19 FLAX: Flexible and open corpus-based language collections development

Alannah Fitzgerald[1], Shaoqun Wu[2], and María José Marín[3]

Abstract

In this case study we present innovative work in building open corpus-based language collections by focusing on a description of the open-source multilingual Flexible Language Acquisition (FLAX) language project, which is an ongoing example of open materials development practices for language teaching and learning. We present language-learning contexts from across formal and informal language learning in English for Academic Purposes (EAP). Our experience relates to Open Educational Resource (OER) options and Practices (OEP) which are available for developing and distributing online subject-specific language materials for uses in academic and professional settings. We are particularly concerned with closing the gap in language teacher training where competencies in materials development are still dominated by print-based proprietary course book publications. We are also concerned with the growing gap in language teaching practitioner competencies for understanding important issues of copyright and licencing that are changing rapidly in the context of digital and web literacy developments. These key issues are being largely ignored in the informal language teaching practitioner discussions and in the formal research into teaching and materials development practices.

Keywords: FLAX, corpora, MOOC, OER, OEP, open access, open-source software, open data, domain-specific language, legal English.

1. Concordia University Department of Education Montreal, Canada; alannahfitzgerald@gmail.com.

2. University of Waikato Department of Computer Science Hamilton, New Zealand; shaoqunyw@gmail.com.

3. Universidad de Murcia Departamento de Filologia Inglesa Murcia, Spain; mariajose.marin1@um.es.

How to cite this chapter: 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. doi:10.14705/rpnet.2015.000281

1. Context/rationale

Corpus-based approaches for language learning, teaching and materials development have featured frequently in the research into Computer Assisted Language Learning (CALL) but they have yet to become mainstream practice in classroom-based language education. Accessibility remains a central issue whereby many existing corpus-based tools and resources are beyond the reach of most language learners and teachers. Restrictions stem from a combination of complex and often outdated user interface designs and still, in many cases, from subscription costs. Usability studies with corpus-based systems have also failed to materialise in the research and development trends from the growing body of literature dedicated to CALL.

Enter Massive Open Online Courses (MOOCs) and OERs where opportunities arise for the re-visioning and re-purposing of corpus-based approaches for the development of language support in online learning. In many ways this case study reflects our growing interest in online learning –an untapped educational environment that would appear to be a natural home for the uptake of web-based corpus tools and resources– where language support needs to be scaled for large numbers of users at minimal cost. It is our objective to bridge new contexts of online research, development and practice in open education with corpus-based approaches within traditional classroom-based language education. We are doing this by reusing open content and data to build domain-specific language collections with the FLAX system.

FLAX is defined as

> "an open-source software system designed to automate the production and delivery of interactive digital language collections [and language exercises. Source] material comes from digital libraries (language corpora, web data, open access (OA) publications, open educational resources) for a virtually endless supply of authentic [linguistic examples] in context. With simple interface designs, FLAX has been designed so that non-expert users –language teachers, language learners, subject specialists,

> instructional design and e-learning support teams– can build their own collections [of language, as well as their own exercises based on a wide pool of linguistic material].
>
> The FLAX software can be freely downloaded to build [language] collections with any text-based content and supporting audio-visual material, for both online and classroom use" (Fitzgerald, 2014, para. 3).

This case study will provide an ongoing example of the collaborative development of the Law Collections on the FLAX website for supporting formal and informal English language learning with corpus-based approaches.

2. Aims and objectives

- To demonstrate how subject-specific language collections can be built with the FLAX open-source software for uses across formal and informal education, as exemplified by the Law Collections development on the FLAX website.

- To engage language teaching and research practitioners in the design process of subject-specific collections development in FLAX, and to research the efficacy of these collections for uptake by learners and teachers in MOOCs as well as in traditional classroom-based language learning.

- To share a methodology for distributing openly available tools and resources for subject-specific language education.

In a research and development project with FLAX for building subject-specific language learning collections, we sourced relevant open content in the area of socio-legal English, including the 8.85 million-word British Law Reports Corpus (BLaRC) (Figure 1), MOOC lectures, OA law journal and PhD theses publications (Fitzgerald, Wu, & Barge, 2014).

