Flexible Open Language Education for a Multilingual World

Alannah Fitzgerald, Shaoqun Wu and Ian H. Witten

Department of Education, Concordia University, Montreal, Quebec.
Canada Computer Science Department, Waikato University, Hamilton, New Zealand {shaoqun, ihw}@cs.waikato.ac.nz, fitzgerald@education.concordia.ca

This research and technology paper will present open language tools and collections that have been developed for supporting domain-specific academic language with the FLAX multilingual open source software. OpenCourseWare (OCW), Massive Open Online Courses (MOOC) and Open Educational Resources (OER) are becoming popular educational vehicles through which well-resourced universities and organisations can reach out to non-traditional audiences, including those from other countries and cultures. For example, the OCW Consortium website states that, “Open Education seeks to scale educational opportunities by taking advantage of the power of the internet, allowing rapid and essentially free dissemination, and enabling people around the world to access knowledge, connect and collaborate” (“About the OCWC,” n.d.).

Specificity in Academic Language

Open education provides a compelling opportunity for domain-specific academic language learning. Online courses supply a large corpus of interesting linguistic material relevant to a particular area, including supplementary images (slides), audio and video. We contend that this corpus can be automatically analysed, enriched, and transformed into a resource that learners can browse and query in order to extend their ability to understand the language used, and help them express themselves more fluently and eloquently in that domain. To illustrate this idea, an existing online corpus-based language learning tool (FLAX) is applied to an English-medium Coursera MOOC offered by Columbia University, entitled Virology I: How Viruses Work.

MOOC participants register for educational courses; they do not sign up as language learners. However, many online learners will encounter a language barrier during their study with many of the open educational offerings being delivered in the world’s presiding lingua francas, namely English, Arabic, French, Chinese, Russian, Spanish, and Portuguese. Beyond the simple translation of lecture transcripts and course readings, learners will be strongly motivated to improve their knowledge of key terms and concepts as they are used in the subject domain, exemplified here with the Virology MOOC collections in FLAX for support with Academic Language. (They are also helpful for native speakers of the target language.)

OER Research Hypotheses for Open Language Support

Research into the development and uses of text analysis tools from corpus linguistics has been primarily carried out in relation to traditional classroom-based university teaching only. This is despite the growing number of higher education offerings in open and distance learning, including the recent surge in OER, OCW and MOOCs in collaboration with universities and educational organisations. Drawing on linguistic data sources from MOOCs, along with survey data from MOOC learners and interview and survey data from course developers and English language education professionals, this paper will present findings from participants based on their perceptions of the effectiveness of the open language tools and collections in FLAX under investigation for Academic Language support. Specific OER research hypotheses have been investigated through this research in collaboration with the OER Research Hub. The following OER hypotheses have been examined through the different data collection instruments in relation to the different research participant groups, including:

Hypothesis A: Use of OER leads to improvement in student performance and satisfaction.
Hypothesis E: Use of OER leads to critical reflection by educators, with improvement in their practice.
Hypothesis H: Informal learners adopt a variety of techniques to compensate for the lack of formal support.
Hypothesis I: Open education acts as a bridge to formal education, and is complementary, not competitive, with it.

Scaling Flexible Open Language Learning

For the purpose of innovating, building and creating multilingual learning support collections for large-scale learning, both online and offline, the flexible tools and resources in FLAX can be applied to content...
in any modern language. Within the context of MOOCs, inroads into the research and development of language support include primarily: translation and transcription of course content for supporting listening and reading comprehension; and written diagnostic assessments and peer review for supporting writing requirements. FLAX is primarily concerned with language learning support and can work in tandem with these existing technologies and approaches.

We use the Greenstone digital library system, which is widely used open source software that enables end users to build collections of documents and metadata and serve them on the Web (Witten et al., 2010). The linguistic enhancements in FLAX described below are all extensions to Greenstone. FLAX takes text documents, automatically extracts important language components—such as academic words and their usage patterns, key concepts, collocations, and lexical bundles—and presents them in ways that draws the attention of learners and gives them opportunities to encounter these components in various authentic contexts.

Augmenting Text for Language Learning

Open educational practitioner, Vincent Racaniello, of Columbia University created Virology I from lectures that were popular across a range of web channels, including iTunesU and YouTube, before being imported into the Coursera MOOC. These lectures, along with Racaniello’s weekly podcast This Week in Virology, his academic Virology blog, and Open Access articles related to his virology courses, are published under a Creative Commons Attribution licence. All of these resources were pre-processed before being built into FLAX collections. The lecture transcripts underwent simple editing, including division into subsections, and were reformatted into manageable chunks as HTML files to decrease cognitive load when listening and viewing. Scientific images and their labels from the lecturer’s PowerPoint slides were re-formatted for readability. Textual documents are searchable, and browsable by title. Videos, audios and images are embedded within the document.

