Investigating an open methodology for designing domain-specific language collections

Alannah Fitzgerald[1], Shaoqun Wu[2], and Martin Barge[3]

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. 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

1. Introduction

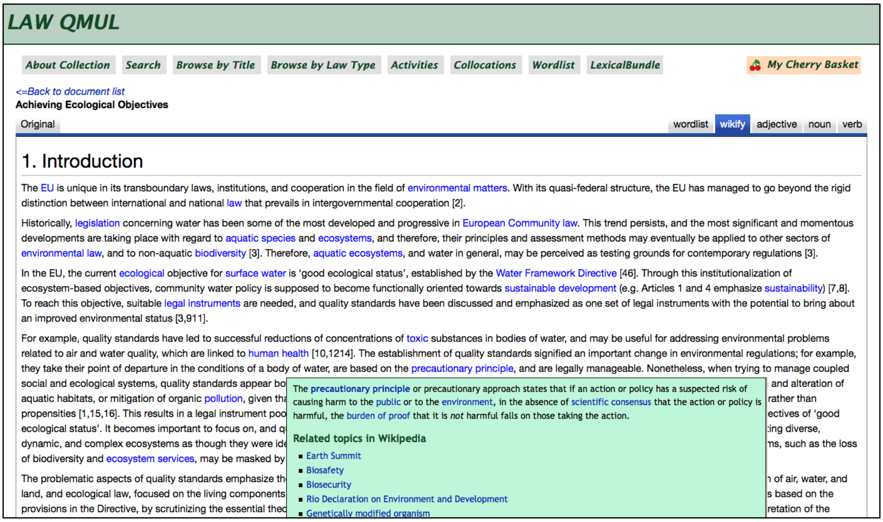
More so than ever, we have increasing access to a range of authentic open content online, such as lectures and podcasts, e-books/textbooks, research publications, blogs, wikis, as well as free and open online tools for their linguistic analyses. Designing easy-to-use interfaces for the use of these linguistic 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



Achieving Ecological Objectives

1. Introduction

The EU is unique in its transboundary laws, institutions, and cooperation in the field of environmental matters. With its quasi-federal structure, the EU has managed to go beyond the rigid distinction between international and national law that prevails in intergovernmental cooperation [2].

Historically, legislation concerning water has been some of the most developed and progressive in European Community law. This trend persists, and the most significant and momentous developments are taking place with regard to aquatic species and ecosystems, and therefore, their principles and assessment methods may eventually be applied to other sectors of environmental law, and to non-aquatic biodiversity [3]. Therefore, aquatic ecosystems, and water in general, may be perceived as testing grounds for contemporary regulations [3].

In the EU, the current ecological objective for surface water is 'good ecological status', established by the Water Framework Directive [46]. Through this institutionalization of ecosystem-based objectives, community water policy is supposed to become functionally oriented towards sustainable development (e.g. Articles 1 and 4 emphasize sustainability) [7,8]. To reach this objective, suitable legal instruments are needed, and quality standards have been discussed and emphasized as one set of legal instruments with the potential to bring about an improved environmental status [3,911].

For example, quality standards have led to successful reductions of concentrations of toxic substances in bodies of water, and may be useful for addressing environmental problems related to air and water quality, which are linked to human health [10,1214]. The establishment of quality standards signified an important change in environmental regulations; for example, they take their point of departure in the conditions of a body of water, are based on the precautionary principle, and are legally manageable. Nonetheless, when trying to manage coupled
social and ecological systems, quality standards appear bo … and alteration of
aquatic habitats, or mitigation of organic pollution, given tha … rather than
propensities [1,15,16]. This results in a legal instrument poo … ectives of 'good
ecological status'. It becomes important to focus on, and q … ing diverse,
dynamic, and complex ecosystems as though they were ide … ns, such as the loss
of biodiversity and ecosystem services, may be masked by

The problematic aspects of quality standards emphasize th … of air, water, and
land, and ecological law, focused on the living components … s based on the
provisions in the Directive, by scrutinizing the essential theo … retation of the

The precautionary principle or precautionary approach states that if an action or policy has a suspected risk of causing harm to the public or to the environment, in the absence of scientific consensus that the action or policy is harmful, the burden of proof that it is *not* harmful falls on those taking the action.

