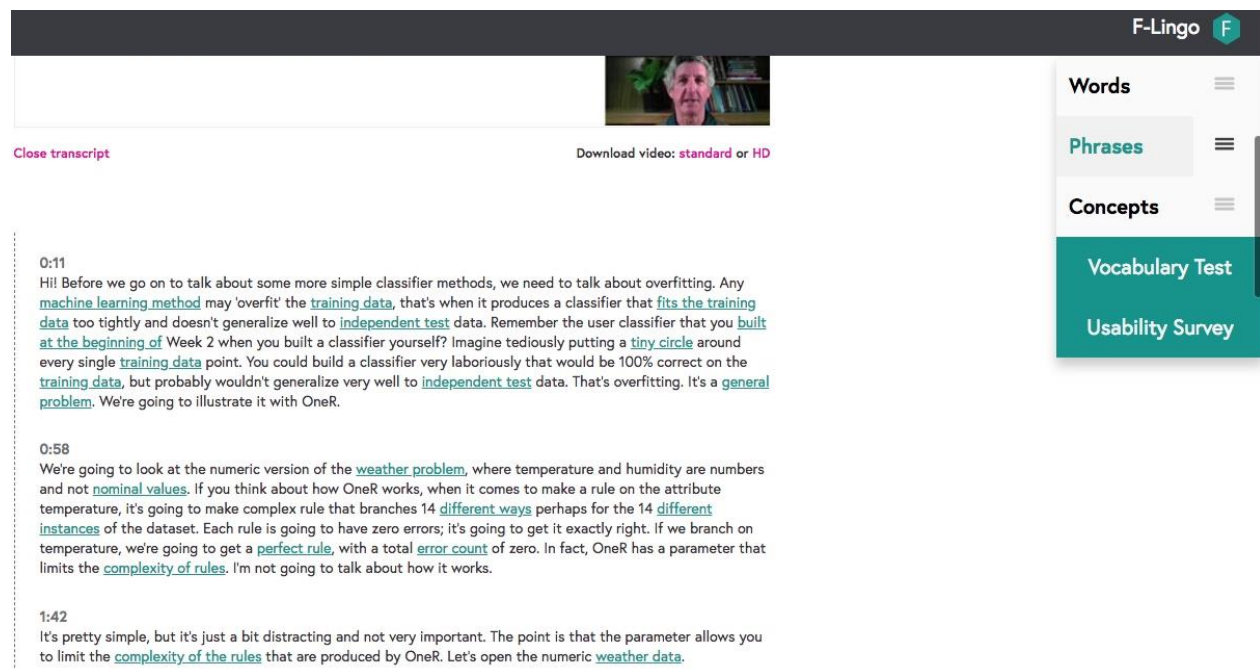





Figure 38 F-Lingo highlighted phrases in FutureLearn MOOC video transcript

<sup>62</sup> <https://www.futurelearn.com/programs/data-mining>

F-Lingo uses established frequency word lists to identify keywords within the text – classifying words as keywords only if they are absent from the General Service List (West, 1953). For keywords, definitions are retrieved from Wiktionary and example sentences are derived from both the content of the course and the PhD Abstract collections in FLAX. Next, F-Lingo uses syntactic patterns to identify collocations as phrases within the text, for example noun + noun (*data mining*), verb + noun (*visualise data*), and so on. For phrases, example sentences are derived from course content to show how they are used on the course with further example sentences and related collocations derived from the FLAX Wikipedia collection as shown in Figure 39. F-Lingo also uses the Wikipedia Miner toolkit (Milne & Witten, 2013) of machine learned approaches to detect and disambiguate Wikipedia concepts within a document as shown in Figure 40. The steps outlined here are all done offline, in a pre-processing stage.

<p><b>How is the phrase "machine learning method" used?</b> </p> <p>Examples (from this course)</p> <ul style="list-style-type: none"> <li>• How do you know how well your machine learning method is doing?</li> <li>• The success of a machine learning method depends on the domain.</li> <li>• We're going to talk about another machine learning method called the nearest neighbor, or instance-based, machine learning method.</li> <li>• The result is logistic regression, a popular and powerful machine learning method that uses the logit transform to predict probabilities directly.</li> <li>• We know that the evaluation of this machine learning method J48 on this dataset, "diabetes", gives 74.5% accuracy, probably somewhere between 73.5% and 75.5%.</li> </ul> <p>3 more</p> <p>Examples (from FLAX)</p> <ul style="list-style-type: none"> <li>• An Alternating Decision Tree (ADTree) is a machine learning method</li> <li>• A nice feature of constructive induction methods such as MDR is the ability to use any data mining or machine learning method to analyze the new representation of the data.</li> <li>• It is part of the machine learning method to reduce the risk for a SAR paradox, especially taking into account that only a finite amount of data is available (see also MVUE).</li> <li>• Generative topographic map (GTM) is a machine learning method that is a probabilistic counterpart of the self-organizing map (SOM), is provably convergent and does not require a shrinking neighborhood or a decreasing step size.</li> </ul> <p>Expanded phrases</p> <p>Machine learning method to reduce Machine learning method to analyze</p>	<p><b>What is "machine learning"?</b> </p> <p><b>Machine learning</b>, a branch of <b>artificial intelligence</b>, is a scientific discipline concerned with the design and development of <b>algorithms</b> that allow <b>computers</b> to evolve behaviors based on empirical <b>data</b>, such as from <b>sensor</b> data or <b>databases</b>. A learner can take advantage of examples (data) to capture characteristics of interest of their unknown underlying probability distribution. Data can be seen as examples that illustrate relations between observed variables. A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data; the difficulty lies in the fact that the set of all possible behaviors given all possible inputs is too large to be covered by the set of observed examples (training data). Hence the learner must generalize from the given examples, so as to be able to produce a useful output in new cases.</p> <p>Related concepts</p> <ul style="list-style-type: none"> <li>• <a href="#">Data mining</a></li> <li>• <a href="#">Natural language processing</a></li> <li>• <a href="#">Supervised learning</a></li> <li>• <a href="#">Artificial neural network</a></li> <li>• <a href="#">Computer vision</a></li> </ul> <p>13 more</p> <p></p>
<p>Figure 39 F-Lingo phrase examples for “machine learning method” derived from MOOC course content and FLAX Wikipedia collection</p>	<p>Figure 40 F-Lingo MOOC concept examples for “machine learning” mined with the Wikipedia Miner Toolkit</p>

In terms of my postdoctoral research with F-Lingo, I am currently in the process of scoping out research sites with universities offering FutureLearn MOOCs to scale the collection of

experimental usage data and usability research with F-Lingo. I intend to contribute to Jemma König's research by carrying out a study focused on self-regulated learning (SLE) in MOOCs with F-Lingo, following the work of Littlejohn et al. (2016) into SLE in the MOOC space, for supporting the learning of academic and professional terminology. Follow-up discussions with volunteer learners (who consent to be contacted via surveys), will employ think aloud techniques (Ericsson & Simon, 1987) and cognitive walkthrough for identifying learner strategies for browsing and querying the F-Lingo system, and for evaluating the usability and learnability of the MOOC content that interfaces with the F-Lingo Chrome extension via the MOOC platform, including: the FLAX ACE collections, the FLAX PhD Abstract collections, Wikipedia, and Wiktionary. My focus with F-Lingo and non-formal learners will be on supporting and investigating self-regulated learning in the MOOC space given the issues with low student retention in MOOCs where language barriers with academic English are well reported.

It is important to question how MOOC providers and MOOC course designers at leading universities around the world, many of whom invariably offer MOOCs in English, can support large and diverse learner groups using automated open data-driven language learning systems. Non-formal online learning is the activity of understanding, gaining knowledge or acquiring skills outside the remit of being a registered student at a formal educational institution. As with most MOOCs, this kind of non-formal learning typically occurs without direct teacher or tutor support, so I am especially interested in investigating whether or not stand-alone automated data-driven learning systems can assist with the learning of domain-specific academic or professional terminology. Informal and non-formal language learning are under-researched areas due to constraints faced with data collection. My planned research in this area, in collaboration with MOOC providers and universities offering MOOCs, will therefore enable data collection into an area of online research, teaching and learning that is of significance to open and distance education as well as language education.

We are also planning to make F-Lingo widely operational through performance improvement campaigns with educational technologists working with the delivery of MOOCs who can be trained in data scraping methods to enable their course content to be parsed by the F-Lingo system software for automated language learning support. Our end goal is to make F-Lingo, which draws on linguistic databases from FLAX, interoperable with any online learning platform. I view F-Lingo as a work-around solution to carrying out TDM methods on All Rights Reserved

course content in online learning that although deemed as open access in the sense of being read-only from an accessible outward-facing online learning platform (with MOOCs being a clear example), has not, and most likely will not, be licensed openly with Creative Commons for reuse and remix by the wider education and learning community due to the paucity of open education policy for content reuse and remix in higher education. The current work with F-Lingo is based on findings from this doctoral research with respects to providing proof of concept that data-driven approaches with MOOC course content are valued by non-formal learners.

*FLAX Learning Collocations system: Analysing user query data*

My planned research and development work with the FLAX research group will also provide yet more powerful tools and robust corpora for one of the most challenging areas of English language learning, collocations (sequences of words that frequently co-occur), where there are literally hundreds of thousands of possibilities for combining words and phrases. In order to achieve impact, my proposed program of research will build on my prior research with Dr. Shaoqun Wu of the FLAX team and my wider professional network within the areas of open education and second language education.

The FLAX LC system currently houses three databases built from the BAWE corpus, the BNC and a Wikipedia corpus comprised of three million articles. We conducted an initial user query analysis study, capturing user query data at scale for the period from June 2016 to June 2017 (Wu, Fitzgerald, Yu & Witten, 2019). This study not only provided suggestions for improving the usability and experience of our system, but it also revealed interesting facts on how the FLAX LC is used.

354,694 queries from 67 countries were recorded with an average of 971 queries per day. Table 16 shows the top 10 countries and corresponding percentages from the study. Queries from 57 other countries are grouped under the “Other” category. About two thirds (65%) of queries were from three English-speaking countries: The United Kingdom (28%), New Zealand (24%) and Australia (13%). The Republic of Korea is at the top of the list among all non-English-speaking countries, followed by China, Russia, Belarus and Israel.

Table 16. *Geographic distribution of FLAX LC users. Reprinted from* Wu, S., Fitzgerald, A., Yu, A., & Witten, I.H. (2019). *Developing and evaluating a learner-friendly collocation*

system with user query data. *International Journal of Computer-Assisted Language Learning and Teaching*, 9(2), pp.53-78.

Country	Percent of queries
United Kingdom	28%
New Zealand	24%
Australia	13%
Republic of Korea	9.5%
China	3.3%
Russia	3.2%
United States	2.8%
Canada	2.7%
Belarus	2.3%
Israel	1.8%
Other	9.4%

The initial study that we conducted into user query data also captured popularity scores of the uptake of the three databases—Wikipedia, BAWE, and BNC—as shown in Table 17, along with the statistics of user preferences by country. The Wikipedia database (53.2%) was the most popular, but this is most likely due to the fact that the Wikipedia corpus is the default corpus offered by the FLAX LC system, i.e. users need to select the BAWE or BNC corpora from the drop-down menu and explicitly switch to query those corpora. The BAWE corpus came in at second place and this may indicate an increased focus on learning academic English by users. The user preferences by country shows that New Zealand users preferred Wikipedia and the BNC, and that users in the Republic of Korea preferred the Wikipedia corpus. The BAWE corpus was the most popular among United Kingdom users (50.9%) where the BAWE corpus was incidentally developed at three UK universities, followed by Australian users (21.1%). The results are mixed and not distinctive among other countries. Due to the surprising popularity of the BAWE corpus, which is derived from university student writing and small in comparison with the Wikipedia corpus and the BNC, we have developed new and extensive databases that make up the ACE collections, which are derived from high-quality academic text in different disciplines. The ACE collections have been developed in response to these findings from the



year-long user query data analysis study showing an increasing preference for academic English corpora.

Table 17. *Database usages and user preferences by country. Reprinted from Wu, S., Fitzgerald, A., Yu. A., & Witten, I.H. (2019). Developing and evaluating a learner-friendly collocation system with user query data. International Journal of Computer-Assisted Language Learning and Teaching, 9(2), pp.53-78.*

Database	Percent of queries	User preferences by country	
Wikipedia	53.2%	New Zealand	26.6%
		Republic of Korea	19.2%
		United Kingdom	19.1%
		Other	35.1%
BAWE	38%	United Kingdom	50.9%
		Australia	21.1%
		New Zealand	14.5%
		Other	13.5%
BNC	8.8%	New Zealand	63.5%
		United Kingdom	6.7%
		United States	5.5%
		Other	24.3%

Our research focus will be on the uptake and utilisation of the ACE collections in the FLAX LC system to boost collocation learning support in formal and informal education. Specifically, we aim to implement and evaluate the largest open access academic English collocations corpora, ACE, with linguistic data harvested from the CORE aggregation service at the Knowledge Media institute, UK Open University with metadata and full-text content from over 135 million open access articles. The ACE collections have just been developed and are now available online alongside and within the existing FLAX LC system.

Further analyses of user query data collected by the FLAX LC system with the new ACE databases in addition to the Wikipedia and BNC databases would provide valuable information

and suggestions for DDL researchers and language teachers when supporting their learners with the study of collocations. These iterative analyses could also go some way toward answering research questions like what makes a word and its derivatives difficult to learn by examining the collocations that students have looked at, or whether the types of queries made by users are different according to different geographical regions. We have recently added new facilities to track user interactions with the system in more detail to identify patterns of users' query reformulation strategies (i.e. site searching strategies). These additional facilities will also allow us to draw a comparison between the analysis study already concluded (Wu, Fitzgerald, Yu & Witten, 2019) and a further one in a year's time, along with a more detailed comparison between users from English speaking and non-English speaking countries. We intend for these results to yield new and in-depth insights for understanding user behavior in corpus consultation.

