



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

© 2014 by Research-publishing.net
Each author retains their own copyright

CALL Design: Principles and Practice
Proceedings of the 2014 EUROCALL Conference, Groningen, The Netherlands
Edited by Sake Jager, Linda Bradley, Estelle J. Meima, and Sylvie Thouèsny

Rights: All articles in this collection are published under the Attribution-NonCommercial -NoDerivatives 4.0 International (CC BY-NC-ND 4.0) licence. Under this licence, the contents are freely available online (as PDF files) for anybody to read, download, copy, and redistribute provided that the author(s), editorial team, and publisher are properly cited. Commercial use and derivative works are, however, not permitted.



Disclaimer: Research-publishing.net does not take any responsibility for the content of the pages written by the authors of this book. The authors have recognised that the work described was not published before, or that it is not under consideration for publication elsewhere. While the information in this book are believed to be true and accurate on the date of its going to press, neither the editorial team, nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, expressed or implied, with respect to the material contained herein. While Research-publishing.net is committed to publishing works of integrity, the words are the authors' alone.

Trademark notice: product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Copyrighted material: every effort has been made by the editorial team to trace copyright holders and to obtain their permission for the use of copyrighted material in this book. In the event of errors or omissions, please notify the publisher of any corrections that will need to be incorporated in future editions of this book.

Typeset by Research-publishing.net
Cover design: © Raphaël Savina (raphael@savina.net)
Fonts used are licensed under a SIL Open Font License

ISBN13: 978-1-908416-19-3 (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: 978-1-908416-20-9 (Ebook, PDF, colour)
ISBN13: 978-1-908416-21-6 (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: décembre 2014.

Investigating an open methodology for designing domain-specific language collections

Alannah Fitzgerald¹, Shaoqun Wu², and Martin Barge³

Abstract. With this research and design paper, we are proposing that Open Educational Resources (OERs) and Open Access (OA) publications give increasing access to high quality online educational and research content for the development of powerful domain-specific language collections that can be further enhanced linguistically with the Flexible Language Acquisition System (FLAX, <http://flax.nzdl.org>). FLAX uses the Greenstone digital library system, which is a widely used open-source software that enables end users to build collections of documents and metadata directly onto the Web (Witten, Bainbridge, & Nichols, 2010). FLAX offers a powerful suite of interactive text-mining tools, using Natural Language Processing and Artificial Intelligence designs, to enable novice collections builders to link selected language content to large pre-processed linguistic databases. An open methodology trialed at Queen Mary University of London in collaboration with the OER Research Hub at the UK Open University demonstrates how applying open corpus-based designs and technologies can enhance open educational practices among language teachers and subject academics for the preparation and delivery of courses in English for Specific Academic Purposes (ESAP).

Keywords: corpus-based language learning, ESAP, OER, open access, user interface design, teacher education, British law reports corpus, MOOC.

-
1. Concordia University; alannahfitzgerald@gmail.com.
 2. University of Waikato; shaoqunyw@gmail.com.
 3. Queen Mary University of London; m.i.barge@qmul.ac.uk.

How to cite this article: 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. Thoušny (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

1. Introduction

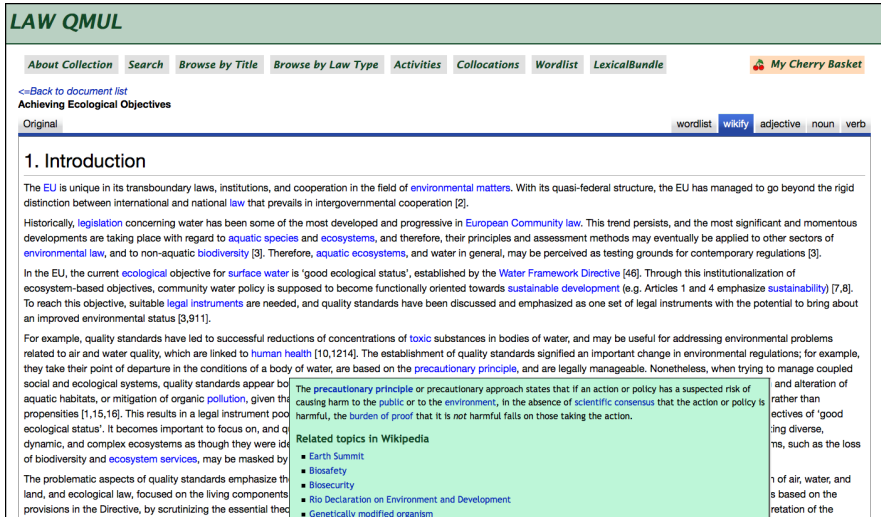
More so than ever, we have increasing access to a range of authentic open content online, such as lectures and podcasts, e-books/textbooks, research publications, blogs, wikis, as well as free and open online tools for their linguistic analyses. Designing easy-to-use interfaces for the use of these linguistic tools is a key requirement for their uptake by non-expert users, namely learners, teachers, subject academics, instructional designers and language resource developers. The Open Educational Resources and Open Access movements within higher education provide a compelling opportunity for the development of derivative domain-specific language learning resources. The field of Computer Assisted Language Learning (CALL) is now presented with a large supply of interesting linguistic material relevant to specific subject areas, including text, supplementary images (slides), audio and video. Such material can be automatically analysed, enriched, and transformed into corpus-based resources that learners can browse and query in order to extend their ability to understand the language used, and help them to express themselves more fluently and eloquently in target subject domains.