Related topics in Wikipedia

- Earth Summit
- Biosafety
- Biosecurity
- Rio Declaration on Environment and Development
- Genetically modified organism

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 (OpenSpires)	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

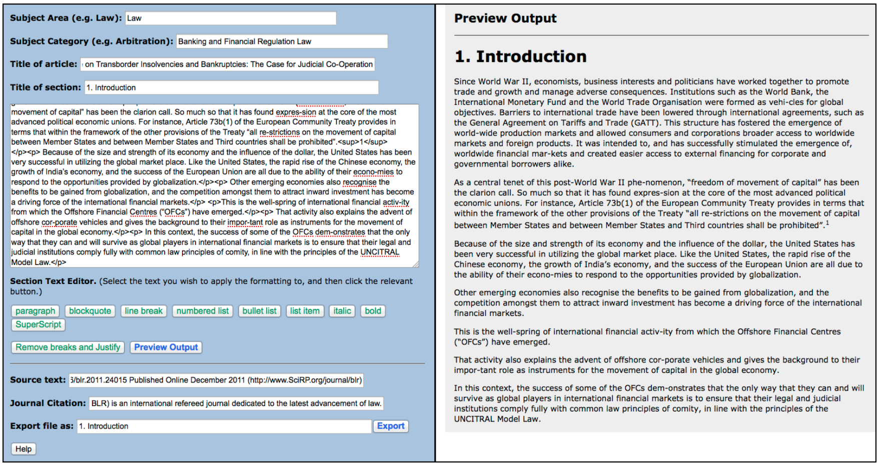
Domain-specific law collections in FLAX were developed for ESAP students taking the Law Pathway on the summer pre-sessional and the Critical Thinking and Writing in Law In-Sessional programmes at QMUL. The law collections in FLAX are centred on the re-use of OER and OA research publications in the target 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



Subject Area (e.g. Law): Law

Subject Category (e.g. Arbitration): Banking and Financial Regulation Law

Title of article: on Transborder Insolvencies and Bankruptcies: The Case for Judicial Co-Operation

Title of section: 1. Introduction

movement of capital" has been the clarion call. So much so that it has found expres-sion at the core of the most advanced political economic unions. For instance, Article 73b(1) of the European Community Treaty provides in terms that within the framework of the other provisions of the Treaty "all re-strictions on the movement of capital between Member States and between Member States and Third countries shall be prohibited".¹</p><p> Because of the size and strength of its economy and the influence of the dollar, the United States has been very successful in utilizing the global market place. Like the United States, the rapid rise of the Chinese economy, the growth of India's economy, and the success of the European Union are all due to the ability of their econo-mies to respond to the opportunities provided by globalization.</p><p> Other emerging economies also recognise the benefits to be gained from globalization, and the competition amongst them to attract inward investment has become a driving force of the international financial markets.</p> <p>This is the well-spring of international financial activ-ity from which the Offshore Financial Centres ("OFCs") have emerged.</p><p> That activity also explains the advent of offshore cor-porate vehicles and gives the background to their impor-tant role as instruments for the movement of capital in the global economy.</p><p> In this context, the success of some of the OFCs dem-onstrates that the only way that they can and will survive as global players in international financial markets is to ensure that their legal and judicial institutions comply fully with common law principles of comity, in line with the principles of the UNCITRAL Model Law.</p>

Section Text Editor. (Select the text you wish to apply the formatting to, and then click the relevant button.)

Source text: 5/blr.2011.24015 Published Online December 2011 (http://www.SciRP.org/journal/blr)

Journal Citation: BLR) is an international refereed journal dedicated to the latest advancement of law.