*FLAX PhD abstract collections: Developing OERs for learning features of lexical paving*

I also plan to further iterate, implement and evaluate the open access PhD Abstract collections (Wu, Fitzgerald, Witten & Yu, 2018) into formal university academic English writing programs. The PhD abstract corpora were developed as part of my PhD research in collaboration with the British Library's Electronic Theses Online Service and EAP practitioners at Queen Mary University of London. Abstracts play a number of important roles in academic text. Identified primarily as a sub-genre (Swales and 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 (Hackin, 2001) with the growing consensus among academics that they may often be the only part of a paper read via abstracts databases. 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 collections in FLAX. Metadata, which currently includes the abstracts of 450,000 doctoral theses from UK universities, was harvested from EThOS to create the PhD abstracts collections in FLAX.

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 lexicogrammar 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 lexicogrammar 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)

I intend to carry out research with knowledge users: teachers and learners engaged in EAP programs. My research will focus on the development and evaluation of supplementary corpus-derived classroom teaching and independent learning resources for EAP programs (to be licensed and distributed as OERs) in collaboration with Doctor Geneviève Bordet of Université Paris Diderot for the uptake and utilisation of the PhD Abstract collections with language teachers and learners in EAP programs with a particular emphasis on aspects of domain-specific terminology found in STEM subjects. In particular, we will be analysing features of PhD Abstract discourse and lexicogrammatical patterns identified in the PhD Abstract collections in comparison with learner writing with reference to Bordet's research into lexical paving.

### ***Concluding remarks***

This doctoral thesis is the culmination of several years of collaborative work surveying the higher education landscape across different countries and different modalities. It has been a great privilege to work alongside thought leaders in the areas of open education, language education and computer science for devising solutions to real-world problems with access differentiation in higher education and in English language education. The topics presented in this thesis represent long standing interests. As a result, I am grateful to have had the opportunity to explore these

topics in greater detail. That being said, I am also excited to continue expanding my efforts and my focus to address additional topics with real-world implications that drive my passion in service to the field of education.

## References

- Ädel, A. (2010). Using corpora to teach academic writing. In M.C. Campoy Cubillo, B. Bellés Fortuño & M.L. Gea-Valor (Eds.), *Corpus-based Approaches in English Language Teaching* (pp. 39–55). London: Continuum.
- Alcorn, B., Christensen, G., & Kapur, D. (2015) Higher education and MOOCs in India and the global south. *Change: The Magazine of Higher Learning*, 47(3), 42–49. DOI: 10.1080/00091383.2015.1040710
- Altbach, P. G., Reisberg, L., & Rumbley, L. E. (2009). *Trends in Global Higher Education: Tracking an Academic Revolution*. A Report prepared for the UNESCO 2009 World Conference on Higher Education. Retrieved from <http://unesdoc.unesco.org/images/0018/001832/183219e.pdf>
- Amiel, T., & Reeves, T. C. (2008). Design-based research and educational technology: Rethinking technology and the research agenda. *Educational Technology & Society*, 11 (4), 29–40.
- Ananiadou, S., & Mcnaught, J. (Eds.) (2006). *Text mining for biology and biomedicine*. Boston, MA/London, UK: Artech House.
- Ananiadou, S., Rea, B., Okazaki, N., Procter, R., & Thomas, J. (2009). Supporting systematic reviews using text mining. *Social Science Computing Review*, 27, 509–523. doi:10.1177/0894439309332293
- Ananiadou, S., Thompson, P., Thomas, J., Mu, T., Oliver, S., Rickinson, M., Sasaki, Y., Weissenbacher, D., & McNaught, J. (2010). Supporting the education evidence portal with text mining. *Philosophical Transactions of the Royal Society*, 368, 3829–3844. doi:10.1098/rsta.2010.0152
- Angus, D., Rintel, S., & Wiles, J. (2013). Making sense of big text: A visual-first approach for analyzing text data using Leximancer and Discursis. *International Journal of Social Research Methodology*, 16(3), 261–267.
- Anderson, T. (2007). Design-Based Research: A new research paradigm for open and distance learning. Keynote address given at the Community of Inquiry Chais Research Centre Annual Education Technology Conference. The Open University of Israel, Ra'anana, Israel. Retrieved from <http://www.slideshare.net/terrya/design-based-research-new-research-paradigm>

- Anderson, T., & Shattuck, J. (2012). Design-Based Research: A decade of progress in education research? *Educational Researcher*, 41(1), 16–25.
- Anthony, L. (2014, July). A view to the future in corpus tools development. 11th Teaching and Language Corpora Conference (TALC 11) Keynote Address. Lancaster University, UK.
- Anthony, L. (2018). *Introducing English for specific purposes*. London: Routledge.
- Argyris, C. & Schön, D. (1991). Participatory action research and action science compared. In W.F. Whyte (Ed.), *Participatory action research* (pp. 85–96). Newbury Park, NJ: Sage.
- Arum, R., & Roska, J. (2011). *Academically adrift: limited learning on college campuses*. Chicago: Chicago University Press.
- 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.
- Atenas, J. (2015). Model for democratisation of the contents hosted in MOOCs. *RUSC Universities and Knowledge Society Journal*, 12(1), 3–14. doi <http://dx.doi.org/10.7238/rusc.v12i1.2031>
- Bainbridge, J., Melitski, J., Zahradnik, A., Lauria, E., Jayaprakash, S., & Baron, J. (2015). Using learning analytics to predict at-risk students in online graduate public affairs and administration education. *Journal of Public Affairs Education*, 21(2), 247–262.
- Balfour, S.P. (2013). Assessing writing in MOOCs: Automated essay scoring and calibrated peer review™. *Research & Practice in Assessment*, 8, 40–48.
- Barab, S., & Squire, L. (2004). Design-based research: Putting a stake in the ground. *The Journal of the Learning Sciences*, 13(1), 1–14.
- Barefoot, B. (2004). Higher education's revolving door: confronting the problem of student drop out in US colleges and universities. *Open Learning*, 19(1), 9–19.
- Baskerville, R.L., & Myers, M.D. (2015). Design ethnography in information systems. *Information Systems Journal*, 25, 23–46.
- Bayne, S., Knox, J., & Ross, J. (2015). Open education: the need for a critical approach. *Learning, Media and Technology*, 40(3), 247–250. DOI: 10.1080/17439884.2015.1065272
- Belcher, D. (2010). What ESP is and can be: An introduction. In D. Belcher (Ed.), *English for specific purposes in theory and practice* (pp. 1-20). Ann Arbor, MI: University of Michigan Press.

- Bell, P. (2004). On the theoretical breadth of design-based research in education. *Educational Psychologist*, 39(4), 243–253.
- Benkler, Y. (2007). *The wealth of networks: How social production transforms markets and freedom*. New Haven, CT: Yale University Press.
- Benson, M., Benson, E., & Ilse, R.F. (1986). *The BBI combinatory dictionary of English: A guide to word combinations*. Amsterdam/Philadelphia: John Benjamins.
- Bereiter, C. (2002). Design research for sustained innovation. *Cognitive Studies, Bulletin of the Japanese Cognitive Science Society*, 9(3), 321–327.
- 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, Netherlands: Rodopi.
- Bernardini, S. (2004). Corpora in the classroom: An overview and some reflections on future developments. In J. McH. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (pp.15–36). Amsterdam: John Benjamins.
- Bhatia, V.K. (1993). *Analysing genre: Language use in professional settings*. London: Longman.
- Biber, D., Conrad, S., & Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25, 371–405.
- Biber, D. (2006). *University language: a corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.
- Biber, D., & Kurjian, J. (2007). Towards a taxonomy of web registers and text types: A multi-dimensional analysis. *Language and Computers*, 59(1), 109–130.
- Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purpose*, 26, 263–286.
- Bichard, J., & Gheerawo, R. (2010). The ethnography in design. In A.J. Clarke (Ed.), *Design anthropology: Object culture in the 21<sup>st</sup> century*. New York: SpringerWien.
- Biesta, G. (2010). *Good Education in an Age of Measurement: Ethics, Politics, Democracy*. Boulder, CO: Paradigm.
- Bishop, H. (2004). The effect of typographical salience on the look up and comprehension of unknown formulaic sequences. In N. Schmidt (Ed.), *Formulaic sequences: acquisition, processing and use* (pp. 227–244). Philadelphia, PA: John Benjamins.

- Blomberg, J. (1993). Ethnographic field methods and their relation to design. In D. Schuler & A. Namioka (Eds.), *Participatory design: principles and practices*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Bollier, D. (2007). The growth of the commons paradigm. In C. Hess & E. Ostrom (Eds.), *Understanding knowledge as a commons: From theory to practice* (pp. 27–40). Cambridge, MA: MIT Press.
- 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). Peter Lang: Bern.
- Bordet, G. (2015). The role of “Lexical Paving” in building a text according to the requirements of a target genre. In P. Thompson & G. Diani (Eds.), *English for Academic Purposes: Approaches and Implications* (pp. 43–66). Newcastle upon Tyne, England: Cambridge Scholars Publishing.
- Borgman, C. (2015). *Big data, little data, no data: Scholarship in the networked world*. Cambridge, Massachusetts: MIT Press
- Borthwick, K., & Gallagher-Brett, A. (2014). Inspiration, ideas, encouragement: teacher development and improved use of technology in language teaching through open educational practice. *Computer Assisted Language Learning*, 27(2), 163–183. DOI: 10.1080/09588221.2013.818560
- Bosman, J. (2012, March 13). After 244 years, Encyclopaedia Britannica stops the presses. *The New York Times*.
- Boulton, A. (2009). Testing the limits of data-driven learning: language proficiency and training. *ReCALL*, 21(1), 37–54.
- Boulton, A. (2010a). Data-driven learning: Taking the computer out of the equation. *Language Learning*, 60(3), 534–572.
- Boulton, A. (2010b). Data-driven learning: On paper, in practice. In T. Harris & M. Moreno Jaén, (Eds.). *Corpus linguistics in language teaching* (pp. 17–52). Bern: Peter Lang.
- Boulton, A. (2011). Data-driven learning: The perpetual enigma. In S. Roszkowski & B. Lewandowska-Tomaszczyk (Eds.), *Explorations Across Languages and Corpora* (pp. 563–580). Frankfurt, Germany: Peter Lang.
- Boulton, A. (2012a). Beyond concordancing: Multiple affordances of corpora in university



- language degrees. *Languages, Cultures and Virtual Communities*, 34, 33–38.
- Boulton, A. (2012b). Hands-on / hands-off: Alternative approaches to data-driven learning. In J. Thomas & A. Boulton (Eds.), *Input, process, and product: Developments in teaching and language corpora* (pp. 152–168). Brno, Czech Republic: Masaryk University Press.
- Boulton, A., & Thomas, J. (2012). Corpus language input, corpus processes in learning, learner corpus product. Introduction to J. Thomas & A. Boulton (Eds.), *Input, Process and Product: Developments in Teaching and Language Corpora* (pp. 7–34). Brno: Masaryk University Press.
- Boulton, A. (2013). Wanted: Large corpus, simple software. No timewasters. In A. Leńko-Szymańska. *Proceedings of TaLC10: 10th International Conference on Teaching and Language Corpora* (pp.1–6). Warsaw, Poland: University of Warsaw.
- Boulton, A., & Pérez-Paredes, P. (2014). ReCALL special issue: Researching uses of corpora for language teaching and learning Editorial Researching uses of corpora for language teaching and learning. *ReCALL*, 26(2), 121-127. doi:10.1017/S0958344014000068
- Boulton, A. (2015). Applying data-driven learning to the web. In A. Lenko-Szymanska, & A. Boulton (Eds.), *Multiple Affordances of Language Corpora for Data-driven Learning* (pp. 267–296). Amsterdam: John Benjamins.
- Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning* 67(2), 348–393.
- Bowen, G. (2006). Grounded theory and sensitizing concepts. *International Journal of Qualitative Methods*, 5, 12–23.
- Brabazon, T. (2006). The Google effect. *Libri*, 56(3), 157–167.
- Brabazon, T. (2013). *Digital dieting: From information obesity to intellectual fitness*. London & New York: Routledge Taylor & Francis Group.
- British Library. (n.d.). EThOS Toolkit | Re-use by external services: EThOS as a data provider: Metadata. Retrieved from <http://ethostoolkit.cranfield.ac.uk/tiki-index.php?page=Re-use+by+external+services>
- British National Corpus, version 3 (BNC XML Edition). (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Retrieved from <http://www.natcorp.ox.ac.uk>

- Brown, A.L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The Journal of the Learning Sciences*, 2(2), 141–178.
- Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., & Seaton, D. T. (2013). Studying learning in the worldwide classroom: Research into edX’s first MOOC. *Journal of Research & Practice in Assessment*, 8, 13–25.
- Budapest Open Access Initiative (BOAI). (2002). BOAI declaration. Retrieved from <http://www.budapestopenaccessinitiative.org/read>
- Burkhardt, H. (2006). From design research to large-scale impact. In J. van den Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *Educational design research* (pp. 185–228). London: Routledge.
- Campbell, C., Pitt, L., Parent, M., & Berthon, P. (2011). Understanding consumer conversations around ads in a web 2.0 world. *Journal of Advertising*, 40(1), 87–102.
- Campbell, L. (2013, June 5). What do FutureLearn’s terms and conditions say about open content? [Blog post]. Retrieved from <http://blogs.cetis.org.uk/lmc/2013/06/05/what-do-futurelearns-terms-and-conditions-say-about-open-content/>
- Cape Town Open Education Declaration. (2007). Cape Town open education declaration: Unlocking the promise of open educational resources. Retrieved from <http://www.capetowndeclaration.org/read-the-declaration>
- Carroll, J.B. (1964). Words, meanings, & concepts. *Harvard Educational Review*, 34, 178–202.
- Castilho, S., Gaspari, F., Moorkens, J., & Way, A. (2017). Integrating machine translation into MOOCs. *Proceedings of the EDULEARN 2017 Conference* (pp. 9360–9365). Barcelona: Spain.
- Chalmers, I. (2003). Trying to do more good than harm in policy and practice: the role of rigorous, transparent, up-to-date evaluations. *The Annals of the American Academy of Political Science*, 589, 22–40. doi:10.1177/ 0002716203254762
- Chambers, A., & O’Sullivan, I. (2004). Corpus consultation and advanced learners’ writing skills in French. *ReCALL*, 16(1), 158–172.
- Chang, J.-Y. (2014). The use of general and specialized corpora as reference sources for academic English writing: A case study. *ReCALL*, 26(2), 243–259.

