Flexible Open Language Education for a Multilingual World

Alannah Fitzgerald, Department of Education, Concordia University, Montreal, Canada &
The Open Educational Resources Research Hub, Institute of Educational Technology,
The Open University, Milton Keynes, United Kingdom
fitzgerald@education.concordia.ca
Shaoqun Wu and Ian Witten, Computer Science Department, Greenstone Digital Library Lab,
The University of Waikato, Hamilton, New Zealand
shaoqun@cs.waikato.ac.nz and ihw@cs.waikato.ac.nz

Abstract

More so than ever we have increasing access to a range of authentic open content online (lectures and podcasts, e-books, e-textbooks, Open Access research publications, blogs, wikis etcetera) and 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. This research track paper at the OCWC Global Conference will present open language tools and collections that have been developed for supporting domain-specific academic language acquisition with the FLAX multilingual Open Source Software (OSS). OpenCourseWare (OCW), Massive Open Online Courses (MOOCs) and Open Educational Resources (OERs) are becoming popular educational vehicles through which well-resourced universities and organisations can reach out to non-traditional audiences, including those from other cultures and language groups. OCW and MOOC participants register for specific 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 official languages or lingua francas of which English is the most dominant globally. Beyond the simple translation of lecture transcripts and course readings, both native-speakers and non-native speakers alike will be strongly motivated to improve their knowledge and usage of key academic terms and concepts as they are used in the language of instruction for a specific subject domain.

We contend that the language challenge that often accompanies learning an academic subject in another language also presents a compelling opportunity for domain-specific language learning that remixes available educational and research content. This content supplies a large corpus of interesting linguistic material relevant to a particular subject domain area, including text, supplementary images (slides), audio and video. We argue 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. (It is also helpful for native speakers of the language of instruction.) To illustrate this idea, an existing online corpus-based language-learning tool (FLAX) is applied to two English-medium Coursera MOOCs, titled *Virology 1: How Viruses Work* and *Virology 2: How Viruses Cause Disease*, offered by Columbia University. A further investigation into an open methodology for developing linguistic support across both formal and informal education is also being trialed at Queen Mary University of London in collaboration with the OER Research Hub at the UK Open University. In this paper, we will also discuss how applying open corpus-based designs and technologies can enhance open educational practices among those working in both formal and informal education, for the preparation and delivery of English for Specific Academic Purposes (ESAP).

Keywords
Domain-Specific Language; Corpus-based Language Learning, FLAX Language, Open Educational Resources, MOOCs, OpenCourseWare, English for Specific Academic Purposes

Scaling Flexible Open Language Learning

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 OERs, OCW and MOOCs in collaboration with universities and educational organisations. For the purpose of innovating, building and creating multilingual learning support collections for large-scale language learning, both online and offline, the flexible tools and resources in FLAX can augment content in any modern language. Within the context of MOOCs and OCW, some inroads have already been made into the research and development of linguistic support, including: translation and transcription technologies for supporting listening and reading comprehension of course lectures and readings; automated text analysis tools for performing language diagnostics for lexico-grammatical features in learners' written answers to quizzes and exams; and human peer review for supporting collaborative writing in extended written assessments. FLAX is primarily concerned with developing interactive language learning support at the lexico-grammatical level and can work in tandem with these existing technologies and approaches for developing powerful open linguistic support.

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 (Wu & Witten, 2013). 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 way 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 and Earth's virology professor, Vincent Racaniello, of Columbia University created *Virology I* and *II* from lectures that were popular across a range of web channels, including iTunesU and YouTube, before being imported into the Coursera MOOC.

Figure 1. Open Educational Practitioner and Earth's Virology Professor, Vincent Racaniello. Source: Vincent Racaniello Virology YouTube channel. https://www.youtube.com/channel/UCyFgCoP4ovsHbt92vM4zN2A

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 license (CC-BY). 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.

