Title of Project:
A New Paradigm for Open Data-driven Language Learning Systems Design in Higher Education

Researcher:
Alannah Fitzgerald
alannahfitzgerald@gmail.com

Current Affiliation
The University of Waikato

Doctoral Institution
Concordia University

Research Supervisor:
Dr. Steven Shaw
Concordia University

Final Report

Motivation for the Research

This doctoral thesis presents three studies in collaboration with the open source FLAX project (Flexible Language Acquisition flax.nzdl.org). This research makes an original contribution to the fields of language education and educational technology by mobilizing knowledge from computer science, corpus linguistics, and open education and proposes a new paradigm for open data-driven language learning systems design in higher education. Furthermore, the research presented in this thesis uncovers and engages with an infrastructure of open educational practices (OEP) that push at the parameters of policy for the reuse of open access research and pedagogic content in the design, development, distribution, adoption, and evaluation of data-driven language learning systems.

A central proposition of this thesis with publications is that where language corpora have been deployed in the research for linguistic analyses by researchers, the knowledge generated has often failed to translate into the design of openly accessible pedagogical applications for data-driven learning (DDL). Instead what we have witnessed is corpus systems that have been designed and developed primarily by and for corpus linguists for research purposes. This failure in knowledge translation is due in no small part to the following issues: copyright restrictions with the texts in corpus building that inhibit text data mining and sharing; subscription costs with natural language processing (NLP) and text analysis software tools that restrict access; and complex user interface designs of NLP and text analysis tools that limit uptake and utilization by
non-expert users, namely language teachers and language learners.

As per the directives of the editors working for the various publishers in preparing these manuscripts, the two journal articles and one book chapter central to Studies 1-3 of the thesis are very concise documents, two of which have been submitted for publication and one of which has been published by Routledge with TIRF as part of the *Global Research on Teaching and Learning English Series* in the area of digital language learning and teaching: research, theory and practice for which this doctoral dissertation grant was awarded. In order to ensure that the thesis reads as a coherent whole, additional bridging material has been included with references to supporting research with the FLAX project that foreground each study and show the relationship between the central studies in the thesis. The main thrust of the thesis demonstrates the iterative nature of all three inter-related research studies and their implications for current as well as planned future research, in addition to implications for policy and pedagogy.

**Research Questions**

For Study 1, the following research questions were investigated:

1. To what extent can open access content foster open educational practices among academic English language stakeholders for designing, developing and evaluating data-driven language learning resources?
2. What impact do the underlying business models and cultural practices of institutions and organizations have on open educational practices for remixing open access content in the design, development, implementation, and dissemination of resources for EAP in higher education?

In response to the language support collections developed for Study 2 for applications in non-formal online learning (MOOCs), the following research questions were devised as a basis to collect data from participants on their perceived experience of using the FLAX system:

1. Are automated domain-specific terminology learning support systems perceived as motivating to use (i.e., user-friendly and efficacious) in non-formal online learning where there is no formal language support provision?
2. Do the affordances of being able to browse and search data-mined course content that has been linked to auxiliary resources positively augment the learning and usability experience of MOOC platforms and learning management systems?

An identifiable lack of corpus-informed resources for teaching and learning specialized varieties of English was a motivating factor for conducting the experiment described in Study 3 in response to the following research questions:

1. To what extent can the digital commons of open and authentic content enrich data-driven learning across formal and informal language learning?
2. What effect does the application of DDL methods for querying open and authentic content have on the acquisition of specialized terminology, as opposed to accessing non-DDL-based online resources?

Research Methodology

Study 1 employs automated content analysis to mine the concept of open educational systems and practices from qualitative reflections spanning 2012-2019 with stakeholders from an ongoing multi-site design-based research study with the FLAX project. Design considerations are presented for remixing domain-specific open access content for academic English language provision across formal and non-formal higher education contexts. Primary stakeholders in this ongoing research collaboration include the following: knowledge organizations – libraries and archives including the British Library and the Oxford Text Archive, universities in collaboration with Massive Open Online Course (MOOC) providers; an interdisciplinary team of researchers; and knowledge users in formal higher education – English for Academic Purposes (EAP) practitioners.

Study 2 presents a data-driven experiment in non-formal higher education by triangulating user query system log data with learner participant data from surveys (N=174) on the interface designs and usability of an automated open source digital library scheme, FLAX. Text and data mining approaches (TDM) common to NLP were applied to pedagogical English language corpora, derived from the content of two MOOCs, (Harvard University with edX, and the University of London with Coursera), and one networked course (Harvard Law School with the Berkman Klein Center for Internet and Society), which were then linked to external open resources (e.g., Wikipedia, the FLAX Learning Collocations system, WordNet), so that learners could employ the information discovery techniques (e.g., searching and browsing) that they have become accustomed to using through search engines (e.g., Google, Bing) for discovering and learning the domain-specific language features of their interests.

Study 3 presents a data-driven experiment in formal higher education from the legal English field to measure quantitatively the usefulness and effectiveness of employing the open Law Collections in FLAX in the teaching of legal English at the University of Murcia in Spain. Informants were divided into an experimental and a control group and were asked to write an essay on a given set of legal English topics, defined by the subject instructor as part of their final assessment. The experimental group only consulted the FLAX English Common Law MOOC collection as the single source of information to draft their essays, and the control group used any information source available from the Internet to draft their essays.