Uses for domain-specific corpora in language learning and teaching are increasing in popularity (Gabrielatos, 2005; Stubbs & Barth, 2003). Salient lexico-grammatical patterns are easily identified and retrieved by corpus tools when corpora are derived from genres and certain types of document that predominate in domain-specific areas.

Many studies have been conducted into the perceived usefulness of corpora and concordancers for the search, analysis, retrieval and transfer of language items in language learning. Usability studies on the design and presentation of linguistic data by concordancers and corpus-based systems for uptake by language learners have not yet featured prominently in the research literature into CALL, however.

Collections in FLAX use an automated scheme that extracts recurrent grammatical patterns and phrases from text and presents them in an augmented text interface, designed for the non-expert corpus user (Wu & Witten, Forthcoming). Rather than relying on complex search commands to query corpora within involved concordancer interfaces (which have been designed by and for the corpus linguist), FLAX links relevant tools and resources into streamlined online interfaces for the language learner. For example, in the ESAP collections, FLAX connects to the open-source Wikipedia Miner toolkit to extract key concepts and their definitions from Wikipedia articles (Milne & Witten, 2013) to assist with reading and vocabulary as can be seen in Figure 1.

Figure 1. FLAX augmented text interface with wikify function in Law Collections



2. Method

2.1. Open domain-specific collections building in FLAX

“Use of OER leads to critical reflection by educators, with [...] improvement in their practice” (OER Research Hub, n.d., para. 1). This is one of a cluster of research hypotheses currently under investigation at the OER Research Hub for the development of open language corpora in FLAX in collaboration with Queen Mary University of London (QMUL).

Table 1. Type, number and source of items in the FLAX Law Collections

Type of media	Number and source of corpus items
Open Access Law research articles	40 Articles (DOAJ - Directory of OA Journals, with Creative Commons for the development of derivatives)
MOOC lecture transcripts/videos (streamed via YouTube/Vimeo)	4 MOOC Collections: Copyright Law (Harvard/edX), English Common Law (University of London/Coursera), Age of Globalization (Texas at Austin/edX), Environmental Law and Politics (OpenYale)
Podcast audio files/transcripts (OpenSpire)	10-15 Lectures (Oxford Law Faculty and the Centre for Socio-Legal Studies)
PhD Law thesis writing	50-70 EThoS Theses (sections: abstracts, introductions, conclusions) at the British Library (OA but not licensed Creative Commons – permissions granted by HEIs)
British Law Reports Corpus (BLaRC)	8.5 million-word corpus developed by María José Marín Pérez. Derived from free legal sources at the British and Irish Legal Information Institute (BAILII) aggregation website

Domain-specific law collections in FLAX were developed for ESAP students taking the Law Pathway on the summer pre-sessional and the Critical Thinking and Writing in Law In-Sessional programmes at QMUL. The law collections in FLAX are centred on the re-use of OER and OA research publications in the target domain of Law, as can be seen in Table 1. It is anticipated that these collections for legal English will be of use across both formal and informal language learning and translation contexts.

2.2. Formatting resources for use in FLAX

Text extracts of longer than 2-3,000 words are likely to halt or crash the FLAX server application, due to the quantity of text parsing that the FLAX server can efficiently process in a given time. Therefore, source texts have to be divided into sections of not more than 2-3,000 words in length.

Source articles are often downloadable in .pdf format, and are often accessible as full web documents. However, text extracts intended for upload to the FLAX website need to be marked up in HTML. Even with knowledge of HTML, the process of marking up each text extract is a time-consuming process. It was therefore decided to develop a web-based formatting tool, implemented using JavaScript, as can be seen in Figure 2 to ease the process of converting sections of text to HTML.