Figure 1. BLaRC in the FLAX Law Collections

Table 1. Open resources featured in and linked to the FLAX Law Collections (Fitzgerald, 2014, section “Law Collections in FLAX”)

Type of Resource	Number and Source of Collection Resources
Open Access Law research articles	40 Articles (DOAJ – Directory of Open Access Journals[1], with Creative Commons licenses for the development of derivatives).
MOOC lecture transcripts and videos (streamed via YouTube and Vimeo)	4 MOOC Collections: English Common Law (University of London with Coursera)[2], Age of Globalization (Texas at Austin with edX)[3], Copyright Law (Harvard with edX)[4], Environmental Politics and Law (OpenYale).
Podcast audio files and transcripts (OpenSpires)	15 Lectures (Oxford Law Faculty and the Centre for Socio-Legal Studies).
PhD Law thesis writing	50-70 EThoS Theses[5] (sections: abstracts, introductions, conclusions) at the British Library (Open Access but not licensed as Creative Commons – permission for reuse granted by participating Higher Education Institutions).

1. http://doaj.org/

2. https://www.coursera.org/course/engcomlaw

3. https://www.edx.org/course/utaustinx/utaustinx-ut-3-02x-age-globalization-2626

4. http://copyx.org/

5. http://ethos.bl.uk/Home.do;jsessionid=4F2E6E1673362D6ED04702DFA665C081

BLaRC	8.85 million-word corpus derived from free legal sources at the British and Irish Legal Information Institute (BAILII)[1] aggregation website.
Legal Terms List	A legal English vocabulary derived from the BLaRC using two Automatic Term Recognition Methods.
FLAX Wikipedia English	Linking in a reformatted version of Wikipedia (English version), providing key terms and concepts as a powerful gloss resource for the Law Collections.
FLAX Learning Collocations	Linking in lexico-grammatical phrases from the British National Corpus (BNC)[2] of 100 million words, the British Academic Written English corpus (BAWE)[3] of 2500 pieces of assessed university student writing from across the disciplines, and a re-formatted Wikipedia corpus in English of approximately 2.5 million articles.
FLAX Web Phrases	Linking in a reformatted Google n-gram corpus (English version) containing 380 million five-word sequences drawn from a vocabulary of 145,000 words.

3. What we did:developing demonstration open Law Collections in FLAX

The following sections outline how we built the Law Collections in FLAX and key aspects of their functionality for language teaching. The features described offer a model of how FLAX can be used. The approach is fully automated and can be applied to any FLAX language collection.

Functionality. Ease of navigation and attractive, simple user interfaces are central to FLAX. Iterations of the FLAX software to create your own stand-alone FLAX server and to implement the FLAX MOODLE module (within the MOODLE virtual learning environment) are available for download on the FLAX website. With the development of the FLAX MOODLE module, new and simpler teacher interfaces were developed to move away from the more complex librarian interfaces used in the standard Greenstone digital library software, which the FLAX open-source software is an extension of (Witten, Wu, & Yu, 2011).

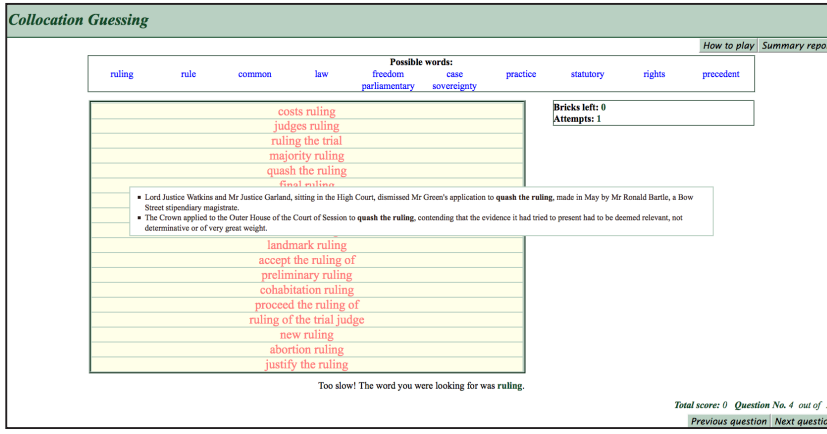
1. http://www.bailii.org/

2. http://www.natcorp.ox.ac.uk/

3. http://www.coventry.ac.uk/research/research-directory/art-design/british-academic-written-english-corpus-bawe/

The free *Book of FLAX* e-book, available on the FLAX website, tells you everything you need to know about building your own interactive FLAX collections featuring game-based activities like the one shown in Figure 2 below. A series of FLAX training videos in Chinese and English are also available on the FLAX website, with the latter featured on the Teacher Training Videos[1] website.