- Charles, M. (2012). Proper vocabulary and juicy collocations: EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes*, 31(2), 93–102.
- Charles, M. (2015). Same task, different corpus: The role of personal corpora in EAP classes. In A. Lenko-Szymanska, & A. Boulton (Eds.) *Multiple affordances of language corpora for data-driven learning* (pp. 131–153). Amsterdam: John Benjamins.
- Charles, M., & Pecorari, D. (2016). *Introducing English for academic purposes*. London: Routledge.
- Chen, H. H. (2011) Developing and evaluating a web-based collocation retrieval tool for EFL students and teachers. *Computer Assisted Language Learning*, 24(1), 59–76.
- Chemers, M. M., Hu, L., & Garcia, B. F. (2001). Academic self-efficacy and first year college student performance and adjustment. *Journal of Educational Psychology*, 93, 55–64.  
doi:10.1037/0022-0663.93.1.55
- Chesbrough, H., Vanhaverbeke, W., & West, J. (2008). *Open innovation: Researching a new paradigm*. Oxford University Press: Oxford.
- Chinnery, G. (2008). You've got some GALL: Google-assisted language learning. *Language Learning & Technology*, 12(1), 3–11.
- Christensen, C.M. (1997). *The innovator's dilemma: When new technologies cause great firms to fail*. Boston, MA: Harvard Business School Press.
- Christensen, C.M., Raynor, M., & McDonald, R. (December 2015): What is disruptive innovation? *Harvard Business Review* 93(12), 44–53.
- Christensen, G., Steinmetz, A., Alcorn, B., Bennett, A., Woods, D., & Emanuel, E.J. (November 6, 2013). The MOOC phenomenon: Who takes Massive Open Online Courses and why? Retrieved from <http://ssrn.com/abstract=2350964>
- Chuang, I., & Ho, A. D. (2016). HarvardX and MITx: Four years of open online courses -- Fall 2012–Summer 2016. Retrieved from SSRN: <https://ssrn.com/abstract=2889436> or <http://dx.doi.org/10.2139/ssrn.2889436>
- Clifford, J. (1990). Notes on (field)notes. In R. Sanjek (Ed.), *Fieldnotes: The makings of anthropology* (pp. 47–70). Ithaca, NY: Cornell University Press.
- Cobb, P., Confrey, J., di Sessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, 32(1), 9–13.
- Cobb, T. (n.d). Compleat Lexical Tutor. Retrieved from <http://www.lex tutor.ca/>

- Cobb, T. (1999). Applying constructivism: A test for the learner-as-scientist. *Educational Technology Research & Development*, 47(3), 15–31.
- Cobb, T. (2006). Internet and literacy in the developing world: Delivering the teacher with the text. *Educational Technology Research & Development*, 54(6), 627–645.
- Cobb, T. & Boulton, A. (2015). Classroom applications of corpus analysis. In D. Biber & R. Reppen (Eds.), *Cambridge handbook of English corpus linguistics* (pp. 478–497). Cambridge, UK: Cambridge University Press.
- Collins, A. (1992). Toward a Design Science of Education. In E. Scanlon & T. O'Shea (Eds.), *New directions in educational technology* (pp. 15–22). New York: Springer-Verlag.
- Colpaert, J. (2004). Transdisciplinarity. *Computer Assisted Language Learning*, 17(5), 459–472
- Colpaert, J. (2015). Reflections on present and future: towards an ontological approach to LMOOCs. In E. Martín-Monje & E. Bárcena (Eds.) *Language MOOCs. Providing Learning, Transcending Boundaries* (pp. 161–172). De Gruyter Open: Berlin. Retrieved from <https://www.degruyter.com/downloadpdf/books/9783110422504/9783110422504.10/9783110422504.10.pdf>
- Colpaert, J. (2016). Big content in an educational engineering approach. *Journal of Technology and Chinese Language Teaching*, 7(1), 1–14. Retrieved from <http://tclt.us/journal/2016v7n1/colpaert.pdf>
- Colpaert, J. (2018): Transdisciplinarity revisited. *Computer Assisted Language Learning*, DOI: 10.1080/09588221.2018.1437111
- COConnectedREpositories (CORE): Aggregating the world's open access papers. (n.d.). Retrieved from <https://core.ac.uk/>
- Cost of Knowledge. (n.d.). Elsevier statement. Retrieved from <https://gowers.files.wordpress.com/2012/02/elsevierstatementfinal.pdf>
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213–238.
- Coxhead, A., & Byrd, P. (2007). Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Second Language Writing*, 16, 129–147.
- Crabtree, A., Rouncefield, M., & Tolmie, P. (2012). *Doing design ethnography*. Springer: London.

- Creative Commons. (2015). *State of the commons*. Retrieved from <https://stateof.creativecommons.org/2015/cc-sotc-2015-xx12.html>
- Daniel, J. (2013). Rankings and online learning: a disruptive combination for higher education? In P.T.M. Marope, P.J. Wells & E. Hazelkorn (Eds), *Rankings and accountability in higher education: Uses and misuses* (pp. 95–112). Paris: UNESCO
- Daskalovska, N. (2015). Corpus-based versus traditional learning of collocations. *Computer Assisted Language Learning*, 28(2), 130–144. doi: 10.1080/09588221.2013.803982
- De Groot, F. O. (2017). Tracing the potential of out-of-class digitally mediated language learning practice back to the classroom: A nexus of practice perspective. In M. Carrier, R. M. Damerow, and K. M. Bailey (Eds.), *Digital language learning and teaching* (pp 25–37). New York, NY: Routledge.
- Dede, C. (2004). If design-based research is the answer, what is the question? A commentary on Collins, Joseph, and Bielaczyc; diSessa and Cobb; and Fishman, Marx, Blumenthal, Krajcik, and Soloway in the JLS special issue on design-based research. *Journal of the Learning Sciences*, 13(1), 105–114. doi: 10.1207/s15327809jls1301\_5
- Design-Based Research Collective (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher*, 32(1), 5–8.
- Digital commons (economics). (2016, March 13). In *Wikipedia, the free encyclopaedia*. Retrieved from [https://en.wikipedia.org/w/index.php?title=Digital\\_commons\\_\(economics\)&oldid=709892273](https://en.wikipedia.org/w/index.php?title=Digital_commons_(economics)&oldid=709892273)
- Ding, A., & Bruce, I. (2017). *The English for academic purposes practitioner: operating on the edge of academia*. Palgrave Macmillan.
- Downes, S. (2004). Nine rules for good technology. In *The learning marketplace: Meaning, metadata, and content syndication in the learning object economy* (pp. 11–15). Retrieved from <http://www.downes.ca/files/book3.htm>
- Downes, S. (2007). Models for sustainable open educational resources. *Interdisciplinary Journal of Knowledge and Learning Objects*, 3. Retrieved from <http://ijklo.org/Volume3/IJKLOv3p029-044Downes.pdf>
- Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1), 99–117.

- Engeström, Y. (2009). Epilogue: the future of activity theory. In A. Sannino, H. Daniels & K. D. Gutierrez (Eds.), *Learning and expanding with activity theory*. Cambridge: Cambridge University Press.
- Eaglestone, B., Ford, N., Brown, G.J., & Moore, A. (2007). Information systems and creativity: an empirical study. *Journal of Documentation*, 63, 443–464.
- Edwards, L., Hilty, R., Hugenholtz, B., Klimpel, P., Kreutzer, T., Lynn, J., Manara, C., Perez, N., & Vivant, M. (2012). *Intellectual property and innovation: A framework for 21<sup>st</sup> century growth and jobs*. In I. Hargreaves & P. Hofheinz (Eds.), Lisbon Council Policy Brief, Vol. VI, No. 2. ISSN 2031 0943 Retrieved from <http://www.lisboncouncil.net/publication/publication/84-intellectual-property-and-innovation-a-framework-for-21st-century-growth-and-jobs-.html>
- Elliott, R., Fischer, C. T., & Rennie, D. L. (1999). Evolving guidelines for publication of qualitative research studies in psychology and related fields. *Evolving Guidelines for Publication of Qualitative Research Studies in Psychology and Related Fields*, 38, 215–229.
- Emery, M., & Devane, T. (2007) Participative design workshop. In P. Holman, T. Devane, S. Cady (Eds.), *The change handbook: The definitive resource on today's best methods for engaging whole systems* (pp. 419– 435). San Francisco: Berrett-Koehler.
- Epstein, K. (2012). Academic Spring sees widening boycott of Elsevier. *BMJ, News*, 344.
- Ericsson, K., & Simon, H. (1987). Verbal reports on thinking. In C. Faerch & G. Kasper (Eds.), *Introspection in Second Language Research* (pp. 24–54). Clevedon, Avon: Multilingual Matters.
- ETIC (1975). *English for academic study: Problems and perspectives*. London: British Council.
- Farrow, R., Pitt, R., de los Arcos, B., Perryman, L.-A., Weller, M., & McAndrew, P. (2015). Impact of OER use on teaching and learning: data from OER Research Hub (2013–2014). *British Journal of Educational Technology*, 46(5), 972–976.
- Finch Group (2012). Accessibility, sustainability, excellence: how to expand access to research publications. *Report of the Working Group on Expanding Access to Published Research Findings*. Retrieved from <http://www.researchinfonet.org/publish/finch/>
- Fisher, W.W. (2014a). CopyrightX: Lecture 12.3, Remedies: Criminal Penalties. Retrieved from [https://www.youtube.com/watch?v=x6W8gk9fOFk&feature=youtube\\_gdata\\_player](https://www.youtube.com/watch?v=x6W8gk9fOFk&feature=youtube_gdata_player)

- Fisher, W.W. (2014b). HLS1X: CopyrightX course report. HarvardX working paper series No. 5. Available at SSRN: <http://ssrn.com/abstract=2382332> or <http://dx.doi.org/10.2139/ssrn.2382332>
- Fitzgerald, A. (2013a). *Openness in English for Academic Purposes*. Open Educational Resources case study with Durham University: Pedagogical development from OER practice. Commissioned by the Higher Education Academy (HEA) and the Joint Information Systems Committee (JISC), United Kingdom. Retrieved from <https://www.heacademy.ac.uk/knowledge-hub/openness-english-academic-purposes>
- Fitzgerald, A. (2013b). *TOETOE International: FLAX Weaving with Oxford Open Educational Resources*. Open Educational Resources International case study with the University of Oxford. Commissioned by the Higher Education Academy (HEA) and the Joint Information Systems Committee (JISC), United Kingdom. Retrieved from <https://www.heacademy.ac.uk/knowledge-hub/toetoe-international-flax-weaving-oxford-open-education-resources>
- Fitzgerald, A. (2013c, February 14). Love is a stranger in an open car to tempt you in and drive you far away ... toward OEP [Blog post] *TOETOE Technology for Open English Toying with Open E-resources* ('tɔɪtɔɪ). Retrieved from <http://www.alannahfitzgerald.org/love-is-a-stranger-in-an-open-car-who-tempts-you-in-and-drives-you-far-away/>
- Fitzgerald, A., Wu, S., & Barge, M. (2014). Investigating an open methodology for designing domain-specific language collections. In S. Jager, L. Bradley, E. J. Meima, & S. Thouësny (Eds.), *CALL Design: Principles and Practice; Proceedings of the 2014 EUROCALL Conference, Groningen, The Netherlands* (pp. 88–95). Dublin: Research-publishing.net.
- Fitzgerald, A., Wu, S., & Marín, M.J. (2015). FLAX: Flexible and open corpus-based language collections development. In K. Borthwick, E. Corradini, A. Dickens (Eds.), *10 years of the LLAS elearning symposium: Case studies in good practice* (pp. 215–227). Dublin: Research-publishing.net.
- Fitzgerald, A. (2015, September 14). Rebalancing English language education: Access, materials writing and copyright. [Blog post] *ELTjam*. Retrieved from <http://eltjam.com/rebalancing-english-language-education-access-materials-writing-and-copyright/>

- Fitzgerald, A., Marín, M.J., Wu, S., & Witten, I. H. (2017). Evaluating the efficacy of the digital commons for scaling data-driven learning. In M. Carrier, R. Damerow, K. Bailey (Eds.), *Digital language learning and teaching: Research, theory and practice*. Global Research on Teaching and Learning English Series (pp. 38–51). New York: Routledge & TIRF.
- Fitzgerald, A. (2017). MOOC linguistic support. Data files available at <https://osf.io/juakn/>
- Fitzgerald, A., Wu, S., König, J., Witten, I.H., & Shaw, S. (Submitted). Designing and evaluating an automated open data-driven language learning support system for MOOCs.
- Fitzgerald, A., Wu, S. & Witten, I.H. (Submitted). Reflections on remixing open access content for data-driven language learning systems design in higher education.
- FLAX. (n.d.). The “Flexible Language Acquisition Project”. Retrieved from <http://flax.nzdl.org/>
- Flowerdew, J., & Peacock, M. (2001). Issues in EAP: A preliminary perspective. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 8–24). Cambridge: Cambridge University Press.
- Flowerdew L. (2009). Applying corpus linguistics to pedagogy: A critical evaluation. *International Journal of Corpus Linguistics*, 14(3), 393–417.
- Ferguson, G., Pérez-Llantada, C., & Plo, R. (2011). English as an international language of scientific publication: A study of attitudes. *World Englishes*, 30(1), 41–59.
- Franken, M. (2014). The nature and scope of student search strategies in using a web derived corpus for writing. *The Language Learning Journal*, 42(1), 85–102.
- Frankenberg-Garcia, A. (2005). A peek into what today’s language learners as researchers actually do. *International Journal of Lexicography*, 18, 335–55.
- Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors? *System*, 32(3), 301–319.
- Gatt, C., & Ingold, T. (2013). From description to correspondence: anthropology in real time. In W. Gunn, T. Otto, R.C. Smith (Eds), *Design anthropology: Theory and practice* (pp. 139–158). London: Bloomsbury Academic.
- Geertz, C. (1973). *The interpretation of cultures: selected essays*. New York: Basic Books.
- Geiller, I. (2014). How EFL students can use Google to correct their “untreatable” written errors. *The Eurocall Review*, 22(2), 26–45.