Mining Wikipedia

“The online encyclopaedia Wikipedia is a vast, constantly evolving tapestry of interlinked articles. For developers and researchers it represents a giant multilingual database of concepts and semantic relations, a potential resource for natural language processing and many other research areas.” (Milne & Witten, 2013 p. 222).

FLAX connects to the Wikipedia Miner tool to extract key concepts and their definitions from Wikipedia articles. Milne and Witten (2013) describe the method used to relate words and phrases in running text to Wikipedia articles. 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, intracellular parasite, cells, organism, genome, nucleic acid, … in the article titled What is a Virus are identified as Wikipedia concepts. This definition for genome is extracted: “In modern molecular biology and genetics, the genome is the entirety of an organism’s hereditary information.”

Learning Collocations and Lexical Bundles

The importance of collocation knowledge in language learning has been widely recognized. Hill (1999) observed that second language writing tends to be cumbersome and error prone because of insufficient collocation knowledge. Studies suggest that an educated native speaker has a vocabulary of around 20,000 word families (Goulden et al., 1990). That is a large number, but still a manageable goal for the most determined and motivated of language learners. However, it pales into insignificance when compared with the total number of items—expressions, idioms, collocations—that native speakers have (Hill, 2000). Collocation knowledge is difficult to acquire simply because there is so much of it. Native speakers carry hundreds of thousands—possibly millions—of lexical chunks in their heads, ready to draw upon in order to produce fluent, accurate and meaningful language (Lewis, 1997). This presents a daunting challenge to language learners.

General and collocations dictionaries offer at best minimal examples of common collocations in a target language, as they are restricted in space as to the examples that can be published for a general audience. Domain-specific collocations are therefore not readily available in commercially published language
reference resources and that is why systems like FLAX offer massive, high-quality resources that help students build up collocation knowledge within specific domain areas. We focus on lexical collocations with noun-based structures *verb + noun, noun + noun, adjective + noun, and noun + of + noun*, because they are the most salient and important patterns in topic-specific text. The system first assigns part-of-speech tags to words in the text and then extracts word combinations that match syntactic patterns. These extracted collocations are grouped by pattern and sorted by frequency.

“Lexical bundles” are multi-word sequences with distinctive syntactic patterns and discourse functions that are commonly used in academic prose (Biber & Barbieri, 2007; Biber et al, 2003, 2004). Typical patterns include *noun phrase + of, prepositional phrase + of, it + verb/adjective phrase, be + noun/adjective phrase,* and *verb phrase + that*. Such phrases fulfill discourse functions such as referential expression (framing, quantifying and place/time/text-deictic), stance indication (epistemic, directive, ability) and discourse organization (topic introduction and elaboration).

**Language activities**

FLAX provides a series of language activities, accessed through the *Activities* button, that focus on words, collocation, sentence or article structures and concepts related to the topics. Each activity has a teacher’s interface and a student interface. In the former, language teachers and instructional designers developing MOOC support can select parameters for exercise creation, and provide hints for students. The exercises are generated automatically, and can be reviewed and modified to discard undesirable language choices before presenting them to learners.

There are many activity types. One example, *Cloze* (“fill-in-the-blanks”) activities are widely used to test knowledge of vocabulary and syntax, as well as reading comprehension. Words are removed from an article and learners must re-insert them. The target words can be content words such as nouns, verbs, adjectives and adverbs; or function words such as prepositions, pronouns, conjunctions and auxiliaries; or Wikipedia concepts that have been identified automatically as sketched above. To create a Cloze activity one selects an article and then decides whether the system should omit words based on a specified gap size, or specified parts of speech, or Wikipedia concepts. Images, audio and video that accompany an article can be added into the exercise at the instructor’s discretion.

**Discussion**

The MOOC language collections we have built demonstrate the affordances of the FLAX software. FLAX is open source and can be downloaded to build language support collections in any language with text-based content and supporting audio-visual material, for both online and classroom use. It is designed so that non-expert developers—whether language teachers, subject specialists, or instructional design and e-learning support teams—can build their own collections. Educational and research content varies in terms of licensing restrictions, depending on the publishing policies and strategies adopted by institutions for their content. FLAX has been designed to offer a flexible suite of linguistic support options for enhancing such content across both open and closed platforms.

This research proposes that the practical contribution of the FLAX tools and language collections, which are openly available for download on the FLAX website and promoted through open channels, will benefit current practice in Academic Language support for both formal and informal open education, as demonstrated here with MOOCs. Open online systems and resources like the collections in FLAX proposed by this research have unique characteristics and challenges with regards to diffusion, adoption and integration as compared with commercial initiatives by well-known language course book publishers which still dominate the traditional language learning resources market. Data is presented regarding participants’ perceptions and interactions with these open educational language systems, as they exist and as we are designing them in this research. In addition to this, a deeper understanding of how to design, iterate, integrate, evaluate and scale open technological systems in support of advanced approaches to language learning and instruction within the specific context of open educational resource initiatives is presented through this research and technology track paper for discussion with OCWC Global 2014 participants.

**References**