Figure 2. FLAX Collocations Guessing Game learner interface populated by the BNC corpus



Building open language collections in FLAX. Available on the FLAX website are completed collections and on-going collections being developed by registered users. All resources are pre-processed before being built into FLAX collections. For example, lecture transcripts and OA publications undergo simple editing, including division into subsections, and are reformatted into manageable chunks as HTML files to decrease the cognitive load for learners when listening and viewing.

How the open datasets are combined and used in the FLAX user interface. FLAX links relevant tools and reusable resources into streamlined online

1. http://teachertrainingvideos.com/

interfaces for language teachers and learners. By reusable resources, this can mean one of two things: those that are openly licenced, and those for which we have gained permission to use for non-commercial purposes. For example, some datasets used in FLAX (the BNC, the BAWE corpus, and the Google web dump n-grams corpus) have restrictive licences, but the open datasets (Wikipedia, the BLaRC, OER, and OA publications) have non-restrictive licences for language resource development purposes for uses in education and research.

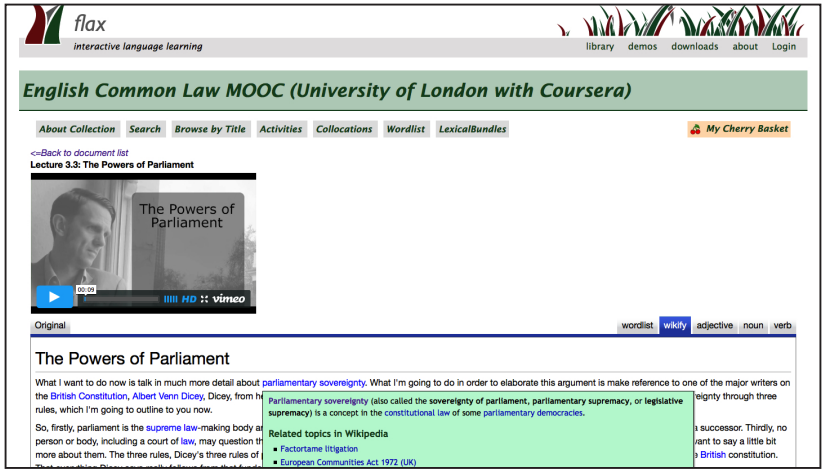
In the formatting stage of pre-processing documents for inclusion in the Law Collections in FLAX, licences originating from the different OER and OA data sources have been reflected accurately in the FLAX system to show the different permissions for reuse by end users. Built into the FLAX software when building collections is an acknowledgement message highlighted in blue for the collections builders to show that they are aware of the licencing permissions of the different resources they are using to make collections [*"Before you include any document in your collection, please ensure that you have copyright permission to do so"*]. However, actual practice with understanding and reusing the variety of copyrighted resources available online is not necessarily something with which language teachers are familiar or confident in handling. This is why we are building public collections with language teachers and learners on the FLAX website to demonstrate, and document through our research, best open educational and design practices for the development of language collections with the FLAX open-source software.

Video streaming and part-of-speech tagging. Audio-visual resources in the form of lectures and podcasts can be either embedded directly into the FLAX software or are streamed through well-known third party providers such as YouTube and Vimeo.

Wikipedia Miner toolkit. FLAX connects to the open-source Wikipedia Miner toolkit, also developed at the University of Waikato, to extract key concepts and their definitions from Wikipedia articles to assist with reading and vocabulary

in subject-specific areas as seen in Figure 3. Key concepts and their definitions are extracted from Wikipedia articles and are linked to documents in FLAX collections. For example, *Factortame litigation*, *European Communities Act 1972 (UK)*, *Thorburn v Sunderland City Council, Human Rights Act 1998*, *Supremacy (European Union Law)*, are identified as related topics in Wikipedia to provide a broader context for understanding the English Common Law MOOC sub-collection in FLAX, and a definition for *parliamentary sovereignty* is also extracted.