- Geluso, J., & Yamaguchi, A. (2014). Discovering formulaic language through data-driven learning: Student attitudes and efficacy. *ReCALL*, 26(2), 225–242. doi: 10.1017/S0958344014000044
- Genzuk, M. (2003). A synthesis of ethnographic research. *Occasional Papers Series*. Center for Multilingual, Multicultural Research. Los Angeles: Rossier School of Education, University of Southern California.
- Gibson, W. (1999, November 30). The science in science fiction. *Talk of the Nation*, NPR. Retrieved from <http://www.npr.org/templates/story/story.php?storyId=1067220>
- Gillani, N., & Eynon, R. (2014). Communication patterns in massively open online courses. *Internet and Higher Education*, 23, 18–26.
- Gillett, A. (2018, April 13). Is EAP ESP? [Blog post] *Uefap*. Retrieved from <http://www.uefap.net/blog/?p=933>
- Godwin-Jones, R. (2012). Challenging hegemonies in online learning. *Language Learning & Technology*, 16(2), 4–13. Retrieved from <http://llt.msu.edu/issues/june2012/emerging.pdf>
- Gonzalez, G.H., Tahsin, T., Goodale, B.C., Greene, A.C., & Greene, C.S. (2016). Recent advances and emerging applications in text and data mining for biomedical discovery. *Brief Bioinform*, 17(1), 33–42.
- Graddol, D. (2006). English Next – why English as a global language may mean the end of ‘English as a Foreign Language’. The British Council: The English Company.
- Graham, C. R. (2006). Blended learning systems: Definition, current trends, and future directions. In C.J. Bonk & C. R. Graham (Eds.), *The handbook of blended learning: Global perspectives, local designs* (pp.3–21). San Francisco, CA: Pfeiffer
- Graham, C. R. (2012). Emerging practice and research in blended learning. In M.G. Moore (Ed.), *Handbook of distance education* (pp. 333–350). New York, NY: Routledge.
- Green, C. (2015, June 2). edX makes it easy for authors to share under Creative Commons [Blog post]. Retrieved from <https://creativecommons.org/2015/06/02/edx-makes-it-easy-for-authors-to-share-under-creative-commons/>
- Groom, J., & Lamb, B. (2014). Reclaiming innovation. *EDUCAUSE Review*, 49(3). Retrieved from <https://www.educause.edu/visuals/shared/er/extras/2014/ReclaimingInnovation/default.html>

- Groot, P.J.M. (2000). Computer assisted second language vocabulary acquisition. *Language Learning & Technology*, 4, 60–81.
- Gunn, W., & Donovan, J. (2013). *Design and anthropology*. Burlington, Vermont: Ashgate.
- Gunn, W., Otto, T., & Smith, R.C. (2013). *Design anthropology: Theory and practice*. London: Bloomsbury Academic.
- Guo, P., & Reinecke, K. (2014). Demographic differences in how students navigate through MOOCs. *Proceedings of the first ACM conference on Learning @ Scale* (pp. 21–30). New York: ACM.
- Hackin, T. (2001). Abstracting from abstracts. In M. Hewings (Ed.), *Academic writing in context: Implications and applications* (pp. 93–103). Birmingham University Press, Birmingham.
- Hadley, G. (2015). *English for Academic Purposes in neoliberal universities: A critical grounded theory* (Vol. 22). Cham: Springer International Publishing. Retrieved from <http://link.springer.com/10.1007/978-3-319-10449-2>
- 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.
- Hake, R. (1998). Interactive engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(1), 64–74.
- Hakkarainen, P. (2009). Designing and implementing a PBL course on educational digital video production: Lessons learned from a design- based research. *Educational Technology, Research and Development*, 57(2), 211–228. doi: 10.1007/s11423- 007- 9039- 4
- Hargreaves, I. (2011). *Digital opportunity – A review of intellectual property and growth*. London: HM Government. Retrieved from [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/32563/ipreview-finalreport.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/32563/ipreview-finalreport.pdf)
- Hearst, M. (1999). Untangling text data mining. In R. Dale & K. Church (Eds.), *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (pp. 3–10). Morristown, New Jersey: ACL. doi:10.3115/1034678.1034679
- Hearst, M. A. (2009). *Search user interfaces*. Cambridge, UK: Cambridge University Press.

- Herrington, J., McKenney, S., Reeves, C. T., & Oliver, R. (2007) Design-based research and doctoral students: guidelines for preparing a dissertation proposal. Retrieved from <http://ro.ecu.edu.au/cgi/viewcontent.cgi?article=2611&context=ecuworks>
- Hill, J. (1999). Collocational competence. *ETP (English Teaching Professional)*, 11, 1–6.
- Hilton III, J. L., & Wiley, D. (2011). Open access textbooks and financial sustainability: A case study on flat world knowledge. *The International Review of Research in Open and Distributed Learning*, 12(5), 18–26.
- Hilton III, J. (2016). Open educational resources and college textbook choices: a review of research on efficacy and perceptions. *Educational Technology Research & Development*, 64(4), 573–590. <https://doi.org/10.1007/s11423-016-9434-9>
- Ho, A. D., Chuang, I., Reich, J., Coleman, C. A., Whitehill, J., Northcutt, C. G., William, J.J., Hansen, J.D., Lopez, G., & Petersen, R. (2015). HarvardX and MITx: Two years of open online courses Fall 2012-Summer 2014 (SSRN Scholarly Paper No. ID 2586847). Rochester, NY: Social Science Research Network. Retrieved from SSRN: <http://papers.ssrn.com/abstract=2586847>
- Howatt, A. P. R. (2004). *A history of English language teaching* (2nd ed.) Oxford: Oxford University Press.
- Hughes, J.A., Randall, D., & Shapiro, D. (1992). Faltering from ethnography to design. In *Proceedings of ACM 1992 Conference on Computer-Supported Cooperative Work: Sharing Perspectives* (pp. 115–123). New York: ACM Press.
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
- Hyland, K. (2000). *Disciplinary discourses: Social interactions in academic writing*. Harlow: Longman.
- Hyland, K. (2002). Specificity revisited: How far should we go now? *English for Specific Purposes*, 21, 385–395.
- Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. London: Continuum.
- Hyland, K. (2006). *English for Academic Purposes: An advanced resource book*. London: Routledge.
- Hyland, K. (2008a). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, 18(1), 41–62.

- Hyland, K. (2008b). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1), 4–21.
- Hyland, K. (2018). Sympathy for the devil: A defense of EAP. *Language Teaching*, 51(3), 383–399.
- Hyland, K. (2019). Participation in publishing: the demoralising discourse of disadvantage. In P. Habibie & K. Hyland (Eds.), *Novice writers and scholarly publication: Authors, mentors, gatekeepers* (pp. 1–33). Cham, Switzerland: Palgrave Macmillan.
- IT Services, University of Oxford, 13 Banbury Road (n.d.). [Oxford Text Archive] [BAWE Terms and Conditions Text]. Retrieved from <http://ota.ox.ac.uk/scripts/download.php?otaid=2539>
- Johns, T. (1991). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Johns & P. King (Eds.), *Classroom concordancing*. *English Language Research Journal*, 4, 27–45.
- Johns, T. (1993). Data-driven learning: An update. *TELL&CALL*, 2, 4–10.
- Johns, T. (1997). Contexts: The background, development and trialing of a concordance-based CALL program. In A. Wichmann, S. Figelston, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 100–115). London, UK: Longman.
- Johns, T. (2002). Data-driven learning: the perpetual challenge. In B. Kettemann & G. Marko (Eds.), *Proceedings of Teaching and Learning by Doing Corpus Analysis: the Fourth International Conference on Teaching and Language Corpora*, Graz 19–24 July 2000 (pp.107–117). Amsterdam: Rodopi.
- Kalman, Y. (2014). A race to the bottom: MOOCs and higher education business models. *Open Learning: The Journal of Open, Distance and eLearning*, 29(1), 5–14.
- Kelty, C.M. (2008). *Two bits: The cultural significance of free software*. Durham, NC: Duke University Press.
- Kilbourn, K. (2013). Tools of movement and engagement: design anthropology's style of knowing. In W. Gunn, T. Otto, R.C. Smith (Eds.), *Design anthropology: Theory and practice* (pp. 68–82). London: Bloomsbury Academic.
- Kilgariff, A., & Grefenstette, G. (2003). Introduction to the social issue on the web as corpus. *Computational Linguistics*, 29(3), 333–347.

- Kincheloe, J. L. (2005). On to the next level: Continuing the conceptualization of the bricolage. *Qualitative Inquiry*, 11(3), 323–350.
- King, M. (2007). General principles of user-oriented evaluation. In L. Dybkjær, H. Hemsén & W. Minker (Eds). *Evaluation of text and speech systems* (pp. 125–161). Berlin/Heidelberg/New York: Springer.
- Kizilcec, R., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: Analysing learner subpopulations in Massive Open Online Courses. *Proceedings of the third international conference on learning analytics and knowledge* (pp. 170–179). New York, NY, USA: ACM.
- Knoth, P., & Zdrahal, Z. (2012). CORE: three access levels to underpin open access. *D-Lib Magazine*, 18(11/12), article number 4.
- Knox, J. (2013). The limitations of access alone: Moving towards open processes in education technology. *Open Praxis*, 5(1), 21–29.
- Koseoglu, S., & Bozkurt, A. (2018). An exploratory literature review on open educational practices, *Distance Education*, 39(4), 441–461.
- Kovanović, V., Joksimović, S., Gašević, D., Siemens, G., & Hatala, M. (2015). What public media reveals about MOOCs: A systematic analysis of news reports. *British Journal of Educational Technology*, 46, 510–527. doi:10.1111/bjet.12277
- Kuper, A., Lingard, L., & Levinson, W. (2008). Critically appraising qualitative research. *BMJ*, 337, a1035. <https://doi.org/10.1136/bmj.a1035>.
- Laakso, M., Welling, P., Bukvova H., Nyman, L., Björk, B-C., & Hedlund, T. (2011). The development of open access journal publishing from 1993 to 2009. *PLoS One*, 6(6), e20961. 10.1371/journal.pone.0020961
- LeCompte, M., & Schensul, J. (1999). *Analyzing and interpreting ethnographic data*. California: AltaMira Press.
- Lederman, D. (2017, April 28). Clay Christensen, doubling down. *Inside Higher Ed*. Retrieved from <https://www.insidehighered.com/digital-learning/article/2017/04/28/clay-christensen-sticks-predictions-massive-college-closures>
- Lepore, J. (2014, June 23). The Disruption Machine: What the gospel of innovation gets wrong. *The New Yorker*. Retrieved from <https://www.newyorker.com/magazine/2014/06/23/the-disruption-machine>

- Lessig, L. (2004). *Free Culture: How Big Media Uses Technology and the Law to Lock Down Culture and Control Creativity*. New York: Penguin.
- Lessig, L. (2008). *Remix: Making art and commerce thrive in the hybrid economy*. New York, NY: Penguin Press.
- Levin, B. (2011). Mobilising research knowledge in education. *London Review of Education*, 9(1), 15–26. doi:10.1080/14748460.2011.550431.
- Levy, M. (1997). *CALL: Context and conceptualisation*. Oxford: Oxford University Press.
- Li, L., Franken, M., & Wu, S. (2017). Bundle-driven metadiscourse analysis: Sentence initial bundles in Chinese and New Zealand postgraduates' thesis writing. In C. Hatipoglu, E. Akbas & Y. Bayyurt (Eds.), *Metadiscourse in written genres: Uncovering textual and interactional aspects of texts*. Amsterdam: Peter Lang Publishing.
- Liao, S., & Lei, L. (2017). What we talk about when we talk about corpus: A bibliometric analysis of corpus-related research in linguistics (2000-2015). *Glottometrics*, 38, 1–20.
- Lincoln, Y., & Guba, E. (1985). *Naturalistic inquiry*. London: Sage Publications Inc.
- Littlejohn, A., Hood, N., Milligan, C., & Mustain, P. (2016). Learning in MOOCs: Motivations and self-regulated learning in MOOCs. *Internet and Higher Education*, 29, 40–48.
- Littlejohn, A., & Hood, N. (2017). How educators build knowledge and expand their practice: The case of open education resources. *British Journal of Educational Technology*, 48(2), 499–510. doi: <https://doi.org/10.1111/bjet.12438>
- Liyanagunawardena, T., Adams, A., & Williams, S. (2013). MOOCs: A systematic study of the published literature 2008–2012. *The International Review of Research in Open and Distance Learning*, 14(3), 202–227.
- Lock, S. (1988). Structured abstracts. *BMJ: British Medical Journal*, 297, 156.
- Lockley, P. (2015, December 9). Re: Open Policies for MOOCs [Discussion-list]. JISCMail - OER-DISCUSS Archives. Retrieved from <https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind1512&L=OER-DISCUSS&F=&S=&P=16754>
- Lugmayr, A., Stockleben, B., Zou, Y., Anzenhofer, S., & Jalonen, M. (2014). Applying “design thinking” in the context of media management education. *Multimedia Tools and Applications*, 71, 119–157.
- Lyotard, J-F. (1984). *The Postmodern Condition: A Report on Knowledge*. Theory and history of literature, v. 10. Minneapolis: University of Minnesota Press.