Figure 2. FLAX HTML resource formatting tool

The user can paste copied text into a main text field, and paste/type the article title and section headings into labelled boxes. HTML tag buttons enable the user to

insert tags at relevant points in the text in order to re-format as required. When the file is exported, using the ‘Export’ button, the tool generates the HTML file, using the text input by the user. The tool is still in early stages of development and can only handle basic text formatting functions. However, further iterations of the tool are planned (e.g. the inclusion of colour-coded tags for enhanced user readability; the ability to insert image links).

3. Discussion

3.1. Learning collocations in FLAX

Among other aspects of language, the ESAP for law collections in FLAX provide an excellent context in which to study collocations, a notoriously challenging aspect of English productive use even for quite advanced learners (Bishop, 2004; Nesselhauf, 2003).

Figure 3. Collocations in Law QMUL Collections linked to FLAX Wikipedia collocations database

153 collocation(s) associated with the word environmental

Adjective (80) | Noun + of (38) | Verb (35)

- [environmental law](#) (44)
- [environmental quality](#) (37)
- [environmental science](#) (10)
- [environmental impact statement](#) (9)
- [environmental tobacco smoke](#) (8)
- [environmental effects](#) (6)

- Now, I'd like to close by having you think about damages and how we calculate damages. And as you think about every case that we've discussed in this term, when we think about [environmental effects](#), you should be thinking about the magnitude of the effect and you should be thinking about the distribution of the effect. So think about this comparison between nuclear power on the one hand and wind energy on the other.
- And what we'll see today is that knowledge about the [environmental effects](#) of national security are produced predominately by the Defense Department. They control sites, they control the technology in weaponry and weapon delivery systems, making it extremely difficult for the public to understand really what the dangers might be.
- And the Environmental Protection Agency basically shuts down the public from understanding what that is. And it's all about how much of the chemical is produced, what the company knows about where it goes, and also what the health effects or [environmental effects](#) might be. So access to data, intellectual property rights, secrecy and confidentiality, these are all things that we really need to think about very carefully when we structure environmental law.
- So we could get a whole new level of understanding of energy and the environmental health of different products, including issues such as protein content, fat content, type of fat, amount of salt, et cetera. Right now, most of us walk through the marketplace really blind to these issues, blind to the [environmental effects](#) of food production, blind also to the energy consumption.

Learning Collocations My Cherry Basket Act

environmental In (contemporary English Wikipedia)

Family words: [environment](#) [environmentalist](#) [environmentalists](#) [environmentally](#) [environments](#) Synonym

used as an adjective

Collocation	Count	Collocation	Count
environmental + noun		environmental issues	2088
		environmental concerns	684
		environmental protection	1146
		environmental concerns	684
		environmental impact	882
		environmental concern	88
		environmental factors	830
		major environmental concern	10
		environmental conditions	794
		local environmental concerns	8
		major environmental concerns	4
adjective + environmental		social and environmental	79
		environmental and social	33
		various environmental	31
		local environmental	27
		significant environmental	26

- One of the **major environmental concerns** facing the park is illegal fishing of Chamo.
- The hills are mostly forested, although deforestation and the formation of fire-**major environmental concerns** in Trinidad.
- Though the route was basically set in stone in Connecticut, many issues remain biggest of which were **major environmental concerns** about how the freeway Reservoir, which is the main drinking water supply for Providence.
- In order to bring global concerns to the forefront of public attention, Pyle has w collaborative relationships between **major environmental concerns** such as

Figure 3 shows the result for the word *environmental*, which returns 153 collocations in the OpenYale lectures. Collocations are grouped under tabs that reflect the syntactic roles of the associated word or words: adjective (shown), noun + *of*, verb. The underlined words, *environmental* and *effects*, are hyperlinked to entries for those words in an external collocations database⁴ built from a Wikipedia-derived corpus of 200 million articles. For example, clicking on the link for *environmental* generates a further collocations popup that lists *environmental issues*, *environmental protection*, etc., along with their frequency and their context in this much larger corpus.

3.2. Lexical bundles, word lists and natural language processing in FLAX

FLAX identifies “lexical bundles” used in the target ESAP law collections, which are multi-word sequences with distinctive syntactic patterns and discourse functions found in academic prose and lectures (Biber & Barbieri, 2007; Biber, Conrad, & Cortes, 2003, 2004). A typical pattern found in spoken corpora is *verb phrase + that (wanted to reemphasise/mention that...)*.

