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.

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.
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>
<thead>
<tr>
<th>Type of media</th>
<th>Number and source of corpus items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Access Law research articles</td>
<td>40 Articles (DOAJ - Directory of OA Journals, with Creative Commons for the development of derivatives)</td>
</tr>
<tr>
<td>MOOC lecture transcripts/videos</td>
<td>4 MOOC Collections: Copyright Law (HarvardX), English Common Law (University of London/Coursera), Age of Globalization (Texas at AustinX), Environmental Law and Politics (OxfordYale)</td>
</tr>
<tr>
<td>(streamed via YouTube/Vimeo)</td>
<td></td>
</tr>
<tr>
<td>Podcast audio files/transcripts</td>
<td>10-15 Lectures (Oxford Law Faculty and the Centre for Socio-Legal Studies)</td>
</tr>
<tr>
<td>(OpenSpires)</td>
<td></td>
</tr>
<tr>
<td>PhD Law thesis writing</td>
<td>50-70 EThos Theses (sections: abstracts, introductions, conclusions) at the British Library (OA but not licensed Creative Commons – permissions granted by HEIs)</td>
</tr>
<tr>
<td>British Law Reports Corpus (BLaRC)</td>
<td>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</td>
</tr>
</tbody>
</table>
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
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 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

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

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


---

5. See the OpenNLP system; [http://opennlp.sourceforge.net](http://opennlp.sourceforge.net)


95