Figure 3. FLAX augmented text interface with wikify function in Law MOOC Collections



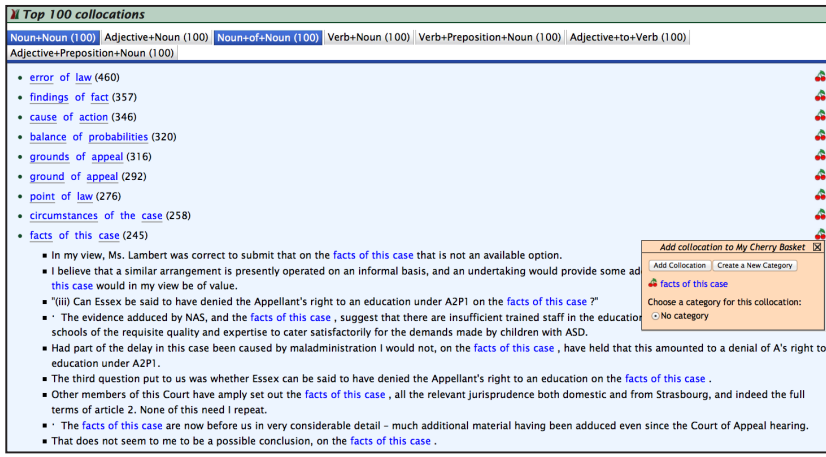
Search capabilities. Search queries in FLAX are highlighted in yellow for ease of recognition and can contain more than one word. For phrase searching, a query can be enclosed by quotation marks; for example, *"doctrine of precedent"* returns sentences containing this exact phrase, while *doctrine of precedent* returns sentences that contain these three words and associated words in any order, e.g. *this idea of the binding doctrine of precedent*.

Keywords and word lists. "The development of *wordlist* and *keyword* interfaces [in FLAX] also allows learners to analyse the range of vocabulary

used in a specified document, including the General Service List (West, 1953), the Academic Word List (AWL) by Coxhead (2000) and Off-list [topic-specific] words" (Fitzgerald, 2013, section "Open Linguistic Support in the Context of Formal and Informal EAP", para. 4). Words can be sorted alphabetically or by frequency; in either case, frequency in the corpus is shown alongside the word.

Collocations. It is possible to focus on lexical collocations with noun based structures noun + noun, adjective + noun, noun + of + noun, verb + noun, verb + preposition + noun, adjective + to + verb and adjective + preposition + noun as seen in Figure 4 because they are the most important and useful patterns for second language learners of subject-specific language.

Figure 4. Collocations in Law Collections linked to FLAX Learning Collocations (Wikipedia) collection



The Cherry Basket. By clicking on the cherry icon, also shown in Figure 4, users can go through the collections, selecting examples of language they wish to store and retrieve. By clicking on the blue hyperlinked words in the subject-specific collections, FLAX will link to a larger collocations database with the BNC, BAWE and Wikipedia corpora.

3.1. Who is using the open Law Collections in FLAX?

Following on from earlier work with the FLAX project into second language learning in the context of MOOCs (Fitzgerald, Wu, & Witten, 2014), we are currently investigating how the Law Collections in FLAX are being reused in the MOOCs listed earlier in this case study in Table 1 and in formal classroom-based language learning and translation contexts. The initial collections design work with language teachers at Queen Mary University of London focused on sourcing open resources that would be of relevance to their pre-sessional EAP law cohorts and more specifically with their postgraduate law students who require language support on their critical thinking and writing in-sessional programme. At the Universidad de Murcia in Spain, legal English translation students are reusing the English Common Law MOOC collection in FLAX to mine key lexico-grammatical patterns and prepare a class presentation and follow-on essay on the differences between the civil and common law systems.

3.2. Research with the open Law Collections in FLAX

To date, we have made the *English Common Law* and the *Age of Globalization* MOOC collections in FLAX available to 35,000+ registered learners in over a hundred different countries. We are reusing OER research instruments (surveys, interview and think aloud protocols) from the OER Research Hub[1] research bank based at the UK Open University to collect data on the following revised OER hypotheses[2] for language education using the FLAX collections in informal online learning and traditional classroom-based learning:

- *Hypothesis A*: use of OER language collections leads to improvement in student performance and satisfaction.

- *Hypothesis E*: use of OER for developing language collections leads to critical reflection by language educators, with improvement in their practice.

1. http://oerresearchhub.org/

2. Revised from Fitzgerald (2013, section "Multi-site Research into Developing Open Linguistic Support", para. 2).

- *Hypothesis K*: informal means of assessment are motivators to learning with OER language collections.

- *Hypothesis H*: informal learners adopt a variety of techniques to compensate for the lack of formal language support.

- *Hypothesis I*: open education acts as a bridge to formal language education, and is complementary, not competitive, with it.