- Malhotra, A., Younesi, E., Gurulingappa, H., & Hofmann-Apitius, M. (2013). 'HypothesisFinder:' a strategy for the detection of speculative statements in scientific text. *PLoS Computational Biololgy*, 9(7), e1003117.
- Marín, M. J., & Rea, C. (2014). Assessing four automatic term recognition methods: Are they domain-dependent? *English for Specific Purposes World*, 42(15), 1–27.
- Marín, M. J. (2014). Evaluation of five single-word term recognition methods on a legal corpus. *Corpora*, 9(1), 83–107.
- Marin, M.J., Ortis Llopaz, M., & Fitzgerald, A. (2017). A Data-driven learning experiment in the legal English classroom using the FLAX platform. *Miscelánea: A Journal of English and American Studies*, 55, 37–64.
- Martin, M. (1984). Advanced vocabulary teaching: The problem of synonyms. *Modern Language Journal*, 69, 130–137.
- McAlpine, J., & Myles, J. (2003). Capturing phraseology in an online dictionary for advanced users of English as a second language: a response to user needs. *System*, 31, 71–84.
- McEnery, T., & Wilson, A. (1997). Teaching and language corpora. *ReCALL*, 9(1), 5–14.
- McEnery, T., & Wilson, A. (2001). *Corpus linguistics*. Edinburgh: Edinburgh University Press.
- McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: an advanced resource book*. London: Routledge.
- McGill, L., Falconer, I., Dempster, J.A., Littlejohn, A., & Beetham, H. (2013). OER journeys: HEFCE OER Review final report. JISC. Retrieved from <https://oersynth.pbworks.com/w/page/60338879/HEFCE-OER-Review-Final-Report>
- McGill, L., Beetham, H., Falconer, I., & Littlejohn, A. (2010) JISC/HEA OER Programme: Pilot Phase Synthesis and Evaluation Report. Retrieved from <https://oersynth.pbworks.com/w/page/29688444/Pilot-Phase-Synthesis-and-Evaluation-Report>
- McKenney, S., & Reeves, T. (2012). *Conducting Educational Design Research*. London and New York: Routledge.
- McLuhan, M. (2005). Fordham University: First lecture (1967). In S. McLuhan and D. Staines (Eds.), *Understanding me: Lectures and interviews*. Cambridge: MIT Press

- McMillan Cottom, T. (2015). Intersectionality and critical engagement with the Internet. In U. Safiya, N. Tynes & B. Tynes (Eds.), *The intersectional Internet: Race, sex, class, and culture online*. Peter Lang Publishing.
- McSherry, C. (2015, October 16). Big win for fair use in Google books lawsuit. *Electronic Frontier Foundation*. Retrieved from <https://www.eff.org/deeplinks/2015/10/big-win-fair-use-google-books-lawsuit>
- Mellinkoff, D. (1963). *The language of the law*. Boston: Little, Brown & Co.
- Merlucci, A. (1996). *Challenging codes: Collective action in the information age*. Cambridge: Cambridge University Press.
- Meyers, C., & Jones, T. (1993). *Promoting active learning: Strategies for the college classroom*. Jossey-Bass: San Francisco.
- Mezynski, K. (1983). Issues concerning the acquisition of knowledge: Effects of vocabulary training on reading comprehension. *Review of Educational Research*, 53, 253–279.
- Milligan, C., Littlejohn, A., & Margaryon, A. (2013). Patterns of engagement in connectivist MOOCs. *Merlot Journal of Online Learning and Teaching*, 9(2).
- Milne, D., & Witten, I. H. (2013). An open-source toolkit for mining Wikipedia. *Artificial Intelligence*, 194, 222–239.
- Moxley, J. (2018, March). Letter from the founder. *Writing Commons*. Retrieved from <https://writingcommons.org/component/acymailing/listid-225-master-wc-list/mailid-105-writing-commons-march-2018>
- Natarajan, J., Berrar, D., Dubitzky, W., Hack, C., Zhang, Y., DeSesa, C., Brocklyn, J.R.V., & Bremer, E.G. (2006). Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line. *BMC Bioinformatics*, 7(1), 373.
- Naughton, John. (2011). *The elusive technological future*. Keynote address given at the Association for Learning Technology Annual Conference. Leeds, United Kingdom. Retrieved from <https://www.youtube.com/watch?v=yUXh-GPa5dI>
- Naur, P. (1983). Program development studies based on diaries. In T. Green, S. Payne, G. Veer (Eds.), *Psychology of computer use* (pp. 159–170). London: Academic Press.
- Nesi, H, Gardner, S., Thompson, P., & Wickens, P. (2007) *The British Academic Written English (BAWE) corpus*, developed at the Universities of Warwick, Reading and Oxford Brookes



- under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800)
- Nesi, H., & Gardner S. (2012). *Genres across the disciplines: student writing in Higher Education*. Cambridge: Cambridge University Press.
- Nuijten, M.B., Hartgerink, C.H.J., van Assen, M.A.L.M, Epskamp, S., & Wicherts, J.M. (2015). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 48(4), 1205–1226. ISSN 1554-351X, 1554-3528.
- O'Brien, B. C., Harris, I. B., Beckman, T. J., Reed, D. A., & Cook, D. A. (2014). Standards for reporting qualitative research: A synthesis of recommendations. *Academic Medicine*, 89(9), 1245e1251. <https://doi.org/10.1097/ACM.0000000000000388>.
- O'Keeffe, A., McCarthy, M., & Carter R. (2007). *From corpus to classroom: language use and language teaching*. Cambridge: Cambridge University Press.
- Okerson, A. (1991). With feathers: Effects of copyright and ownership on scholarly publishing. *College & Research Libraries*, 52(5), 425–38.
- Onwuegbuzie, A. J., & Leech, N. L. (2005). On becoming a pragmatic researcher: The importance of combining quantitative and qualitative research methodologies. *International Journal of Social Research and Methodology*, 8, 375–387.
- O'Reilly, T. (2005, September 30). What is Web 2.0? End of the software release cycle. *O'Reilly Media, Inc*. Retrieved from <https://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html?page=4>
- Orlikowski, W.J. (1991). Integrated information environment or matrix of control? The contradictory implications of information technology. *Accounting, Management and Information Technologies*, 1, 9–42.
- O'Sullivan, I., & Chambers, A. (2006). Learners' writing skills in French: Corpus consultation and learner evaluation. *Journal of Second Language Writing*, 15(1), 49–68.
- Otto, T., & Smith, R.C. (2013). Design anthropology: a distinct style of knowing. In W. Gunn, T. Otto, R.C. Smith (Eds.), *Design anthropology: Theory and practice* (pp. 1–29). London: Bloomsbury Academic.

- Pappano, L. (2012, November 2). The year of the MOOC. *The New York Times*. Retrieved from <http://www.nytimes.com/2012/11/04/education/edlife/massive-open-online-courses-are-multiplying-at-a-rapid-pace.html>
- Park, K. (2012). Learner-corpus interaction: A locus of microgenesis in corpus-assisted L2 writing. *Applied Linguistics*, 33(4), 361–385.
- Parr, C. (2013, May 10). Not staying the course, *Times Higher Education*. Retrieved from <https://www.insidehighered.com/news/2013/05/10/new-study-low-mooc-completion-rates>.
- Peachey, N. (2005). Concordancers in ELT. *British Council Teaching English*. Retrieved from <http://www.teachingenglish.org.uk/think/articles/concordancers-elt>
- Peter, S., & Deimann, M. (2013). On the role of openness in education: A historical reconstruction. *Open Praxis*, 5(1), 7–14. doi: <http://dx.doi.org/10.5944/openpraxis.5.1.23>
- Pérez-Paredes, P., Ordoñana Guillamón, C., & Aguado Jiménez, P. (2018): Language teachers' perceptions on the use of OER language processing technologies in MALL. *Computer Assisted Language Learning*, 31 (5-6), 522-545. DOI: 10.1080/09588221.2017.1418754
- Rapoport, R. (1970). Three dilemmas of action research. *Human Relations*, 23, 499–513.
- Raymond, E. (1997, May 27). Release early, release often. *The cathedral and the bazaar: Musings on Linux and open source by an accidental revolutionary*. Retrieved from <http://www.catb.org/~esr/writings/cathedral-bazaar/cathedral-bazaar/ar01s04.html>
- Reason, P., & Bradbury, H. (2007). *Handbook of Action Research*, 2nd Edition. London: Sage.
- Renkle, A., Atkinson, R., Maier, U. & Staley, R. (2002). From example study to problem solving: Smooth transitions help learning. *Journal of Experimental Education*, 70(4), 293–315.
- Reppen, R. (2010). *Using Corpora in the Language Classroom*. Cambridge: Cambridge University Press.
- Rittel, H.W.J., & Webber, M.M. (1973). Dilemmas in a general theory of planning. *Policy Sciences*, 4(2), 155–169. <https://doi.org/10.1007/BF01405730>
- Robinson, P. (1995). Attention, memory, and the “noticing” hypothesis. *Language Learning*, 45(2), 283–331.
- Rodgers, O., Chambers, A., & Le Baron-Earle, F. (2011). Corpora in the LSP classroom: A learner-centred corpus of French for biotechnologists. *International Journal of Corpus Linguistics*, 16(3), 391–411.

- Römer, U. (2010). Using general and specialized corpora in English language teaching: Past, present and future. In M.C. Campoy-Cubillo, B. Belles-Fortuño, & L. Gea-Valor (Eds.), *Corpus-based Approaches to English Language Teaching* (pp. 18–38). London: Continuum.
- Römer, U. (2011). Corpus research applications in second language teaching. *Annual Review of Applied Linguistics* 31, 205–225.
- Rott, S. 2004. A comparison of output interventions and un-enhanced reading conditions on vocabulary acquisition and text comprehension. *The Canadian Modern Language Review*, 61, 169–202.
- Salvador, T., Bell, G., & Anderson, K. (1999). Design ethnography. *Design Management Journal (Former Series)*, 10, 35–41.
- Sanjek, R. (2004). Going public: responsibilities and strategies in the aftermath of ethnography. *Human Organization*, 63, 444–456.
- Santiago-Delefosse, M., Gavin, A., Bruchez, C., Roux, P., & Stephen, S. L. (2016). Quality of qualitative research in the health sciences: Analysis of the common criteria present in 58 assessment guidelines by expert users. *Social Science & Medicine*, 148, 142–151. <https://doi.org/10.1016/j.socscimed.2015.11.007>.
- Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge: Cambridge University Press.
- Schmitt, N., & Schmitt, D. (1995). Vocabulary notebooks: theoretical underpinnings and practical suggestions. *ELT Journal*, 49(2), 133–143.
- Schmitt, N. (1997). Vocabulary learning strategies. In N. Schmitt, and M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp.199–227). Cambridge: Cambridge University Press.
- Scott, M. (2008). *WordSmith Tools version 5*. Liverpool, UK: Lexical Analysis Software.
- Seimens, G. (2005). Connectivism: A learning theory for the digital age. *International Journal of Instructional Technology & Distance Learning*, 2(1).
- Seimens, G. (2011, October 13). The race to platform education [Blog post]. *elearnspace*. Retrieved from <http://www.elearnspace.org/blog/2011/10/13/the-race-to-platform-education/>

- Seimens, G. (2012, July 25). MOOCs are really a platform [Blog post]. *elearnspace*. Retrieved from <http://www.elearnspace.org/blog/2012/07/25/moocs-are-really-a-platform/>
- Selwyn, N. (2014). Data entry: towards the critical study of digital data and education. *Learning, Media and Technology*, 40(1), 64–82. <https://doi.org/10.1080/17439884.2014.921628>
- Selwyn, N. (2015). Minding our language: why education and technology is full of bullshit ... and what might be done about it. *Learning, Media and Technology*, 41(3), 437–443. <http://doi.org/10.1080/17439884.2015.1012523>
- Shah, D. (2016, December 29). Monetization over massiveness: Breaking down MOOCs by the numbers in 2016. *EdSurge*. Retrieved from <https://www.edsurge.com/news/2016-12-29-monetization-over-massiveness-breaking-down-moocs-by-the-numbers-in-2016>
- Shah, D. (2018a, December 11). By the numbers: MOOCs in 2018. Retrieved from <https://www.class-central.com/report/mooc-stats-2018/>
- Shah, D. (2018b, January 22). A product at every price: A review of MOOC stats and trends in 2017. *Class Central*. Retrieved from <https://www.class-central.com/report/moocs-stats-and-trends-2017/>
- Shah, D. (2018c, June 3). The second wave of MOOC hype is here and it's online degrees. Retrieved from <https://www.class-central.com/report/second-wave-of-mooc-hype/>
- Shah D. (2018d, January 17). MOOC trends in 2017: Content paywalls. *Class Central*. Retrieved from <https://www.class-central.com/report/mooc-trends-content-paywalls/>
- Shapiro, G. (2006, October 26). A rebellion erupts over journals of academia. New York The Sun. Retrieved from <https://www.nysun.com/arts/rebellion-erupts-over-journals-of-academia/42317/>
- Shapiro, H.B., Lee, C.H., Wyman Roth, N.E., Li, K., Çetinkaya-Rundel, M., & Canelas, D.A. (2017). Understanding the massive open online course (MOOC) student experience: An examination of attitudes, motivations, and barriers. *Computers & Education*, 110, 35–50.
- Shaw, G. B. (1913). *Getting married: A disquisitory play*. London: Constable & Company Ltd.
- Shei, C. C. (2008). Discovering the hidden treasure on the Internet: using Google to uncover the veil of phraseology. *Computer Assisted Language Learning*, 21(1), 67–85.
- Shneiderman, B. & Plaisant, C. (2004). *Designing the user interface: strategies for effective human–computer interaction*. Boston, MA: Addison-Wesley Longman.
- Simon, H.A. (1996). *The sciences of the artificial*. Cambridge, Mass: MIT Press.