Table 1. Number of Items in the FLAX Language Virology MOOC Collection.

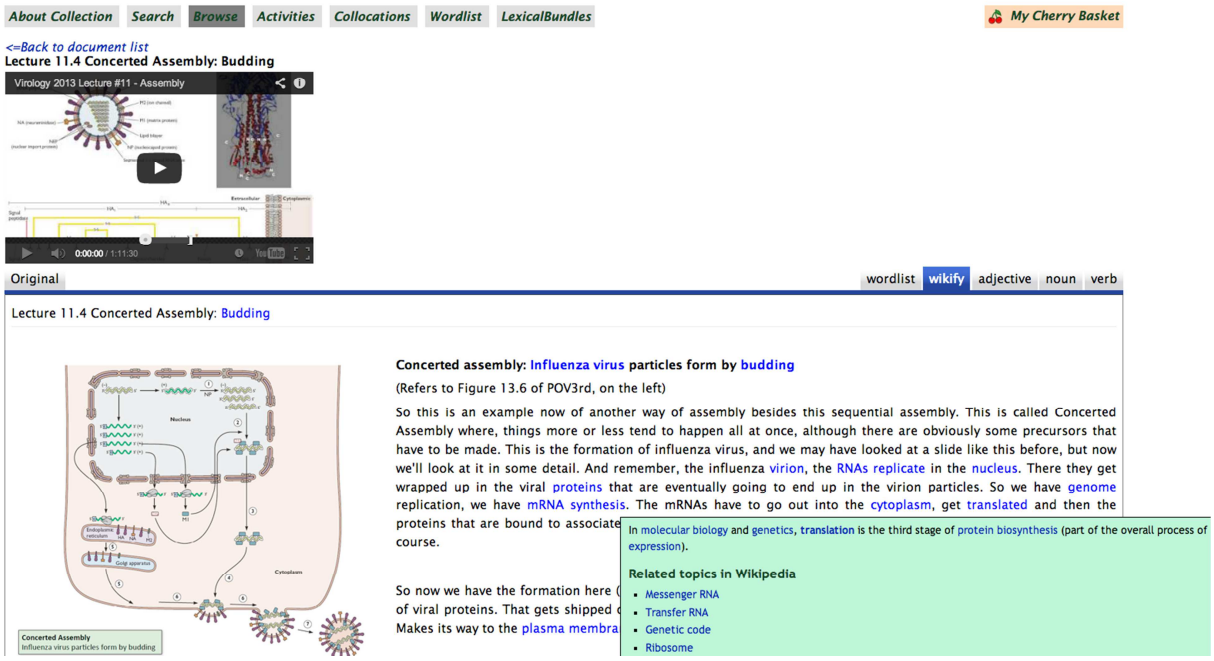
Items	**Number**
Podcast audio transcripts (*This Week in Virology*)	130
YouTube video lecture transcripts (Virology Course 2013 playlist)	110
Academic blog posts (*Virology Blog)*	540
Open Access reference articles (specific to virology research)	40

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.

Figure 2. Wikify Function in FLAX Language Collections.
Source: Lecture 11.4 on Concerted Assembly: Budding. In Coursera MOOC at Columbia University, *Virology 1: How Viruses Work.*



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 be included in the output. For example, *budding*, *mRNA synthesis*, *virion*, *nucleus*, cyto*plasm* and *translated*, … in the article pictured above titled, *Concerted Assembly: Budding* are identified as Wikipedia concepts. This definition for *translation* is extracted by the Wikipedia Miner tool*: "In molecular biology and genetics, translation is the third stage of protein biosynthesis (part of the overall process of expression)"*.

