Second Language Learning in the Context of MOOCs

Shaoqun Wu[2], Alannah Fitzgerald[1] and Ian H. Witten[2]
[1]*Department of Education, Concordia University, Montreal, Quebec, Canada*
[2]*Computer Science Department, Waikato University, Hamilton, New Zealand*

Keywords: Second Language Learning, Corpus-based Language Learning, English for Academic Purposes, MOOCs, FLAX, Open Educational Resources.

Abstract: Massive Open Online Courses are becoming popular educational vehicles through which universities reach out to non-traditional audiences. Many enrolees hail from other countries and cultures, and struggle to cope with the English language in which these courses are invariably offered. Moreover, most such learners have a strong desire and motivation to extend their knowledge of academic English, particularly in the specific area addressed by the course.

Online courses provide a compelling opportunity for domain-specific language learning. They 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 a Coursera MOOC entitled *Virology 1: How Viruses Work*, offered by Columbia University.

1 INTRODUCTION

Massive Open Online Courses (MOOCs) are becoming popular educational vehicles through which universities reach out to non-traditional audiences. They are generally offered by English-speaking universities in the US and UK, and proponents often express an explicit desire to reach out to other countries and cultures. For example, Coursera aspires to provide a "meaningful learning experience for the millions of students around the world who would oCtherwise never have access to education of this quality" (Ng and Koller, 2013). Clearly, many MOOC students will encounter a language barrier during their study. Moreover, they will be strongly motivated to improve their knowledge of English for Academic Purposes (Dudley-Evans and St. John 1988); (Hyland, 2006) as it is used in the MOOC's subject domain.

The use of domain-specific corpora is a growing trend in language teaching and learning (e.g. Gabrielatos, 2005). Most corpora are based on particular domains, genres, or collections of certain types of document from which recurrent phrases and grammatical patterns can easily be retrieved (Stubbs and Barth, 2003). Among other aspects of language, a corpus provides 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).

We have developed an automated scheme called "FLAX" that extracts salient linguistic features from academic text and presents them in an interface designed for second-language students who are learning academic writing (Wu and Witten, 2013). The design is guided by several common ways of utilizing corpus technology. An extraction method is included that identifies typical lexico-grammatical features of any word or phrase in a corpus. Collocations and lexical bundles are automatically extracted; students can explore them by searching and browsing, and inspect them along with contextual information. FLAX also presents learners with common words, and academic words, hyperlinked to their usage and collocates in authentic contexts.

Typical MOOCs constitute a vast corpus of multimedia information, consisting predominantly of text but supplemented by images in the form of slides, audio, audio transcripts, and video; all (usually) in the English language. This paper argues

Wu S., Fitzgerald A. and H. Witten I..
Second Language Learning in the Context of MOOCs.
DOI: 10.5220/0004924703540359
In *Proceedings of the 6th International Conference on Computer Supported Education* (CSEDU-2014), pages 354-359
ISBN: 978-989-758-020-8

that the very same corpus, pre-processed appropriately and presented in a different way, provides a focused resource that allows second-language learners to improve their linguistic knowledge in the domain addressed by the MOOC. (It is also helpful for native speakers of English.)

This paper uses as an example a Coursera offering entitled *Virology 1: How Viruses Work*, from which we have built a FLAX collection. We illustrate in the next section how this resource has been augmented for language learning, and then review how learners can use it to explore language usage. Having established a specific context, we elaborate our position by showing how this approach might be used to facilitate language learning, and what organizational and teaching structures would be suitable to put such a proposal into practice.

2 BUILDING THE COLLECTION

2.1 Selecting and Preparing Materials

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 articles related to his virology courses, are published under a Creative Commons Attribution licence.

All 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.

2.2 Building Digital Library Collections

The FLAX Virology collection has the four components listed in Table 1. Textual documents are searchable, and browsable by title. Videos, audios and images are embedded within the document.

Table 1: Number of items in the collection.

Podcast audio transcripts	130
YouTube video lecture transcripts	110
Virology blog posts (lectures)	280
Open Access reference articles	40

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 described below are all extensions to Greenstone.

3 AUGMENTING TEXT FOR LANGUAGE LEARNING

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 students and gives them opportunities to encounter these components in various authentic contexts.

3.1 Words and Usage Patterns

The lexico-grammatical patterns of each word in the collection (excluding *a*, *an* and *the*) are extracted and grouped by position in sentence—near the beginning or in the middle—because these provide different views of the word's usage patterns.