- Sinclair, J. McH. (2004a). *How to use corpora in language teaching*. Amsterdam: John Benjamins.
- Sinclair, J. McH. (2004b). *Trust the text: Language, corpus and discourse*. London: Routledge.
- Smith, A. E. (2000a). Machine learning of well-defined thesaurus concepts. In A.-H. Tan & P. S. Yu (Eds.), *Proceedings of the International Workshop on Text and Web Mining (PRICAI 2000)* (pp. 72–79). Melbourne: *PRICAI*.
- Smith, A. E. (2000b). Machine mapping of document collections: The Leximancer system. In *Proceedings of the Fifth Australasian Document Computing Symposium*. Sunshine Coast, Australia: DSTC.
- Smith, A. E. (2003). Automatic extraction of semantic networks from text using Leximancer. In *HLT-NAACL 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: Companion volume* (pp. Demo23- Demo24). Edmonton: ACL.
- Smith, A. E., & Humphreys, M. S. (2006). Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping. *Behavior Research Methods*, 38(2), 262–279.
- Sosoni, V. (2017). Casting some light on experts' experience with translation crowdsourcing. *The Journal of Specialised Translation*, 28, 362–384.
- SPARC. (2018, March). Congress funds \$5 million open textbook grant program in 2018 spending bill. Retrieved from <https://sparcopen.org/news/2018/open-textbooks-fy18/>
- Stahl, S.A., & Fairbanks, M.M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research*, 56, 72–110.
- Stallman, R.M. (2002). *Free software, free society: Selected essays of Richard M. Stallman*. J. Gay (Ed.). Boston: The Free Software Foundation.
- Stich, A.E., & Reeves, T.D. (2017). Massive open online courses and underserved students in the United States. *Internet and Higher Education*, 32, 58–71.
- Stevens, P. (1988). ESP after twenty years: a reappraisal. In M. Tickoo (Ed.), *ESP: state of the art* (pp. 1–13). Singapore: SEAMEO Regional Language Centre.
- Stubbs, M. (1996). *Text and corpus analysis*. Oxford: Basil Blackwell.
- Stubbs, M., & Barth, I. (2003) Using recurrent phrases as text-type discriminators. A quantitative method and some findings. *Functions of Language*, 10(1), 61–104.

- Swales, J. M. (1985). *Episodes in ESP*. Oxford: Pergamon Press.
- Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: 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: University of Michigan Press.
- Swan, A. (2012). *Policy guidelines for the development and promotion of open access*. Paris: UNESCO. Retrieved from <http://unesdoc.unesco.org/images/0021/002158/215863e.pdf>
- Swartz, A. (2008). *Guerilla open access manifesto*. Eremo, Italy: The Internet Archive. Retrieved from [https://archive.org/stream/GuerillaOpenAccessManifesto/Goamjuly2008\\_djvu.txt](https://archive.org/stream/GuerillaOpenAccessManifesto/Goamjuly2008_djvu.txt)
- Tennant, J.P., Waldner, F., Jacques, D.C., Masuzzo, P., Collister, L.B., & Hartgerink, C.H.J. (2016). The academic, economic and societal impacts of Open Access: an evidence-based review [version 2; referees: 4 approved, 1 approved with reservations]. *F1000Research*, 5:632. doi: 10.12688/f1000research.8460.2
- Thomas, J. (2017). *Discovering English with Sketch Engine* (2<sup>nd</sup> Edition). Versatile.
- Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins.
- Tomlinson, B., & Masuhara, H. (2010). Published research on materials development for language learning. In B. Tomlinson & H. Masuhara (Eds.), *Research for materials development in language learning*. London: Continuum.
- Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care*, 19(6), 349–357. <https://doi.org/http://dx.doi.org/10.1093/intqhc/mzm042>.
- Tribble, C. (2015). Teaching and language corpora: Perspectives from a personal journey. In A. Lenko-Szymanska, & A. Boulton, A. (Eds.) *Multiple Affordances of Language Corpora for Data-driven Learning* (pp. 37–64). Amsterdam: John Benjamins,
- Turner, J. (2004). Language as academic purpose. *Journal of English for Academic Purposes*, 3, 95–109.
- UNESCO. (2008). Education for All by 2015. Will we make it? *Education for All Global Education Monitoring Report 2008*. United Nations Education Scientific Cultural Organisation. Retrieved from [www.efareport.unesco.org](http://www.efareport.unesco.org)

- UNESCO. (2015). Education for All 2000-2015: Achievements and challenges. *Education for All Global Education Monitoring Report 2015*. United Nations Education Scientific Cultural Organisation. Retrieved from <http://unesdoc.unesco.org/images/0023/002322/232205e.pdf>
- UNESCO. (2017/2018). Accountability in education: Meeting our commitments. *Education for All Global Education Monitoring Report 2017/2018*. United Nations Education Scientific Cultural Organisation. Retrieved from <http://unesdoc.unesco.org/images/0025/002593/259338e.pdf>
- University of California Academic Senate. (2019, February 28). University of California Academic Council statement on the university's negotiations with Elsevier Publishing. Retrieved from [https://senate.universityofcalifornia.edu/\\_files/reports/academic-council-statement-elsevier-feb28.pdf](https://senate.universityofcalifornia.edu/_files/reports/academic-council-statement-elsevier-feb28.pdf)
- University of California Office of the President. (2019, February 28). UC terminates subscriptions with world's largest scientific publisher in push for open access to publicly funded research. Retrieved from <https://www.universityofcalifornia.edu/press-room/uc-terminates-subscriptions-worlds-largest-scientific-publisher-push-open-access-publicly>
- Uvalić-Trumbić, S., & Daniel, J. (2011). Let a thousand flowers bloom! *UNESCO Global Forum on Rankings and Accountability in Higher Education: Uses and Misuses*. Retrieved from <http://www.col.org/resources/speeches/2011presentation/Pages/2011-05-16.aspx>
- Varley, S. (2009) I'll just look that up in the concordancer: Integrating corpus consultation into the language learning environment. *Computer Assisted Language Learning*, 22(2), 133–152.
- Vollmer, T. (2012, November 1). Keeping MOOCs open [Blog post]. Retrieved from <https://creativecommons.org/2012/11/01/keeping-moocs-open/>
- Vyatkina, N. (2016). Data-driven learning of collocations: Learning performance, proficiency, and perceptions. *Language Learning & Technology*, 20(3), 159–179.
- Wakkary, R. (2005). Framing complexity, design and experience: a reflective analysis. *Digital Creativity*, 16, 65–78.
- Walker, D. (2006). Towards productive design studies. In J. van den Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *Educational design research* (pp. 9–18). London: Routledge.

- Wang, T. (2013, May 13). Big data needs thick data. Retrieved from <http://ethnographymatters.net/blog/2013/05/13/big-data-needs-thick-data/>
- Watters, A. (2013, November 7). The Education Apocalypse. 2013 Open Education Conference Keynote Address. In *Hack Education* #opened13 [Blog post]. Retrieved from <http://hackededucation.com/2013/11/07/the-education-apocalypse>
- Watters, A. (2014, November 16). From “open” to justice. 2014 OpenCon Keynote Address. In *Hack Education* #OpenCon14 [Blog post]. Retrieved from <http://hackededucation.com/2014/11/16/from-open-to-justice>
- Watters, A. (2015, February 19). The history of the future of education [Blog post]. In *Hack Education*. Retrieved from <http://hackededucation.com/2015/02/19/the-history-of-the-future-of-education>
- Watters, A. (2016). *The curse of the monsters of educational technology*. Tech Gypsies: Amazon Digital Services LLC.
- Weber, R.P. (1990). *Basic Content Analysis*. (Second edition). Newbury Park, California: Sage Publications.
- Weick, K. (1995). Organizational redesign as improvisation. In G.P. Huber & W.H. Glick (Eds.), *Organizational change and redesign: Ideas and insights for improving performance*, (pp. 346-379). London: Wiley-Blackwell.
- Weller, M. (2011). *The digital scholar: How technology is changing scholarly practice*. Bloomsbury Academic: London. doi: <http://dx.doi.org/10.5040/9781849666275>  
[http://www.bloomsburyacademic.com/view/DigitalScholar\\_9781849666275/book-ba-9781849666275.xml](http://www.bloomsburyacademic.com/view/DigitalScholar_9781849666275/book-ba-9781849666275.xml)
- Weller, M. (2014). *The battle for open: How openness won and why it doesn't feel like victory*. Ubiquity Press. doi: <http://dx.doi.org/10.5334/bam>
- West, M. (1953). *A General Service List of English Words*. London: Longman, Green and Co.
- West, R. (2018). Developing an open textbook for learning and instructional design technology. *Tech Trends*, 1–10. <https://doi.org/10.1007/s11528-018-0263-z>
- Whitaker, M. P. (1996). Ethnography as learning: A Wittgensteinian approach to writing ethnographic accounts. *Anthropological Quarterly*, 69(1), 1–13.
- Widdowson, H. G. (1983). *Learning purpose and language use*. Oxford: Oxford University Press.



- Widdowson, H. (2000). The limitations of linguistics applied. *Applied Linguistics*, 21(1), 3–25.
- Wiley, D. (n.d.). Defining the “open” in open content. Retrieved from <http://opencontent.org/definition/>
- Wiley D. (2011, July 27). Openwashing – the new Greenwashing [Blog post]. *Iterating toward openness*. Retrieved from <http://opencontent.org/blog/archives/1934>
- Willinsky, J. (2002). Copyright contradictions in scholarly publishing. *First Monday*, 7(11).
- Willis, J. (1998). Concordances in the classroom without a computer. In Tomlinson, B. (Ed.). *Materials Development in Language Teaching* (pp. 44–66). Cambridge: Cambridge University Press,
- Wise, A. F., Cui, Y., & Vytasek, J. (2017). Mining for gold: Identifying content-related MOOC discussion threads across domains through linguistic modelling. *Internet and Higher Education*, 32, 11–28.
- Witten, I.H., Paynter, G., Frank, E., Gutwin, C., & Neville-Manning, C.G. (1999). KEA: Practical automatic keyphrase extraction. In *Proceedings of Digital Libraries* (pp. 254–255).
- Witten, I.H., Bainbridge, D., & Nichols, D.M. (2009). *How to build a digital library* (Second Edition). Morgan Kaufmann.
- Witten, I.H., Wu, S., & Yu.X. (2011). Linking digital libraries to courses...with particular application to language learning. In *Proceedings of the 4th International Conference on Computer Supported Education* (pp. 5–14). Noordwijkerhout, The Netherlands: SciTePress.
- Witten, I.H., Wu, S., Li, L., & Whisler, J. (2013). *The book of FLAX: A new approach to computer assisted language learning*. (Second Edition). University of Waikato, New Zealand. Retrieved from [http://flax.nzdl.org/BOOK\\_OF\\_FLAX/BookofFLAX%20up.pdf](http://flax.nzdl.org/BOOK_OF_FLAX/BookofFLAX%20up.pdf)
- Witten, I.H., Eibe, F., Hall, M., & Pal, C.J. (2016). *Data mining: Practical machine learning tools and techniques* (Fourth Edition). Morgan Kaufmann.
- Witten, I.H. (September 2017). Mining MOOCs to assist second language learners. Seminar given at the Knowledge Media Institute, Berrill Building, The Open University, September 2017. Milton Keynes, United Kingdom. Retrieved from <http://stadium.open.ac.uk/stadia/preview.php?s=29&whichevent=2905>

- Wu, S., Franken, M., & Witten, I. H. (2009). Refining the use of the web (and web search) as a language teaching and learning resource. *Computer Assisted Language Learning*, 22(3), 249–268.
- Wu, S., Franken, M., & Witten, I. H. (2010). Supporting collocation learning with a digital library. *Computer Assisted Language Learning*, 23(1), 87–110.
- Wu, S., Witten, I. H., & Franken, M. (2010). Utilizing lexical data from a Web-derived corpus to expand productive collocation knowledge. *ReCALL*, 22(1), 83–102.
- Wu, S. (2010). Supporting collocation learning. *Doctoral Thesis, Computer Science Department, University of Waikato, New Zealand*. Retrieved from the Research Commons at the University of Waikato.
- Wu, S., Franken, M., & Witten, I.H. (2012). Collocation games from a language corpus. In H. Reinders (Ed.), *Digital games in language learning and teaching* (pp. 209–229). England: Macmillan Publishers.
- Wu, S., Fitzgerald, A., & Witten, I. H. (2014). Second language learning in the context of MOOCs. In S. Zvacek, M. T. Restivo, J. Uhomobhi, & M. Helfert (Eds.), *Proceedings of the 6th International Conference on Computer Supported Education* (pp. 354–359). Barcelona: SciTePress.
- 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 (IJCALLT)*, 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.
- Wu, S., Fitzgerald, A., Witten, I.H. & Yu, A. (2018). Automatically augmenting academic text for language learning: PhD abstract corpora with the British Library. In B. Zou, M. Thomas (Eds.), *Integrating Technology into Contemporary Language Learning and Teaching*, (pp. 512–537). IGI Global.
- Wu, S., Fitzgerald, A., Yu, A., & Witten, I.H. (2019). Developing and evaluating a learner-friendly collocation system with user query data. *International Journal of Computer-Assisted Language Learning and Teaching*, 9(2), 53–78.

- Yeh, Y., Li, Y.-H., & Liou, H.-C. (2007). Online synonym materials and concordancing for EFL college writing. *Computer Assisted Language Learning*, 20(2), 131–152. doi: 10.1080/09588220701331451
- Yoon, H., & Hirvela, A. (2004). ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing*, 13(4), 257–283.
- Yu, A., Wu, S., Witten, I. H., & König, J. L. (2016). Learning Collocations with FLAX Apps. In L.E. Dyson, W. Ng & J. Fergusson (Eds.), *Proceedings of the 15<sup>th</sup> World Conference on Mobile and Contextual Learning (mLearn 2016) – Sustaining Quality Research and Practice in Mobile Learning*, (pp. 291–294). Sydney: Australia.
- Zahar, R., Cobb, T., & Spada, N. (2001). Acquiring vocabulary through reading: effects of frequency and contextual richness. *Canadian Modern Language Review*, 15, 541-572.
- Zajacova, A., Lynch, S., & Espenshade, T. (2005). Self-efficacy, stress, and academic success in college. *Research in Higher Education*, 46, 677–706. doi:10.1007/s11162-004-4139-z
- Zimmerman, B. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist*, 25(1), 3–17.