4. Discussion

In terms of detailed feedback on the FLAX system with regards to using the Law Collections, the face-to-face research contexts are likely to yield more reliable findings into the actual efficacy of the system for impacting language learning. This will involve controlled and experimental groups to discern the impact of the FLAX system on learner writing and vocabulary acquisition for legal English through qualitative discourse analysis approaches. However, the type of data we can collect from MOOC learners will be quantitative. MOOC survey questions are matched to the OER research hypotheses to identify learners' perceptions of and use of the FLAX MOOC collections to support vocabulary, reading and listening comprehension of course content and for instances of language transfer into course discussions and peer-reviewed writing.

5. Conclusion

FLAX is committed to opening access in English language education through digital innovation. The FLAX system's capabilities for building language collections with comprehensive facilities for search and retrieval, and customised interactive learning of key subject terms and concepts, addresses the needs of both native and non-native speakers of English who are interested in engaging deeply with open subject-specific resources in English from the OER and OA movements. Furthermore, learners benefit from the enhancement of these open

resources with FLAX's affordances for linking in datasets derived from massive online sources, namely Wikipedia and Google, and from large pre-formatted research corpora such as the BNC, the BLaRC and the BAWE.

Acknowledgements. We would like to thank the OER Research Hub, the Global OER Graduate Network, and The International Research Foundation (TIRF) for English Language Education Doctoral Dissertation Grant, for funding this research collaboration between the FLAX project at the University of Waikato in New Zealand, the Department of Education at Concordia University in Canada, and the Departamento de Filologia Inglesa, Universidad de Murcia in Spain.

References

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238. Reprinted in 2007 in Corpus linguistics by W. Teubert & R. Krishnamurthy (Eds), *Critical concepts in linguistics* (pp. 123-149). Oxford, England: Routledge. doi:10.2307/3587951

Fitzgerald, A. (2013, March 18). Educating in beta. *OER Research Hub*. Retrieved from http://oerresearchhub.org/2013/03/18/educating-in-beta/

Fitzgerald, A. (2014, October 6). Wow! The FLAX language system – So much open data. *TOETOE Technology for Open English - Toying with Open E-resources (ˈtɔɪtɔɪ)*. Retrieved from http://alannahfitzgerald.org/2014/10/06/vici-competition/

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. doi:10.14705/rpnet.2014.000200

Fitzgerald, A., Wu, S., & Witten, I. H. (2014). Second language learning in the context of MOOCs. *Proceedings of the 6th International Conference on Computer Supported Education* (pp. 354-359).

Witten, I. H., Wu, S., & Yu, X. (2011). Linking digital libraries to courses with particular application to language learning. *Proceedings of the 3rd International Conference on Computer Supported Education, Volume 1, Noordwijkerhout, Netherlands, 6-8 May, 2011* (pp. 5-14).

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

FLAX website resource links

FLAX (Flexible Language Acquisition) project website: http://flax.nzdl.org

FLAX Age of Globalization MOOC collection (University of Texas at Austin): http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=lawlecture&if=

FLAX British Law Reports Corpus (BLaRC) collection: http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=BlaRC&if=

FLAX English Common Law MOOC collection (London University with Coursera): http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=englishcommonlaw&if=

FLAX Learning Collocations Wikipedia English collection: http://tinyurl.com/nrc4or5

FLAX training videos: https://www.youtube.com/user/FlaxLanguageLearning/featured

Published by Research-publishing.net, not-for-profit association
Dublin, Ireland; Voillans, France, info@research-publishing.net

10 years of the LLAS elearning symposium: case studies in good practice
Edited by Kate Borthwick, Erika Corradini, & Alison Dickens

Typeset by Research-publishing.net
Cover design and frog picture by Raphaël Savina
Illustration of the retro-themed birthday greetings (id# 129712892) by "© Hermin/www.shutterstock.com"

ISBN13: 978-1-908416-22-3 (Paperback - Print on demand, black and white)
Print on demand technology is a high-quality, innovative and ecological printing method; with which the book is never 'out of stock' or 'out of print'.

ISBN13: 978-1-908416-23-0 (Ebook, PDF, colour)
ISBN13: 978-1-908416-24-7 (Ebook, EPUB, colour)

Legal deposit, Ireland: The National Library of Ireland, The Library of Trinity College, The Library of the University of Limerick, The Library of Dublin City University, The Library of NUI Cork, The Library of NUI Maynooth, The Library of University College Dublin, The Library of NUI Galway.

Legal deposit, United Kingdom: The British Library.
British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: janvier 2015.