Journal of English for Academic Purposes 15 (2014) 1–13

Contents lists available at ScienceDirect

Journal of English for Academic Purposes

journal homepage: www.elsevier.com/locate/jeap

Multiword constructions in first year business and engineering university textbooks and EAP textbooks

David C. Wood [a,*], Randy Appel [b,1]

[a] School of Linguistics and Language Studies, Carleton University, 1125 Colonel By Drive, Ottawa, Canada K1S 5B6
[b] Concordia University, Department of Education, LB-579, 1455 de Maisonneuve Blvd, West Montreal, Canada H3G 1M8

ARTICLE INFO

Article history:
Received 17 April 2013
Received in revised form 18 March 2014
Accepted 24 March 2014
Available online 6 May 2014

Keywords:
English for academic purposes
Multiword constructions
Corpus research
First year textbooks

ABSTRACT

The body of research into formulaic language, multiword meaning and functional units, has shed light on their role in academic discourse. Some studies have examined which units might be of greatest utility for teaching English for academic purposes (EAP), dealing with corpora of academic language. The reality of most EAP programs is a focus on preparing students for first year university courses, in which textbooks play the key role. The present study uses a corpus analysis method focusing on units called multiword constructions (MWC), and addresses three hitherto unexplored key issues for EAP teachers and materials developers: What MWC exist in first year textbooks used in business and engineering, the most popular EAP student majors at a large Canadian university; whether the first year textbook MWCs are present in the reading texts in popular EAP textbooks; whether the EAP textbooks contain pedagogical treatment of MWC. The results show that the MWC are weakly present in the EAP materials, and that they are not dealt with pedagogically. This knowledge has important implications for EAP practitioners and materials developers.

1. Introduction

Teachers and materials designers in the field of English for academic purposes (EAP) are engaged in helping students acquire the language skills necessary to successfully navigate their education in an English language environment. A major concern in this process is to identify which lexical items will be of greatest benefit to the learner. Believing that the most frequently occurring words are also the most useful items to teach, corpus studies can help more accurately define which items should be presented to the learner. The use of large-scale corpora has provided a useful tool to help in this search. Early efforts to identify frequently occurring words within the academic register include Xue and Nation (1984), who used frequency counts to help develop their 'University Word List'. Building off this and other early studies, more recent efforts to identify important academic vocabulary have led to the development of the Academic Word List (AWL). The AWL (Coxhead, 2000) is a widely used frequency based listing of academic vocabulary. The corpus analyzed for the creation of the AWL contained approximately 3.5 million words of academic text from across a range of academic disciplines and fields.

Developments in corpus-based research have allowed for increasingly large-scale lexical studies that can more accurately reveal general patterns within the language system. Within corpus-based research, studies looking at longer sequences of

* Corresponding author. Tel.: +1 613 520 2600x6684.
E-mail addresses: david_wood@carleton.ca (D.C. Wood), r_appel@eduation.concordia.ca (R. Appel).
[1] Tel.: +1 514 848 2424x2004.

http://dx.doi.org/10.1016/j.jeap.2014.03.002
1475-1585/

recurrent word strings have developed into a growing area of investigation (Cortes, 2006; Hyland, 2008; Partington, 1998). For multiword units, often labeled using the umbrella term of *formulaic language* (FL), studies of *lexical bundles* (LB), corpus-driven word combinations extracted using frequency and range criteria (e.g. Biber & Conrad, 1999; Biber, Conrad, & Cortes, 2004) have revealed previously hidden types of frequent word combinations. Research conducted by Simpson-Vlach and Ellis (2010) on formulaic language units used in academic prose, and multiword constructions (MWC) by Liu (2012) has furthered our knowledge of these important language elements which serve key functions in the construction of academic discourse and represent a potentially valuable tool to use in academic English education.