For sentence-initial fragments, the part-of-speech tags of the opening words (except for conjunctions) plus one word following the query term are used to generate patterns. For mid-sentence fragments, the query term's syntactic type—verb, noun, adverb, or adjective—is used to select a stretch of text surrounding the query term, whose tags are used as patterns.

3.2 Key Concepts and Definitions

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

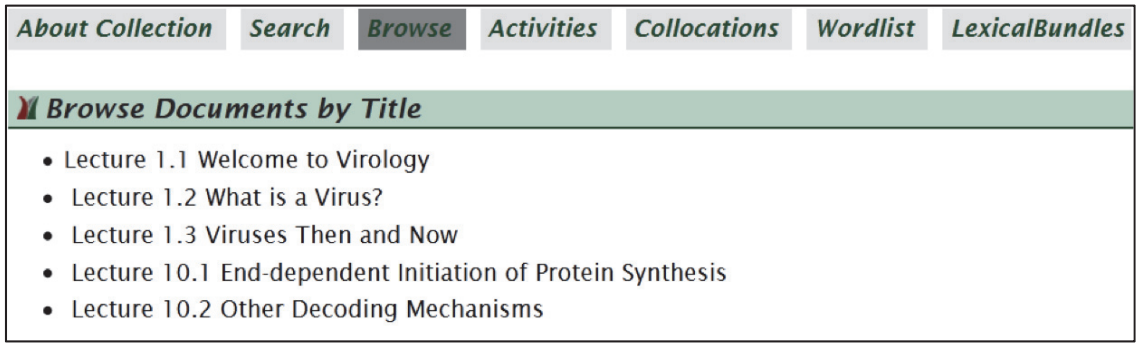

Figure 1: Main page of the Virology collection.

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."

3.3 Collocations and Lexical Bundles

The importance of collocation knowledge in language learning has been widely recognized. Hill (1999) observes that students with good ideas often lose marks in academic essays because they do not know the four or five most important collocations of a key word that is central to what they are writing about. Student text tends to be cumbersome and error prone because of insufficient collocation knowledge. The *Virology* collection contains massive, high-quality resources that help students build up collocation knowledge within the area.

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 and 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 fulfil 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).

To help users explore lexical bundles, FLAX extracts all short phrases that appear in the collection and sorts them by frequency. We chose four-word phrases because at any rate, this is the length of discourse bundle that appears most often in the literature. Bundles at the beginning of sentences ("head bundles") are treated separately from ones in the middle ("middle bundles").

4 HOW LEARNERS EXPLORE LANGUAGE USAGE

FLAX provides learners with simple interfaces to explore language features extracted from the course material. Learners can encounter and inspect words in their original context, or search for them by simply typing a word of interest, or browse them. Language activities built from course material reinforce what they have learnt.

Figure 1 shows the main page of the *Virology* collection of blog posts, which in fact correspond to lectures (this page is the same for the other three collections). The buttons at the top (*Search*, *Browse*, *Activities* etc.) link to language features described in earlier sections.

4.1 Exploring Articles

Users view articles by clicking *Browse* and selecting the article's title. Each article has four different views. One contains the original text, and students can watch or listen to the accompanying video or audio. The other three draw students' attention to the language features described above.

The *wordlist* view allows learners to analyze the range of vocabulary used in the article. It highlights the most frequent 1000 and 2000 words, taken from wordlists used in language teaching (West, 1953); academic words included in the list by Coxhead

(1998); and keywords. Clicking a highlighted word leads to a page that shows all sentences in the collection containing that word. Keywords (identified by the TF-IDF heuristic commonly deployed in information retrieval) are shown in the *keyword* view. The user can control the system's selectivity by adjusting a slider to reveal more or less keywords.

The *Wikipedia* view relates the terminology used in the article to the Wikipedia, highlighting concepts that are defined there to helps learners grasp their meaning. Clicking any highlighted phrase in the document brings up its definition, hyperlinked to the Wikipedia article itself; followed by a list of related topics in Wikipedia that can also be clicked.

The *collocation* view allows students to examine lexical compounds that occur in the article, divided into collocations that involve adjectives, nouns, prepositions, and verbs. Collocations are highlighted in the text to help students notice them and study their context. The system makes it easy for learners to study collocations in different contexts by connecting to two external collocation databases that are built from text in the British National Corpus, and from Wikipedia articles.

4.2 Search and Browsing

The *Search* button in Figure 1 displays usage patterns or collocations of a particular word. For example, one can study the patterns of the word *sequence*, a common academic word in the *Virology* collection, under verb + *sequence* + *of* and verb + *sequence* + *that* clause; or collocations of the word *virus*, the most frequent word. Collocations are grouped by syntactic pattern, e.g. noun + *virus*, verb + *virus* and noun + *of* + *virus*, and sorted by frequency. For example, the most frequent noun + *virus* or *virus* + noun collocations are *virus particle(s)* and following by *influenza virus*, *rabies virus*, *tobacco mosaic virus*. Clicking one reveals how it is used in context.

