Prizing Open and Enhancing Research Corpora for Language Learning

Alannah FITZGERALD
Department of Education
Concordia University
Montreal, Canada
Support Centre for Open Resources in Education
The Open University
Milton Keynes, The United Kingdom
fitzgerald@education.concordia.ca

This paper presentation will discuss open educational resources (OER) for training English Language Teaching (ELT) and English for Academic Purposes (EAP) practitioners in the use of open corpora and open language analysis tools for English language teaching and learning. This will include how and why these corpora, linking both open and proprietary content, are managed and presented for more accessible uses in the language classroom or in independent language learning.

Organizations responsible for managing valuable research corpora such as the one hundred million-word British National Corpus (BNC) along with the British Academic Written English (BAWE) corpus of 2671 proficient pieces of assessed university student writing gathered from across the disciplines, both managed by Oxford University Computing Services, are cognizant of the fact that proprietary corpora such as these do not benefit from being closed to further research and educational resources development. To further exemplify this trend in openness for English language resources, Google has released collections of n-grams from web pages and made them available on the Linguistic Data Consortium's website. It seems that even Google realizes the value of releasing linguistic data as part of their web archiving activity for the purpose of enhancing computational linguistics research into the present and changing nature of modern languages as they are captured on the ever-expanding Web. Taking proprietary linguistic content beyond release for research purposes to enable further development as OER for uses in language teaching and learning is the focus of this paper.

This presentation will be of particular interest to ELT and EAP practitioners, along with corpus developers interested in exploring the issue of openness in relation to data-driven language learning resources development. Different open resources will be highlighted from a consortia of open projects for language learning and teaching, including the Flexible Language Acquisition (FLAX) project of large language collections based on linguistic content derived from Wikipedia, Google n-gram and Open Access research article corpora that have been linked to proprietary corpora such as the BNC and the BAWE. These collections, which present useful collocations and phrases based on complex handling of search queries, feature a simple concordancing user interface developed at the University of Waikato in New Zealand for non-specialist corpus users, namely language teachers and learners. (For further information on the development of the FLAX language collections and tools, and for research carried out using the FLAX software highlighting the success of the

simple user interface, please see Wu, Franken & Witten, 2009 & 2010; Wu, Witten & Franken, 2010).

Learning support collections will also be presented for exploiting the FLAX collections based on the TOETOE (Technology for Open English - Toying with Open E-resources) project at the Support Centre for Open Resources in Education (SCORE) at the Open University. Further OER from the OpenSpires project at the University of Oxford, including podcasts of lectures and transcripts, that have embedded data-driven learning features will also be presented, including vocabulary lists with links to sample sentences, frequent word combinations, an online dictionary and explanations of concepts in Wikipedia.

As a takeaway, participants will leave this presentation with a clear understanding of the benefits of developing an open digital infrastructure for corpus-based resources in language teaching and learning, along with information pertaining to support networks for encouraging good practice with the use of corpus-based OER among their colleagues at their home institutions.

References

British National Corpus, version 3 (BNC XML Edition). (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Available from: http://www.natcorp.ox.ac.uk/ [May 19, 2012].

Nesi, H, Gardner, S., Thompson, P. & Wickens, P. (2007) The British Academic Written English (BAWE) corpus.

Wu, S., Franken, M., & Witten. I. H. (2010). 'Supporting collocation learning with a digital library.' *Computer Assisted Language Learning.* 23: 87-110.

Wu, S., Witten, I. H. & Franken, M. (2010). 'Utilizing lexical data from a Web-derived corpus to expand productive collocation knowledge.' *ReCALL* 22: 83-102.

Wu, S., Franken, M. & Witten, I. H. (2009). 'Refining the use of the web (and web search) as a language teaching and learning resource.' *Computer Assisted Language Learning* 22: 249-268.