Export file as: 1. Introduction

Preview Output

1. Introduction

Since World War II, economists, business interests and politicians have worked together to promote trade and growth and manage adverse consequences. Institutions such as the World Bank, the International Monetary Fund and the World Trade Organisation were formed as vehi-cles for global objectives. Barriers to international trade have been lowered through international agreements, such as the General Agreement on Tariffs and Trade (GATT). This structure has fostered the emergence of world-wide production markets and allowed consumers and corporations broader access to worldwide markets and foreign products. It was intended to, and has successfully stimulated the emergence of, worldwide financial mar-kets and created easier access to external financing for corporate and governmental borrowers alike.

As a central tenet of this post-World War II phe-nomenon, "freedom of movement of capital" has been the clarion call. So much so that it has found expres-sion at the core of the most advanced political economic unions. For instance, Article 73b(1) of the European Community Treaty provides in terms that within the framework of the other provisions of the Treaty "all re-strictions on the movement of capital between Member States and between Member States and Third countries shall be prohibited".[1]

Because of the size and strength of its economy and the influence of the dollar, the United States has been very successful in utilizing the global market place. Like the United States, the rapid rise of the Chinese economy, the growth of India's economy, and the success of the European Union are all due to the ability of their econo-mies to respond to the opportunities provided by globalization.

Other emerging economies also recognise the benefits to be gained from globalization, and the competition amongst them to attract inward investment has become a driving force of the international financial markets.

This is the well-spring of international financial activ-ity from which the Offshore Financial Centres ("OFCs") have emerged.

That activity also explains the advent of offshore cor-porate vehicles and gives the background to their impor-tant role as instruments for the movement of capital in the global economy.

In this context, the success of some of the OFCs dem-onstrates that the only way that they can and will survive as global players in international financial markets is to ensure that their legal and judicial institutions comply fully with common law principles of comity, in line with the principles of the UNCITRAL Model Law.

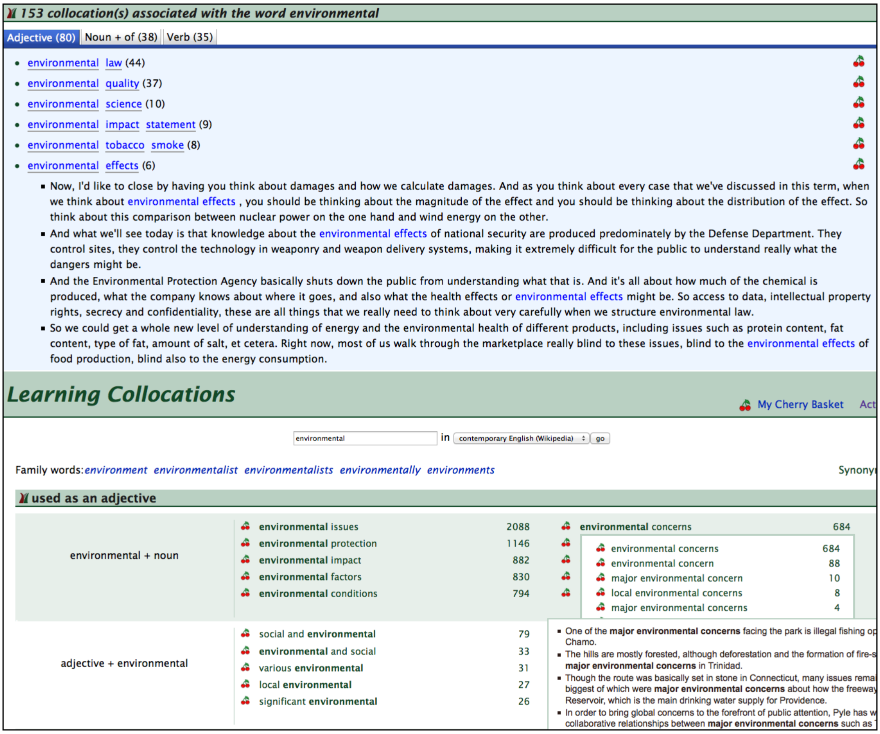
The user can paste copied text into a main text field, and paste/type the article title and section headings into labelled boxes. HTML tag buttons enable the user to insert tags at relevant points in the text in order to re-format as required. When the file is exported, using the ‘Export’ button, the 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

