Chapter 25
Automatically Augmenting Academic Text for Language Learning: PhD Abstract Corpora With the British Library

Shaoqun Wu
University of Waikato, New Zealand

Alannah Fitzgerald
Concordia University, Canada

Ian Witten
University of Waikato, New Zealand

Alex Yu
Waikato Institute of Technology, New Zealand

ABSTRACT

This chapter describes the automated FLAX language system (flax.nzdl.org) that extracts salient linguistic features from academic text and presents them in an interface designed for L2 students who are learning academic writing. Typical lexico-grammatical features of any word or phrase, collocations, and lexical bundles are automatically identified and extracted in a corpus; learners can explore them by searching and browsing, and inspect them along with contextual information. This chapter uses a single running example, the PhD abstracts corpus of 9.8 million words derived from the open access Electronic Theses Online Service (EThOS) at the British Library, but the approach is fully automated and can be applied to any collection of English writing. Implications for reusing open access publications for non-commercial educational and research purposes are presented for discussion. Design considerations for developing teaching and learning applications that focus on the rhetorical and lexico-grammatical patterns found in the abstract genre are also discussed.

DOI: 10.4018/978-1-5225-5140-9.ch025

INTRODUCTION

A growing body of research aims to understand the linguistic features of academic text and their bearing on the problems faced by students learning to write. To support this work, corpora of academic writing have been built as a reference and research base. These include lists of academic words, syntactic patterns characteristic of academic writing, and distinctive linguistic characteristics of multi-word sequences that fulfill discourse functions. The research raises many pedagogical implications, and it should be possible to apply the findings to academic teaching practice. How, then, can we bridge the gap between expert and student writing? Suggestions include helping students understand the importance of learning common collocates or recurrent lexical or grammar patterns in different contexts (Coxhead, 2007), making commonly used lexical bundles more accessible (Hafner & Candlin 2007), and providing students with more realistic writing models (Hyland, 2008a).

There are computer tools that help teachers and learners analyze and study language features. Some allow users to upload text and examine the vocabulary it uses. Concordancers (for example, COCA, Compleat Lexical Tutor, and SKELL), initially designed for linguists, are frequently used to explore corpora with a view to exposing linguistic patterns. Some present short snippets of text, while others present items in paragraph-length units; some limit the items that are retrievable. Teachers and students have been using concordances or alike to obtain, organize, and study real language data derived from corpora. This approach is called data-driven learning (Johns, 1991), and is advocated by many researchers (for instance, Boulton & Thomas, 2012; Boulton & Pérez-Paredes, 2014; Chang, 2014; Cobb & Boulton, 2015; Vyatkina, 2016; Boulton & Cobb, 2017).

We have designed and constructed a language learning system called FLAX that takes academic texts, and automatically extracts linguistic features that have been identified in the research literature. Our design is principled and underpinned by two theories: noticing hypothesis (Robinson, 1995; Schmidt, 2001) and inductive (discovery) learning (Bernardini, 2002). Noticing is facilitated through input enhancement and enrichment that has been proven to be effective in students' recognition and recall of language features, for example, collocations (Sharwood, 1993; Sonbul & Schmitt, 2013; Szudarski & Carter, 2016). FLAX presents important components in academic texts−academic words, key concepts, collocations, and lexical bundles−in a way that draws them to the attention of students. External resources (Wikipedia) are linked to these components to gives students opportunities to encounter them in various authentic contexts, and repeatedly. Simple interfaces are developed so that students can use information discovery techniques (e.g., searching and browsing) that they have become familiarized with through search engines (e.g., Google, Bing) to discover and study the language features of their interests.

The aim of this chapter is not to explain how the FLAX system works behind the scenes: that would be a technical discussion that is relatively uninteresting from the point of view of language education. Instead, we aim to illustrate what it does by describing the result of processing the PhD abstract corpora hosted by the British Library. It works entirely automatically, without any human input, and can be applied to any collection of academic text—for example, samples of writing collected by an individual teacher; an entire textbook; or essays written by students (provided that any texts intended for use are all available electronically).

This chapter is structured as follows. The next section summarizes the research background, including a brief survey of existing corpora and their interfaces, and research on the identification and use of wordlists, collocations and lexical bundles for teaching and learning. The next section describes the PhD abstract corpora that are used as examples throughout the chapter. This section will present a brief