## Appendices

### Appendix A.

#### Major historical milestones in the progress of Open Access publishing

Year	Milestone
1454	Invention of <b>printing</b>
1665	January 5: First issue of The <i><b>Journal des sçavans</b></i> (later spelled <i>Journal des savants</i> ), the earliest academic journal published in Europe and established by Denis de Sallo.
1807	25-year-old <b>Charles Wiley</b> opens a small printing shop at 6 Reade Street in lower Manhattan.
1842	May 10: <b>Julius Springer</b> founded what is now Springer Science+Business Media in Berlin.
1848	<b>John Wiley</b> (son of Charles Wiley) gradually started shifting his focus away from literature toward scientific, technical, medical, and other types of nonfiction publishing.
1880	Foundation of <b>Elsevier</b> .
1936	First scientific book published by <b>Elsevier</b> .
1990	First <b>web page</b> .
1991	An online repository of electronic preprints, known as e-prints, of scientific papers is founded in Los Alamos by the American physicist Paul Ginsparg. It was renamed to <b>ArXiv.org</b> in 1999. The total number of submissions by May 11st, 2016 (after 24.8 years) is 1,143,129 ( <a href="http://arxiv.org/stats/monthly_submissions">arxiv.org/stats/monthly_submissions</a> ).
1993	Creation of the <b>Open Society Institute</b> (renamed to the Open Society Foundations [OSF] since 2001) by the progressive liberal business magnate George Soros. The OSF financially supports civil society groups around the world, with a stated aim of advancing justice, education, public health and independent media.
1997	Launch of <b>SciELO</b> in Brazil. There are currently 14 countries in the SciELO network and its journal collections: Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica, Cuba, Mexico, Peru, Portugal, South Africa, Spain, Uruguay, and Venezuela.
1998	<b>Public Knowledge Project</b> (PKP) is founded by John Willinsky in the Faculty of Education at UBC, with Pacific Press Professorship endowment, dedicated to improving the scholarly and public quality of research.
	PKP has created the <b>Open Conference Systems</b> (2000), <b>Open Journal Systems</b> (2001), <b>Open Harvester Systems</b> (2002) and the <b>Open Monograph Press</b> (2013).

2000	<b>BioMed Central</b> , the self-described first and largest OA science publisher and <b>PubMed Central</b> , a free digital repository for biomedical and life sciences journal, is founded. In 2008, Springer announces the acquisition of BioMed Central, making it, in effect, the world's largest open access publisher.
2001	An online petition calling for all scientists to pledge that from September 2001 they would discontinue submission of papers to journals which did not make the full-text of their papers available to all, free and unfettered, either immediately or after a delay of several months is released. The petition collected 34,000 signatures but publishers took no strong response to the demands. Shortly thereafter, the <b>Public Library of Science (PLOS)</b> was founded as an alternative to traditional publishing. <i>PLOS ONE</i> is currently the world's largest journal by number of papers published (about 30,000 a year in 2015).
	December 1–2: <b>Conference convened in Budapest</b> by the Open Society Institute to promote open access – at the time also known as Free Online Scholarship. Where the Budapest Open Access Initiative (BOAI) was born.
2002	February 14th: Release of the <b>Budapest Open Access Initiative (BOAI)</b> , a public statement of principles relating to OA to the research literature. This small gathering of individuals is recognised as one of the major defining events of the OA movement. On the occasion of the 10th anniversary of the initiative, it was reaffirmed in 2012 and supplemented with a set of concrete recommendations for achieving "the new goal that within the next ten years, Open Access will become the default method for distributing new peer-reviewed research in every field and country."
	Start of the Research in Health - <b>HINARI</b> programme of the World Health Organization and major publishers to enable developing countries to access collections of biomedical and health literature online at reduced subscription costs. Together with Research in Agriculture - <b>AGORA</b> , Research in the Environment - <b>OARE</b> and Research for Development and Innovation - <b>ARDI</b> programmes, it currently forms <b>Research4Life</b> that provides developing countries with free or low-cost access to academic and professional peer-reviewed content online.
2008	The <b>National Institutes of Health (NIH)</b> Public Access Policy, an OA mandate requiring that research papers resulting from NIH funding must be freely and publicly available through PubMed Central within 12 months of publication, is officially recorded.
	The <b>Fair Copyright in Research Works Act</b> (Bill H.R 801 IH, also known as the "Conyers Bill") is submitted as a direct response to the National Institutes of Health (NIH)

	Public Access Policy; intending to reverse it. The bill's alternate name relates it to U.S Representative John Conyers (D-MI), who introduced it at the 111th United States Congress on February 3, 2009.
2011	Arrest of <b>Aaron Swartz</b> after he systematically downloaded articles from JSTOR, for alleged copyright infringement.
	In reaction to the high cost of research papers behind paywalls, <b>Sci-Hub</b> , the first known website to provide automatic and free, but illegal, access to paywalled academic papers on a massive scale, is founded by Alexandra Elbakyan from Kazakhstan.
2012	Start of the <b>Academic Spring</b> , a trend wherein academics and researchers began to oppose restrictive copyright in traditional academic journals and to promote free online access to scholarly articles.
	Start of the <b>Cost of Knowledge</b> campaign which specifically targeted Elsevier. It was initiated by a group of prominent mathematicians who each made a commitment to not participate in publishing in Elsevier's journals, and currently has over 15,933 co-signatories.
	Start of the United States-based campaign <b>Access2Research</b> in which open access advocates (Michael W. Carroll, Heather Joseph, Mike Rossner, and John Wilbanks) appealed to the United States government to require that taxpayer-funded research be made available to the public under open licensing. This campaign was widely successful, and the directive and FASTR (the Fair Access to Science and Technology Research Act) have become defining pieces in the progress of OA in the USA at the federal level.
	Launch of <b>PeerJ</b> , an OA journal that charges publication fees through researcher memberships, not on a per-article basis, resulting in what has been called "a flat fee for 'all you can publish'". Note that as of October 2015 <i>PeerJ</i> also have a flat rate APC of \$695.
2013	January: The suicide of <b>Aaron Swartz</b> draws new international attention for the Open Access movement.
	November: <b>Berlin 11 Satellite Conference</b> for students and early career researchers, which brought together more than 70 participants from 35 countries to engage on Open Access to scientific and scholarly research.
2014	First <b>OpenCon</b> in Washington DC, an annual conference for students and early career researchers on Open Access, Open Data, and Open Educational resources.
	Open Access is embedded the European Commission's <b>Horizon 2020</b> Research and Innovation programme.

2015	Academic publisher <b>Elsevier</b> makes a complaint in New York City for copyright infringement by <b>Sci-Hub</b> . Sci-Hub is found guilty and ordered to shut down. The website re-emerges under a different domain name as a consequence. A second hearing in March 2016 is delayed due to failure of the defendant to appear in court, and to gather more evidence for the prosecution.
------	--

Note: Reprinted from Tennant, J.P., Waldner, F., Jacques, D.C., Masuzzo, P., Collister, L.B., & Hartgerink, C.H.J. (2016). The academic, economic and societal impacts of Open Access: an evidence-based review [version 2; referees: 4 approved, 1 approved with reservations]. *F1000Research*, 5:632. doi: 10.12688/f1000research.8460.2

Appendix B  
Non-formal Learning Support (Type A)

Survey question: “*Learning Support (Type A): In addition to using FLAX, which, if any, of the following learning support did you/your learners use?*”

<b>Learners OER Hub study (N=1921)</b>	<b>Learners FLAX study (N=163)</b>	<b>ECL MOOC Learners (N=60 of 163)</b>	<b>Contracts X MOOC Learners (N=57 of 163)</b>	<b>Copyright X Learners (N=46 of 163)</b>	<b>Copyright X Teachers (N=11)</b>
Discussion with learning peers via social networks e.g. Facebook, Twitter:					
26.20%	38.05%				0.00%
Discussion with learning peers via video chat:					
NA	11.66%				0.00%
Writing my/their own study notes:					
50.50%	47.24%				18.18%
Use of a learning journal /diary/blog:					
25%	15.95%				9.09%
Use of a study calendar/plan:					
24.2%	32.52%				0.00%
<b>Additional Dedicated Learning Support in the English Common Law MOOC</b>					
Discussion with tutors and learning peers in the online forums of the Coursera MOOC platform:					



NA	43.33%	NA	NA	NA
Consulting the web links provided in the Coursera MOOC platform:				
NA	31.67%	NA	NA	NA
Engaging with the weekly practice test questions, Professor's Questions and Challenges:				
NA	3.33%	NA	NA	NA
<b>Additional Dedicated Learning Support in the ContractsX MOOC</b>				
Discussion with teaching fellows and learning peers in the online forums in the edX MOOC platform:				
NA	NA	35.09%	NA	NA
Using the peer assessment tool in the edX MOOC platform:				
NA	NA	33.33%	NA	NA
Completing the weekly unit tests:				
NA	NA	10.53%	NA	NA
<b>Additional Open Learning Support in CopyrightX</b>				
Discussion with teaching fellows and learning peers in the CopyrightX online forums and weekly tutorials via AdobeConnect:				
NA	NA	NA	50.00%	54.55%
Consulting extra resources on the CopyrightX website:				
NA	NA	NA	45.65%	18.18%

Appendix C  
Non-formal Learning Support (Type B)

Survey question: *“Learning Support (Type B): In addition to FLAX, which of the following features, if any, do you believe motivated you/your learners to study?”*

<b>Learning Support (Type B)</b>	<b>Non-formal Learners (N=163)</b>	<b>CopyrightX Teachers (N=11)</b>
Being issued with a certificate for completing the course	47.85%	54.55%
Having access to the CopyrightX website / edX/Coursera platform for information about and content related to the course	71.78%	45.45%
Having access to the discussion forum online to raise any relevant issues or questions	25.15%	0.00%
Having weekly tutorials with a teaching fellow (CopyrightX) / unit tests (ContractsX MOOC) / practice questions (ECL MOOC) to provide support with the course	33.13%	54.54%
Being provided with resources or advice about how to succeed on the course e.g. previous exam materials / questions to the professor / peer assessments	35.58%	27.27%
Being able to discuss with other learners on the course about my experiences (e.g. through forums, Facebook groups, Twitter, meet-ups etc.)	19.02%	0.00%
Being required to successfully complete the final exam to earn a certificate or passing grade	60.74%	63.64%

## Appendix D

### CopyrightX collection in FLAX log data

Total number of clicks: 13157 (2015 and 2016 courses)

User action	User sub-action	FLAX function pathway	No. of clicks	Clicks in %
<b>About Collection</b>				
Click on link to FLAX from CopyrightX LMS to arrive on the landing page or “About Collection” page	Click on YouTube CopyrightX FLAX training videos: 200 views (vid. 1) 105 views (vid. 2)	CollectionAbout	2032	NA
<b>Browse</b>				
Click the “Lectures” or “Readings” buttons in the main menu to browse course documents		ClassifierBrowse	3384	25.72
	Click within a course document (e.g. lecture or reading) on the wordlist, wikify and collocation part of speech (adjective, noun, verb) tabs	FlaxWordListDocument-Retrieve WikifyArticle FlaxCollocationDocument-Retrieve	7625	57.95
	Look up a term or concept definition in a wikified course document	WikipediaArticleDefinitionRetrieve	81	0.61
Click the “collocations” button in the main menu		FlaxCollocationBrowse	101	0.76

Click a collocation [see <b>List A</b> of words]	Click “top 100” collocations (7)	FlaxCollocationRetri eve	99	0.75
	Look up the context of a collocation	FlaxCollocationCont ext-Retrieve	6	0.04
Click highlighted phrase in document to activate “Collocation Notepad” with cherry icon.	Click the “My Cherry Basket” button	CherryPicking	65	0.49
	Add cherry (collocational phrase)		1	0.007
	Add category		1	0.007
Click the “wordlist” button on the main menu		FlaxWordListBrowse	107	0.81
Click the “lexical bundles” button		FlaxLexicalBundleBr owse	86	0.65
	Click on a bundle to view the context	FlaxSampleRetrieve	26	0.19
<b>Search</b>				
Click the “search” button in main menu to query keywords and phrases at the corpus (article) level [see <b>List B</b> for query words]		AdvancedFieldQuery	270	2.05
	Query a keyword or phrase at the course document (article) level	AdvancedFieldQuery with a query word s1.fqv	157	1.19
Click the “search” button in the main menu to query keywords and phrases at the sentence level Or		FlaxWordQuery	129	0.98
	Click on green arrow to reveal how search term(s) are used in the wider context of the course documents	FlaxTextRetrieve	51	0.38

Click the “wordlist” button in the main menu to browse and then query keywords at the sentence level [see <b>List C</b> for query words]				
<b>Activities</b>				
Click the “activities” button in the main menu		CollectionActivity	433	3.29
	ContentWordGuessing		22	0.16
	CollocationalFillinBlanks		116	0.88
	CollocationGuessing		25	0.19
	RelatedWords		17	0.12
	ScrambleSentence		77	0.58
	CollocationDominoes		79	0.60
	SplitSentences		60	0.45
Click the “design activity” button		DesignActivity	139	1.05

### List A

'assignment', 'scenes', 'photo', 'address', 'author', 'affidavit', 'arrangement', 'algorithm', 'abrogation', 'case', 'record', 'alternatively', 'british', 'welfare', 'abbreviation', 'above', 'above', 'able', 'quote', 'generate', 'requirement', 'create', 'publisher', 'enable', 'version', 'fee', 'so', 'cultural', 'legal', 'abbreviation', 'age', 'computer', 'computer', 'key', 'program', 'identify', 'publish', 'access', 'approach', 'benefit', 'issue', 'author', 'welfare', 'involve', 'consist', 'section', 'require', 'seek', 'consumer', 'principle', 'grant', 'individual', 'lecture', 'constitute', 'revenue', 'license', 'distribute', 'context', 'creative', 'legal', 'theory', 'design', 'factor', 'available', 'The', 'issue', 'indication', 'abrogation', 'moral', 'fixation', 'borderline', 'box', 'breakfast', 'bright', 'academic'