The present study represents an effort to shed light on a new area of focus in the research on formulaic language in academic discourse. It identifies formulaic language extracted from a corpus of first year university textbooks, which are the key sources of exposure to academic prose encountered by EAP students. The first year textbooks in this study are representative of the two most popular majors for EAP students, engineering and business, and are therefore more grounded in student reality than some of the corpora used in previous research, which has tended to focus on specific academic disciplines and a range of text types, including journal articles. This focus on first year textbooks in two major areas has resulted in a corpus composed of texts from a range of disciplines but more immediately relevant to EAP instruction than a concentration on just one discipline. Furthermore, the methods used in extracting formulaic language from the corpus differ from those used in previous research, and popular commercially available EAP textbooks are scanned for the presence of the key sequences found in the first year textbooks. The result is a reflection of the reality of EAP learners' academic reading challenges, the nature of the texts they encounter in first year studies, and the relative utility of the EAP materials often used to aid them in coping with reading academic texts.

2. Lexical bundles

Lexical bundles (LBs) are combinations of three or more words identified by means of corpus analysis software. These bundles are identified using a minimum frequency cutoff, and range (minimum number of texts in which the bundle appears). Biber et al. (2004) proposed three categories of functions of lexical bundles: *stance*, expressing epistemic certainty, attitude, or modality toward subsequent propositions; *discourse*, marking relationships among parts of discourse such as introducing topics or elaborating and extrapolating from topics at hand; *referential*, referring directly to temporal, spatial, physical context.

Research shows that bundles play key roles in particular norms for communication (Cortes, 2004; Hyland, 2008), establishing relationships between readers and authors, presenting propositional content, showing evaluation of ideas, engaging readers, explaining facts, and organizing discourse. Hyland (2008) remarks that writing in electrical engineering and biology employs large numbers of directive stance bundles, addressing the reader directly (e.g. *we can see that*, *it is important that*), and a high proportion of referential bundles to describe research procedures. Cortes (2004) notes that published research in biology employs referential bundles to indicate timing and location (e.g. *at the beginning of*), describes physical attributes (*the depth of*) and quantities (*a large number of*), and that university students generally tend not to use the bundles which are common in their fields.

Lexical bundles are discipline-bound, and associated with distinct communicative conventions (Hyland & Hamp-Lyons, 2002). Hyland (2008), in a study of lexical bundles across four disciplines, found that less than half of the fifty most frequent bundles in his corpus occurred in all four disciplines, and that there were clear differences in the types of bundles across the four disciplines. Biber (2006), in a corpus-based analysis of university language, examined lexical bundles in textbooks. He found that academic disciplines differed in their use of lexical bundles, with natural and social sciences relying on them more than the humanities.

While the research uncovering LBs has proven fruitful in attempts to characterize academic discourse in general, it has been noted (e.g. Simpson-Vlach & Ellis, 2010; Liu, 2012) that the LB research tends to produce lists of units which appear semantically and structurally incomplete, such as *to do with the*, or *I think it was*. The main issue with these items is that they are "neither terribly functional nor pedagogically compelling" (Simpson-Vlach & Ellis, 2010, p 493). In light of this, the present study utilizes a somewhat different method of analysis from the LB research in identifying important multiword sequences, and we have adopted the term multiword constructions (MWC) as a label for the items we are identifying.

3. Formulaic language and multiword constructions

In an attempt to deal with some concerns associated with prior corpus-based research into formulaic language, researchers have begun to develop new methods of identifying multiword constructions in various corpora. In a recent effort to develop a list of formulaic language from academic discourse, Simpson-Vlach and Ellis (2010) created the Academic Formulas List (AFL). Using a corpus of 2.1 million words of academic speech and writing, and comparing it to a non-academic corpus, the AFL attempts to delineate which formulas are truly academic. The corpora were scanned for formulaic language at a frequency cut-off of 10 per million and a range of three out of four academic disciplines was set for the written corpus. In an effort to address the issue of low psychological salience and pedagogical inutility of the lists of units uncovered in the LB research, the authors used mutual information (MI) scores as a measure of collocation strength, combined with frequency data and a rating by instructors and testers, to produce a composite score which determined the final lists.

Liu (2012) notes that the work of Simpson-Vlach and Ellis (2010) did result in lists of sequences which seem more complete and less semantically and syntactically incomplete than LB research, however their work still included a number of