Summary of Findings

Themes arising from the qualitative dataset in Study 1 point to affordances as well as barriers with the adoption of open policies and practices for remixing open access content for data-driven language learning applications in higher education. These themes are discussed against the backdrop of different business models and cultural practices present within participating
knowledge organizations. The design research interventions and findings captured in Study 1 are further supported and evolved by the mixed methods of research inquiry employed in Studies 2 and 3.

Findings from Study 2 indicate a positive user experience with interfaces that include advanced affordances for course content browse, search and retrieval that transcend the MOOC platform and learning management system (LMS) standard. Further survey questions derived from an open education research bank from the Hewlett Foundation are reused in this study and presented against a larger dataset from the Hewlett Foundation (N=1921) on motivations for the uptake of open educational resources.

Findings from an analysis of the two learner corpora of essays in Study 3 indicate that members of the experimental group appear to have acquired the specialized terminology of the area better than those in the control group, as attested by the higher term average obtained by the texts in the FLAX-based corpus (56.5) as opposed to the non-FLAX-based text collection, at 13.73 points below.

**Implications**

*A new paradigm for open data-driven language learning systems design in higher education:*

A basic premise underpinning the new research paradigm presented in this thesis, and demonstrated by the FLAX project, is that open data-driven language learning systems design as an approach is learner-centric and operates with the interface to the learner. Whether the learner is operating fully online in non-formal or informal learning mode or in a blended modality that is based both within and beyond the formal language classroom, this approach requires that the tools and interfaces, and indeed the corpora, be openly accessible and remixable for development or adaptation to meet this specific learner requirement. This method is different from existing DDL approaches which assume specialised knowledge or experience with DDL tools, interfaces and strategies, operating on mostly inaccessible corpora in terms of cost or design, or alternatively assuming training to, hopefully, compensate for this lack of knowledge and experience.

From a research and development (R&D) standpoint, the paradigm presented here also operates with the interface to knowledge organisations (universities, libraries, archives) and researchers who are engaging with open educational practices to push at the parameters of open policy for the non-commercial reuse and remix of authentic research and pedagogic content that is increasingly abundant in digital open access format for TDM purposes. This open access content is highly relevant to learning features of specialist varieties of English from across the academy but is otherwise off limits for development into proprietary learning materials by the commercial education publishing industry. Indeed, the open corpus development work presented in this thesis would not have been possible had it not been for the campaigners for copyright reform, the Internet activists, the open policy makers, the open-source software developers, and the advocates for open access, open data, and open education that have made these resources available for reuse and remix.
This paradigm leads down several paths, including research into understanding how users actually perceive, appropriate and use the approach based on the open tools and resources provided. This inquiry informs their design and development, in an R&D process that is presented here through the methodological lens of design-based research and design ethnography. This approach will be fundamentally different than if we assume that the user is actually a DDL or linguistics expert or that such an expert will be the learner’s interface to the system, by preparing output for the learner to experience and learn from. This approach will also be necessarily different than if we assume the user is always a formally registered student at a university with access to EAP support that may or may not offer DDL or linguistics expertise for learning the language features of specific discourse communities from across the academy.

The assumption behind this new paradigm that the right tools and resources can allow the end-learner to drive the processes autonomously is fundamentally revolutionary. This premise goes to the original contribution to knowledge of this thesis, but it also challenges and directs researchers and practitioners in the field to consider and take up this new direction with open data-driven language learning systems design for applications that can be scaled in higher education to meet the increasing numbers of learners who are coming online.

The focus on domain-specific terminology learning support via data-driven approaches is of course also decidedly different from the current EAP paradigm which in mainstream practice has been steadily evolving away from its roots in English for Specific Purposes (ESP), domain specificity and DDL processes towards the generic skills and knowledge programs currently in vogue that are arguably being steered by generic EAP coursebook publications from the commercial education publishing industry.

Thus, this is also a new paradigm based on DDL approaches, driving domain-specific terminology learning support for EAP across formal, non-formal and informal learning modalities in higher education. It will transform, potentially, the focus of DDL systems design developments in language support and learning in general toward the non-specialist end-learner, but also hopefully help re-establish the centrality of language specificity to the field of EAP.

The new paradigm is necessarily rooted in greater inter- or multi-disciplinarity. Given the goal of facilitating, in particular, the increasing number of learners who are coming online, and users of large-scale MOOC platforms who are trying to function in domain-specific subject areas that are invariably offered in the English language, the approach requires collaboration and cooperation among platform providers, subject academics and instructors, educational technologists, software developers, educational researchers, linguists and EAP practitioners with expertise in corpus-based and DDL approaches, and policy makers in knowledge organizations, such as in libraries, universities, and archives.
References


COnnectedREpositories (CORE): Aggregating the world’s open access papers. (n.d.). Retrieved from https://core.ac.uk/


Green, C. (2015, June 2). edX makes it easy for authors to share under Creative Commons [Blog post]. Retrieved from https://creativecommons.org/2015/06/02/edx-makes-it-easy-for-authors-to-share-under-creative-commons/


Nesi, H, Gardner, S., Thompson, P., & Wickens, P. (2007) *The British Academic Written English (BAWE) corpus*, developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800)


http://dx.doi.org/10.1093/intqhc/mzm042.


Watters, A. (2016). *The curse of the monsters of educational technology*. Tech Gypsies: Amazon Digital Services LLC.


Willinsky, J. (2002). Copyright contradictions in scholarly publishing. First Monday, 7(11).