## List B

'substantial', 'criminal', 'vicarious', 'sega', 'doctrinal', 'doctrinal', 'welfare', 'welfare', 'Fair', 'altai', 'altai', 'altai', 'altai', 'altai', 'altai', 'altai', 'altai', 'altai', 'altai', 'altai', 'TRIPS', 'locke', 'Visual', 'Formalities', 'fragmented', 'fragmented', 'fragmented', 'fragmented', 'fragmented', 'fragmented', 'copyright', 'Moral', 'traditional', 'three', 'michael', 'harper', 'deivative', 'moral', 'blue', 'blue', 'blue', 'blue', 'blue', 'blue', 'blue', 'blue', 'blue', 'photo', 'photo', 'blue', 'blue', 'Craig', 'Craig', 'Craig', 'formalities', 'direct', 'direct', 'right', 'traditional', 'traditional', 'traditional', 'traditional', 'traditional', 'traditional', 'traditional', 'traditional', 'traditional', 'One', 'One', 'One', 'One', 'dignity', 'dignity', 'dignity', 'dignity', 'dignity', 'moral', 'Martin', 'Luther', 'moral', 'originality', 'timing', 'Moral', 'originality', 'rendition', 'waldron', 'public', 'public', 'public', 'public', 'political', 'VARA', 'VARA', 'visual', 'a', 'Code', 'Sound', 'Scope', 'fair', 'fair', 'integrity', 'playwright', 'integrity', 'integrity', 'integrity', 'visual', 'visual', 'visual', 'visual', 'vara', 'personality', 'visual', 'first', 'moral', 'derivative', 'integrity', 'snow', 'derivative', 'derivative', 'derivative', 'fair', 'fair', 'joint', 'work', 'bedamax', 'fairey', 'mechanical', 'foolishness', 'napster', 'fairness', 'proportional', 'heirs', 'In', 'In', 'cable', 'related', 'related', 'wheal'

### List C

diminution', 'diminution', 'copyrightx', 'Borrowed', 'agreement', 'alluded', 'mannion', 'feist', 'feist',  
'scenes', 'scenes', 'scenes', 'scenes', 'scenes', 'scenes', 'predictability', 'prediction', 'transformation',  
'photo', 'photo', 'pursue', 'issue', 'VARA', 'Dastar', 'prevailing', 'prevailing', 'prevailing', 'derivative',  
'issue', 'grant', 'creative', 'recreate', 'exclusive', 'exclusive', 'principle', 'creative', 'create', 'creator',  
'creative', 'creative', 'creative', 'creative', 'negotiations', 'creation', 'theory', 'author', 'creative',  
'compensate', 'concept', 'With', 'To', 'author', 'license', 'unlicensed', 'license', 'author', 'potentially',  
'adopted', 'The', 'grumbling', 'grumbling', 'grumbling', 'amendment', 'impose', 'behaviour',  
'behavior', 'To', 'exceptions', 'de', 'fisher', 'Lecture', 'The', 'transcript', 'copyright', 'deterrence',  
'copyright', 'legal', 'Dramatic', 'Terry', 'Dramatic', 'Fisher', 'lecture', 'work', 'william', 'principles',  
'principle', 'meaning', 'performance', 'labor', 'transmission', 'matter', 'Craig', 'Craig', 'Craig', 'Craig',  
'Craig', 'Craig', 'Martin', 'Luther', 'Luther', 'Moral', 'photograph', 'fair', 'publish', 'test', 'To',  
'System', 'visual', 'creative', 'recreate'

## Appendix E

### English Common Law MOOC collection in FLAX log data

Total number of clicks: 8494 (2014, 2015 and 2016 courses)

User action	User sub-action	FLAX function pathway	No. of clicks	Clicks in %
<b>About Collection</b>				
Click on link to FLAX from English Common Law Coursera MOOC platform to arrive on the landing page or “About Collection” page	Click on YouTube ECL MOOC FLAX training videos: 561 views (vid. 1) 214 views (vid. 2) 147 views (vid. 3)	collectionAbout	1863	NA
<b>Browse</b>				
Click the “Lectures”, “Quizzes” or “Extras” buttons in the main menu to browse course documents		ClassifierBrowse	2183	25.70
	Click within a course document (e.g. lecture or reading) on the wordlist, wikify and collocation part of speech (adjective, noun, verb) tabs	FlaxWordListDocument-Retrieve WikifyArticle FlaxCollocationDocument-Retrieve	2890	34.02
	Look up a term or concept definition	WikipediaArticleDefinitionRetrieve	660	7.77

	in a wikified course document			
Click the “collocations” button in the main menu		FlaxCollocationBrowse	200	2.35
Click a collocation [see <b>List A</b> of words]	Click “top 100” collocations (24)	FlaxCollocationRetrieve	225	2.64
	Look at the context of a collocation	FlaxCollocationContextRetrieve	6	0.07
Click highlighted phrase in document to activate “Collocation Notepad” with cherry icon.	Click the “My Cherry Basket” button	CherryPicking	63	0.74
	Add Cherry (collocational phrase)		0	0.0
	Add category		1	0.01
Click the “wordlist” button in the main menu		FlaxWordListBrowse	138	1.62
Click the “lexical bundles” button in the main menu		FlaxLexicalBundleBrowse	134	1.57
	Click on a bundle to view the context	FlaxSampleRetrieve	44	0.51
<b>Search</b>				
Click the “search” button in main menu to query keywords and phrases at the corpus (article) level [see <b>List B</b> for query words]		AdvancedFieldQuery	152	1.78



	Query a keyword or phrase at the course document (article) level	AdvancedFieldQuery with a query word sl.fqv	57	0.67
Click the “search” button in the main menu to query keywords and phrases at the sentence level Or Click the “wordlist” button in the main menu to browse and then query keywords at the sentence level [see <b>List C</b> for query words]		FlaxWordQuery	157	1.84
	Click on green arrow to reveal how search term(s) are used in the wider context of the course documents	FlaxTextRetrieve	14	0.16
<b>Activities</b>				
Click the “activities” button in the main menu		CollectionActivity	535	6.29
	ContentWordGuessing		219	2.57

	CollocationalFillin Blanks		120	1.41
	CollocationGuessi ng		135	1.58
	RelatedWords		139	1.63
	ScrambleSentence		30	0.35
	CollocationDomin oes		32	0.37
	SplitSentences		34	0.40
Click the “design activity” button		DesignActivity	139	1.63

### List A

'supreme', 'abrogate', 'absence', 'parliament', 'absence', 'show', 'abrogate', 'account', 'legal', 'say', 'abuse', 'instance', 'case', 'court', 'structure', 'amount', 'first', 'spirit', 'concept', 'principle', 'court', 'party', 'unwritten', 'common', 'degree', 'ability', 'baby', 'good', 'transparency', 'able', 'appeal', 'administration', 'apply', 'influence', 'conservative', 'ambiguity', 'bind', 'bind', 'bind', 'bind', 'appellant', 'avoidance', 'avoidance', 'avoidance', 'avoidance', 'avoidance', 'avoidance', 'avoidance', 'avoidance', 'avoidance', 'avoidance', 'writ', 'lecture', 'avoidance', 'abrogate', 'adjudicate', 'misrepresentation', 'abandon', 'control', 'judicial', 'abandon', 'absolute', 'agency', 'activist', 'analogous', 'abusive', 'abrogate', 'young', 'adjustment', 'ability', 'abandon', 'assembly', 'claim', 'batter', 'family', 'family', 'family', 'absurdity', 'accord', 'mischief', 'academic', 'absolutely', 'debt', 'edition', 'feel', 'black', 'boy', 'breach', 'decisis', 'equity', 'eu', 'gay', 'interpretative', 'ius', 'interpretive', 'avoidance', 'precedent', 'acknowledge', 'Equity', 'appearance', 'inaction', 'devise', 'legal', 'evoke', 'appellate', 'abrogate', 'alteram', 'Common', 'ability', 'absolute', 'federalism', 'heterosexual', 'master', 'avoidance', 'pur', 'partisan', 'legal', 'legal', 'ability', 'legal', 'application', 'able', 'accord', 'abuse', 'antifascist', 'appellant', 'formal', 'information'

### List B

'legal', 'common', 'case', 'precedent', 'human', 'darcy', 'buckmaster', 'Hound', 'literal', 'literal', 'justice', 'rosset', 'English', 'plaintiff', 'prerequisite', 'norway', 'seeing', 'Pepper', 'FAMILY',



Appendix F  
ContractsX MOOC collection in FLAX log data

Total number of clicks: 1769 (2016 course)

User action	User sub-action	FLAX function pathway	No. of clicks	Clicks in %
<b>About Collection</b>				
Click on link to FLAX from ContractsX edX MOOC platform to arrive on the landing page or “About Collection” page	Click on YouTube ContractsX FLAX training videos: 279 views (vid. 1) 132 views (vid. 2) 81 views (vid. 3)	collectionAbout	716	NA
<b>Browse</b>				
Click the “Browse by Title” button in the main menu to browse course documents		ClassifierBrowse	191	10.79
	Click within a course document (e.g. lecture or reading) on the wordlist, wikify and collocation part of speech (adjective, noun, verb) tabs	FlaxWordListDocument-Retrieve WikifyArticle FlaxCollocationDocument-Retrieve	420	23.74
	Look up a term or concept definition in a wikified course document	WikipediaArticleDefinitionRetrieve	289	16.33
Click the “collocations” button in the main menu		FlaxCollocationBrowse	42	2.37

Click a collocation [see <b>List A</b> of words]	Click “top 100” collocations (10)	FlaxCollocationRetri eve	34	1.92
	Look up the context of a collocation	FlaxCollocationCont extRetrieve	5	0.28
Click highlighted phrase in document to activate “Collocation Notepad” with cherry icon.	Click the “My Cherry Basket” button	CherryPicking	41	2.31
	Add cherry (collocational phrase)		9	0.50
	Add category		6	0.33
Click the “wordlist” button in the main menu		FlaxWordListBrows e	41	2.31
Click the “lexical bundles” button in the main menu		FlaxLexicalBundleB rowse	32	1.80
	Click on a bundle to view the context	FlaxSampleRetrieve	28	1.58
<b>Search</b>				
Click the “search” button in main menu to query keywords and phrases at the corpus (article) level [see <b>List B</b> for query words]		AdvancedFieldQuer y	101	5.70
	Query a keyword or phrase at the course document (article) level	AdvancedFieldQuer y with a query word sl.fqv	98	5.53
Click the “search” button in the main menu to query keywords and phrases at the sentence level		FlaxWordQuery	60	3.39
	Click on green arrow to reveal how search term(s) are used in the	FlaxTextRetrieve	40	2.26

Or Click the “wordlist” button in the main menu to browse and then query keywords at the sentence level [see <b>List C</b> for query words]	wider context of the course documents			
<b>Activities</b>				
Click the “activities” button in the main menu		CollocationActivity	132	7.46
	ContentWordGuessing		4	0.22
	CollocationalFillinBlanks		31	1.75
	CollocationGuessing		14	0.79
	RelatedWords		44	2.48
	ScrambleSentence		5	0.28
	CollocationDominoes		28	1.58
Click the “design activity” button		DesignActivity	74	4.18

### List A

'access', 'analogous', 'interpretation', 'Frolic', 'promise', 'age', 'average', 'agreement', 'able',  
'acceptance', 'beneficiary', 'airport', 'mutuality', 'option', 'account'

### List B

'promise', 'dead', 'lumber', 'deadweight', 'manuscript', 'option', 'Offer', 'promises', 'buying', 'now',  
'offer', 'acceptance', 'deadweight', 'buying', 'buying', 'Now', 'mutual', 'acceptance', 'implicit',  
'circuit', 'original', 'charitable', 'charitable', 'intent', 'subscriptions', 'subscriptions', 'subscriptions',  
'subscriptions', 'but', 'subscription', 'charitable', 'charitable', 'Reliance', 'charitable', 'reliance',  
'Reliance', 'Hoffman', 'Reliance', 'Pennzoil', 'reliance', 'comcast', 'reliance', 'reliance', 'meeting',

'gambling', 'gambling', 'mutual', 'Fraud', 'Duty', 'Hypotheticals', 'POM', 'Krell', 'Krell',  
 'Frustration', 'Premises', 'Taxi', 'Hypothetical', 'shipping', 'music', 'lumley', 'identification',  
 'bookstore', 'impracticability', 'impracticability', 'impracticability', 'gamble', 'lumber', 'lumber',  
 'lumber', 'interpretation', 'interpretation', 'Unit', 'what', 'interpretation', 'Part', 'interpretation', 'the',  
 'twin', 'World', 'shoveling', 'expectation', 'silver', 'deadweight', 'deadweight', 'detrimental',  
 'estoppel', 'beneficiary', 'reliance', 'time', 'snow', 'snow', 'shoveling', 'Snow', 'Batsakis',  
 'deadweight', 'deadweight', 'deadweight', 'trust'

### **List C**

'contract', 'case', 'fluctuates', 'contracts', 'specific', 'we', 'interpretation', 'interpretation', 'mutuality',  
 'reliance', 'performance', 'promise', 'performance', 'expectation', 'specific', 'mutual', 'quo',  
 'violation', 'violate', 'violate', 'detrimental', 'party', 'appeal', 'beneficiary', 'principle', 'unilateral',  
 'following', 'mistake', 'reliance', 'enforce', 'reliance', 'mutuality', 'mutuality', 'reliance', 'enforce',  
 'after', 'exchange', 'was', 'clause', 'mutuality', 'reliance', 'option', 'option', 'reliance', 'offeror',  
 'reasonable', 'principle', 'webb', 'option', 'Batsakis', 'Batsakis', 'Batsakis', 'detrimental',  
 'commercial', 'commercial', 'taken', 'contract'

## Appendix G

### Essay topic list for legal English translation studies

<b>FLAX-BASED TEXTS</b>
1. Judicial Decisions: The Meaning of Precedent in Common Law
2. Parliament and Statutes
3. History and Peculiarities of the Common Law
4. Introduction to the Civil and Common Courts, The European Court, Parliaments and Europe.

<b>NON-FLAX-BASED TEXTS</b>
1. Family Law: A comparison between the Spanish, British and American Systems
2. Civil and Criminal Law in the Spanish and Common Law Systems
3. International Law
4. Powers of Attorney in the Spanish and Common Law Systems
5. An overview on Legal Translation in English and Spanish
6. Probate Law: Wills in Civil and Common Law Systems
7. Contracts: A Comparative Study
8. Royal Assent
9. Delegated legislation in the UK, USA and Spain