- 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

Family words: *environment environmentalist environmentalists environmentally environments*

used as an adjective

environmental + noun	**environmental** issues	2088	**environmental** concerns	684
	environmental protection	1146		
	environmental impact	882		
	environmental factors	830		
	environmental conditions	794		
adjective + environmental	social and **environmental**	79		
	environmental and social	33		
	various **environmental**	31		
	local **environmental**	27		
	significant **environmental**	26		

environmental concerns	684
environmental concern	88
major environmental concern	10
local environmental concerns	8
major environmental concerns	4

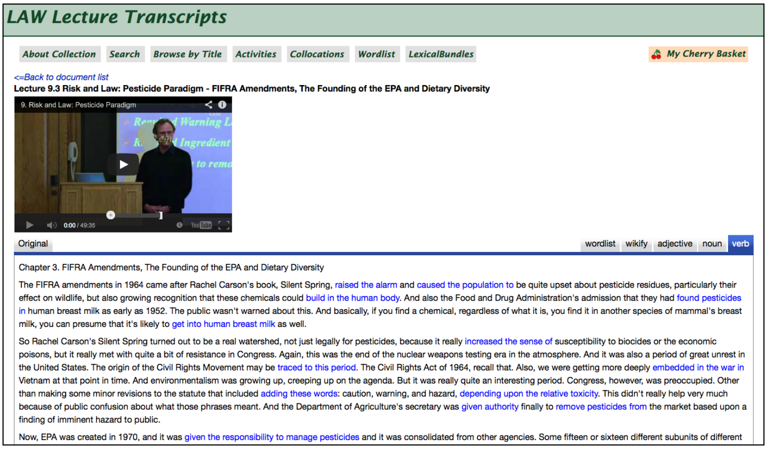
- One of the **major environmental concerns** facing the park is illegal fishing op Chamo.
- The hills are mostly forested, although deforestation and the formation of fire-s **major environmental concerns** in Trinidad.
- Though the route was basically set in stone in Connecticut, many issues remai 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 database[4] 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



Lecture 9.3 Risk and Law: Pesticide Paradigm - FIFRA Amendments, The Founding of the EPA and Dietary Diversity

Chapter 3. FIFRA Amendments, The Founding of the EPA and Dietary Diversity

The FIFRA amendments in 1964 came after Rachel Carson's book, Silent Spring, raised the alarm and caused the population to be quite upset about pesticide residues, particularly their effect on wildlife, but also growing recognition that these chemicals could build in the human body. And also the Food and Drug Administration's admission that they had found pesticides in human breast milk as early as 1952. The public wasn't warned about this. And basically, if you find a chemical, regardless of what it is, you find it in another species of mammal's breast milk, you can presume that it's likely to get into human breast milk as well.

So Rachel Carson's Silent Spring turned out to be a real watershed, not just legally for pesticides, because it really increased the sense of susceptibility to biocides or the economic poisons, but it really met with quite a bit of resistance in Congress. Again, this was the end of the nuclear weapons testing era in the atmosphere. And it was also a period of great unrest in the United States. The origin of the Civil Rights Movement may be traced to this period. The Civil Rights Act of 1964, recall that. Also, we were getting more deeply embedded in the war in Vietnam at that point in time. And environmentalism was growing up, creeping up on the agenda. But it was really quite an interesting period. Congress, however, was preoccupied. Other than making some minor revisions to the statute that included adding these words: caution, warning, and hazard, depending upon the relative toxicity. This didn't really help very much because of public confusion about what those phrases meant. And the Department of Agriculture's secretary was given authority finally to remove pesticides from the market based upon a finding of imminent hazard to public.

Now, EPA was created in 1970, and it was given the responsibility to manage pesticides and it was consolidated from other agencies. Some fifteen or sixteen different subunits of different

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 tagging[5] 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/lllt.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*.