Learning Collocations

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 given 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 to build up collocation knowledge within specific domain areas. A corpus study into university language by Biber characterizes natural science as:

"... a discipline of discovery, identifying and describing entities that had not been previously considered. As a result, natural science employs a large set of highly technical words, like dextrinoid, electrophoresis, and phallotoxins. Most of these words do not have commonplace synonyms, because they refer to entities, characteristics, or concepts that are not normally discussed in everyday conversation." (Biber, 2006)

To compensate for the extensive use of highly technical words in ESAP collections like the one we have developed for the virology MOOCs, we focus on lexical collocations with noun-based structures because they are the most salient and important patterns in domain-specific text:

- verb + noun — *e.g. detect virus particles*
- noun + noun — *e.g. tobacco mosaic virus*
- adjective + noun — *e.g. negative strand virus*
- noun + *of* + noun — *e.g. genome of the virus*

The FLAX 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 as shown in Figure 3. below.

Figure 3. Collocations Associated with Virus in FLAX (Noun + of).
Source: Virology MOOC Collection in FLAX.

Lexical Bundles

Academic prose, in writing and in lectures, commonly contains multi-word sequences or "lexical bundles" with distinctive syntactic patterns and discourse functions (Biber & Barbieri, 2007; Biber et al, 2003, 2004). Typical patterns in the virology MOOC lectures include:

- noun phrase + *of* — *e.g. a DNA copy of*
- prepositional phrase + *of* — *e.g. at the end of*
- *it* + verb/adjective phrase — *e.g. it turns out that*
- *be* + noun/adjective phrase — *e.g. is an example of*
- verb phrase + *that* — *e.g. you can see 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, traditional language instructors developing self-access ESAP resources for their students, or instructional designers

developing linguistic support for MOOC or OCW academic content, can select parameters for exercise creation, and provide hints for learners through the FLAX instructor interface. The exercises are generated automatically, and can be reviewed and modified to discard undesirable language choices before presenting them to learners via the FLAX leaner interface.

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. FLAX has been 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 also been devised to offer a flexible suite of linguistic support options for enhancing such content across both open and closed platforms. It is anticipated that this open methodology for domain-specific language collections building will be of value to wider academic language communities across formal and informal education, with outputs available in the form of OSS and OERs.

Open online systems and resources like the language collections in FLAX proposed by this research have unique characteristics and challenges with regards to diffusion, adoption and integration. Resource quality and flexibility for adaptation are key drivers for (re)use and mash-ups in traditional language education. Observing how OERs are produced and shared by open digital scholars like Professor Racaniello at Columbia University is important for modeling newfound open digital literacies among the traditional language teaching community. 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. What is more, building collections with collaborators across formal and informal education will enable us to arrive at 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.

References

Biber, D., Conrad, S., & Cortes, V. (2003). “Lexical bundles in speech and writing: an initial taxonomy.” In A. Wilson et al. (Eds.), Corpus linguistics by the lune (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, 371–405.

Biber, D. (2006). University Language, A corpus-based study of spoken and written registers. John Benjamins, Amsterdam.

Biber, D., Barbieri F. (2007). “Lexical bundles in university spoken and written registers.” English for Specific Purposes, 26, 263–286.

Goulden, R., Nation, P. & Read, J. (1990). “How large can a receptive vocabulary be?” Applied Linguistics, 11, 341–363.

Hill, J. (1999). “Collocational competence.” ETP, 11.

Hill, J. (2000) “Revising priorities: form grammatical failure to collocational success.” In M. Lewis (Ed.), Teaching collocation, 70–87, LTP, England.

Lewis, M. (1997). Implementing the lexical approach: putting theory into practice. Hove: Language Teaching Publications.

Milne, D. and Witten, I.H. (2013) “An open-source toolkit for mining Wikipedia.” Artificial Intelligence, (194), pp. 222-239, January.

Witten, I.H., Bainbridge, D. and Nichols, D.M. (2010). How to Build a Digital Library. Morgan Kaufmann, Burlington, MA (second edition).

Wu, S. and Witten, I.H. (2013) “Transcending concordance: Augmenting academic text for L2 writing.” Submitted to Computer Assisted Language Learning

License and Citation