Figure 4. FLAX open natural language processing of verb phrases in Law QMUL Collections

The screenshot displays the FLAX interface for 'LAW Lecture Transcripts'. It features a navigation bar with tabs for 'About Collection', 'Search', 'Browse by Title', 'Activities', 'Collocations', 'Wordlist', and 'LexicalBundles'. A video player is embedded, showing a lecture with a slide titled 'Warning Labels for Ingredients in Human Breast Milk'. Below the video, the text of the transcript is shown, with several phrases highlighted in blue to indicate they are part of a lexical bundle. These highlighted phrases include: 'raised the alarm', 'caused the population to be quite upset', 'found pesticides in human breast milk', 'get into human breast milk', 'increased the sense of susceptibility', 'deeply embedded in the war in Vietnam', 'adding these words: caution, warning, and hazard, depending upon the relative toxicity', 'given authority finally to remove pesticides', and 'given the responsibility to manage pesticides'. A 'wordlist' tab is visible at the bottom right of the text area.

User-friendly interfaces have been developed in FLAX to enable learners to analyse collection documents against well-known word lists such as Coxhead’s (2000) Academic Word List and West’s (1953) General Service List. Topic-specific

4. The database is available at <http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=collocations>

words are also extracted from the documents to highlight recurrent vocabulary and a keyword slider tool function has been designed to identify the keyness and frequency of certain lexical items as they occur in specific texts. Keyness refers to the frequency of words as they occur in specific documents as a text feature rather than in relationship to other words as a language feature in the case of collocations, for example. The FLAX system also uses Open Natural Language Processing for the syntactic tagging⁵ of texts, as can be seen in Figure 4 with verb phrases from one of the environmental law lectures.

4. Conclusions

Content varies in terms of licensing restrictions, and FLAX has been designed to offer flexible linguistic support options for enhancing such content across both open and closed platforms. While we anticipate that this open methodology for domain-specific collections building in FLAX will be of value to language communities across formal and informal education, usage studies will be conducted at QMUL to suggest further directions for development.

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, Concordia University in Canada, and Queen Mary University of London in the United Kingdom. We are particularly grateful to Dr. Saima Sherazi and William Tweddle at the Queen Mary Language Centre for their contributions with the planning of this project collaboration and with Dr. Sherazi's presentation of the work at the EuroCALL 2014 conference event.

References

- Biber, D., & Barbieri F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purpose*, 26(3), 263-286. doi:10.1016/j.esp.2006.08.003
- Biber, D., Conrad, S., & Cortes, V. (2003). Lexical bundles in speech and writing: An initial taxonomy. In A. Wilson, P. Rayson, & T. McEnery (Eds), *Corpus linguistics by the lune: A festschrift for Geoffrey Leech* (pp. 71-92). Frankfurt/Main: Peter Lang.
- Biber, D., Conrad, S., & Cortes, V. (2004). If you look at ... : lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371-405. doi:10.1093/applin/25.3.371

5. See the OpenNLP system; <http://opennlp.sourceforge.net>

- Bishop, H. (2004). The effect of typographic 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, USA: John Benjamins. doi:10.1075/Illt.9.12bis
- 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
- Gabrielatos, C. (2005). Corpora and language teaching: Just a fling or wedding bells? *Teaching English as a Second or Foreign Language*, 8(4). Retrieved from <http://tesl-ej.org/ej32/a1.html>
- Milne, D., & Witten, I. H. (2013). An open-source toolkit for mining Wikipedia. *Artificial Intelligence*, 194, 222-239. doi:10.1016/j.artint.2012.06.007
- Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, 24(2), 223-242. doi:10.1093/applin/24.2.223
- OER Research Hub. (n.d.). *Hypothesis E – Reflection* [website]. Retrieved from <http://oerresearchhub.org/collaborative-research/hypotheses/hypothesis-e-reflection/>
- Stubbs, M., & Barth, I. (2003). Using recurrent phrases as text-type discriminators. *Functions of Language*, 10(1), 61-104. doi:10.1075/fol.10.1.04stu
- West, M. (1953). *A general service list of English words*. London: Longman, Green & Co.
- Witten, I. H., Bainbridge, D., & Nichols, D. M. (2010). *How to build a digital library* (2nd ed.). Burlington, MA: Morgan Kaufmann.
- Wu, S., & Witten, I. H. (Forthcoming). Transcending concordance: Augmenting academic text for L2 writing. Submitted to the *Journal of English for Academic Purposes*.