The *LexicalBundles* button lists bundles used at the beginning and in the middle of sentences under separate tabs. Clicking a bundle show the contexts in which it is used. In this collection, the most frequent bundles are conversational such as *And you can see*, *And this is a*, which indicate the spoken nature of this collection.

4.3 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: here are two. *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 students 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 teacher's discretion.

In a *Completing collocations* activity, certain words are again omitted from a document and users fill in the gaps. Here, however, missing words are chosen from collocations that have been identified in the document. FLAX chooses sentences, and highlights selected collocations. If the paragraph contains preceding and following sentences, they are shown as well, to provide context. Many teachers prefer to focus on certain types of collocation, e.g. noun+ noun, adjective + noun or verb + noun. This helps learners focus on sets of words that share similar meanings but have different usage (e.g. *problem* or *issue*) or word combinations specific to a topic (e.g. *virus infection* or *influenza virus*).

5 DISCUSSION

MOOC participants register for educational courses; they do not sign up as language learners. Columbia's virology MOOC is based on mastery learning (Bloom, 1984). Course content builds from week to week, and learners must master previous content before progressing. Assessments match this philosophy: weekly quizzes build towards a final exam.

Of course, the world of MOOCs is fluid: a great deal of experimentation is taking place in terms of

the educational theories and approaches that underpin the range of courses hosted by different institutions. Critics of methodology and terminology divide MOOCs into two camps (Daniel, 2012): ones based on traditional modes of instruction, typically hosted by proprietary learning platforms (like our virology example), and ones based on connectivist peer-learning approaches, typically built on open source platforms (Siemens, 2005).

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 with any 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.

Content varies in terms of licensing restrictions, depending on the publishing 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.

6 INTO THE FUTURE

A recent review commissioned by the UK Department for Business Innovation and Skills (2013) tracks the progress of the MOOC phenomenon as it moves from experimentation into maturity. Current work focuses on meeting the accreditation needs of learners, and on devising and developing new pedagogical models to better support online learning.

FLAX's capabilities for building language collections with comprehensive facilities for search and retrieval, and customized interactive learning of key domain terms and concepts, addresses the needs of both native and non-native speakers who are interested in engaging deeply with specific academic resources in English while developing their receptive reading and listening skills.

We plan to investigate further MOOC collections to determine whether FLAX can assist not only with mastery approaches to learning and assessment like those employed in the *Virology* course, but also with constructivist approaches that support peer learning and assessment—where collections will be derived from student texts, seminar discussions, and peer-review texts, as well as from expert text and lecture transcripts. This will promote the aggregation of crowd-sourced content for collaborative peer learning.

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

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.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher, 13 (6), 4-16.

Coxhead, A. (2000). A new academic word list. TESOL Quarterly, 34(2), 213–238.

Daniel, J. (2012). Making sense of MOOCs: Musings in a maze of myth, paradox and possibility. *Journal of Interactive Media in Education*. Retrieved on Nov 17, 2013 from http://jime.open.ac.uk/2012/18

Dudley-Evans, T., St John, M.J. (1988). Developments in English for Specific Purposes: A multidisciplinary approach. Cambridge: Cambridge University Press.

Gabrielatos, C. (2005) "Corpora and language teaching: Just a fling or wedding bells?" *Teaching English as a second or foreign language*, 8(4). Retrieved Oct 21 2013 from http://tesl-ej.org/ej32/a1.html.

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

Hyland, K. (2006). English for Academic Purposes: An advanced resource book. London: Routledge.

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

Nesselhauf, N. (2003) "The use of collocations by advanced learners of English and some implications for teaching." *Applied Linguistics*, 24(2), 223-242.

Ng, A. and Koller, D. (2013) "The online revolution: education for everyone." *Proc ACM SIGKDD Int Conf on knowledge discovery and data mining*, p.2.

Siemens, G. (2005). Connectivism: A learning theory for the digital age. International Journal of Instructional Technology & Distance Learning, 2(1).

Stubbs, M., and Barth, I. (2003) "Using recurrent phrases as text-type discriminators." *Functions of Language*, 10(1), 61-104.

UK Government Department of Business Innovation &

Skills. (2013). The maturing of the MOOC. London: UK Government Publications.
West, M. (1953). A general service list of English words. Longman, Green & Co., London.
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 Journal.*