English for Specific Purposes 51 (2018) 1–17

Contents lists available at ScienceDirect

English for Specific Purposes

journal homepage: http://ees.elsevier.com/esp/default.asp

The language of civil engineering research articles: A corpus-based approach☆

Alexander Gilmore [a,*], Neil Millar [b]

[a] Department of Civil Engineering, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[b] Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan

ARTICLE INFO

Article history:

Keywords:
Corpus analysis
Specialized corpora
Civil engineering research articles
Keywords analysis
Formulaic sequences
Materials design

ABSTRACT

This paper describes a corpus-based investigation of the 8 million word Specialized Corpus of Civil Engineering Research Articles (SCCERA), developed at the University of Tokyo. A keyword analysis was first performed in order to identify words associated with civil engineering research articles and of potential pedagogic value. These were then compared with established external wordlists (the New General Service List and the New Academic Word List) to categorize keywords into those: (i) commonly occurring in general English; (ii) commonly occurring in academic English, and (iii) not occurring in either the NGSL or NAWL. Keywords in the 11 sub-disciplines of civil engineering displayed marked heterogeneity, raising questions about exactly how specialized a corpus needs to be in order to be of pedagogic value. In a separate 'cluster analysis', 3-, 4-, 5- and 6-word combinations were extracted in order to identify fixed expressions common to the field. These were found to typically belong to one of five categories: (i) cause and effect language; (ii) comparison and contrast language; (iii) language of quantification; (iv) deictic language; (v) language showing the writer's stance. The pedagogic implications of these findings are discussed.

1. Introduction

1.1. The role of specialized corpora in ESP contexts

The 'corpus revolution' which took place in linguistics in the 1980s and 1990s (Rundell & Stock, 1992) has had a major impact on language learning, particularly with respect to the design of dictionaries and reference grammars, which are now typically 'corpus-informed' (Gilmore, 2015). However, while the large 'mega-corpora' available today have been crucial in providing a solid foundation for our understanding of more general lexico-grammatical patterning in English, they are less helpful for the analysis of language used in specific academic or professional contexts such as civil engineering, where large variability has been found to exist between different academic disciplines in terms of word frequencies, collocational patterns and rhetorical moves. For example, Hsu (2014) found that the vocabulary necessary to reach 95% lexical coverage in the 20 sub-corpora of her Engineering Textbook Corpus ranged from 3500 to 8500 word families. Hyland (2008), comparing 4-word lexical bundles from the fields of Biology, Electrical Engineering, Applied Linguistics and Business Studies, calculated that over

☆ Research conducted at the University of Tokyo, Japan.
* Corresponding author.
E-mail addresses: alexgilmore@mac.com (A. Gilmore), kansaineil@gmail.com (N. Millar).

https://doi.org/10.1016/j.esp.2018.02.002
0889-4906/

half of the extended collocations in each discipline did not occur in the other subject areas examined: 4-word bundles like *as shown in figure or it can be seen* appeared to be unique to the Electrical Engineering sub-corpus in his data. While he points out that it is the use of this kind of genre-specific language that identifies writers as expert members of their own particular discourse communities, the disciplinary variability commonly observed has meant that for English for Specific Purposes (ESP) practitioners just 'working out basic items to be dealt with is a key teaching problem' (Gavioli, 2006: 23). Given the fact that publishers are often unwilling to invest in ESP textbooks because of the restricted target audience and limited potential profits (Bennett, 2010; Boulton, 2012), ESP teachers generally have to rely on their own resources for the creation of discipline-specific materials. Specialized corpora can be easily constructed in-house and provide an effective and convenient way to identify key language patterns of relevance to specific disciplines (Mudraya, 2006).

Civil engineers typically have to produce a wide variety of written genres, reflecting both the range of contexts in which they tend to work (academic institutions, construction sites, business, etc.) and the multiple audiences they address (engineering experts, governmental bodies, the general public, etc.). The Civil Engineering Writing Project at Portland State University in the USA, for example, identified at least 10 different genres in their corpus of student/practitioner writing, including site visit reports, cover letters, project-related emails and technical memoranda (Civil Engineering Writing Project 2017). We acknowledge the importance of these genres, but chose to limit our corpus to civil engineering research articles for two main reasons. Firstly, we wished to focus on the immediate needs of our target audience of post-graduate students, researchers and academic staff who are required to publish empirical research in academic journals. Secondly, we wished to work with single source data (i.e. research articles) to ensure that any empirical claims are securely grounded. The inclusion of multiple genres would not allow this – a point also made by Hoey (2005).

1.2. Creating wordlists from corpus data

Second language learners in academic environments inevitably need a large vocabulary in order to function effectively. For example, it is estimated that fluid reading requires understanding of somewhere between 95% and 98% (e.g. Hsu, 2014; Laufer, 1992; Nation, 2006) of the tokens within a text,[1] and that between 8000 and 9000 word families are necessary to provide 98% coverage of an academic text (Nation, 2006).[2] Although an effective vocabulary will include items that typically crop up in the general language, and are therefore likely to be familiar to students, it will also include words that they are less likely to encounter outside an academic setting, ranging from general academic lexis to discipline-specific technical terms. The use of corpora has facilitated the compilation of wordlists containing the vocabulary learners are most likely to encounter in an academic setting, which, in turn, can have applications for both language learning and teaching.

One of the most widely used wordlists predates modern corpus linguistics. The General Service List (GSL), published in 1953 (West, 1953) contains 2000 'word families' (i.e. base form plus inflected forms) that, based on frequency and other factors, were considered most useful to learners of English as a Second Language (ESL). Both research carried out at that time and more recently (Schonell, Meddleton, & Shaw, 1956; Adolphs & Schmitt, 2003; Brezina & Gablasova, 2013) indicates that the list provides substantial coverage of general texts (90%–99% for speech; 80%–85% for writing), as well as academic texts (70%–75%). An updated version of the GSL, 'NGSL' (Browne, Culligan, & Phillips, 2013), derived from a 273 million word sample of the Cambridge English Corpus (CEC), has been found to provide around 5–6% more coverage than West's original GSL with 800 fewer lemmas and was therefore used in our analysis here.

The Academic Word List (AWL) (Coxhead, 2000), derived from a 3.5 million word multi-disciplinary corpus covering 28 disciplines, contains 570 general word families that students in tertiary education are most likely to encounter. Similar to the GSL, an updated version of Coxhead's AWL, the New Academic Word List (Browne et al., 2013), has been produced, based on a carefully selected 288 million word corpus of academic English (for more information see: http://www.newgeneralservicelist.org/nawl-new-academic-word-list/). Since the NAWL is derived from a considerably larger corpus and is designed to work in conjunction with the NGSL, we also use this in our analysis. The list is largely non-technical and can be seen as representing a core vocabulary that students are likely to meet, irrespective of their particular area of study. This contrasts with discipline specific vocabulary (also referred to as 'technical' or 'specialised' vocabulary) which people from outside a given field are unlikely to be familiar with. Research by Chung and Nation (2004) suggests that up to 30% of academic texts can be technical in nature.

However, the existence of a core academic vocabulary, common to a wide range of disciplines, is questioned by Hyland and Tse (2007). Although they find that the AWL provides similar levels of coverage, they show items on the list often vary across disciplines in terms of range, frequency, collocation and meaning. Take, for example, the word *analyse* – in the social sciences the nominal form predominates, while in engineering the form *analytical* is six times more frequent. They conclude that "the different practices and discourses of disciplinary communities undermine the usefulness of such lists" and suggest that "teachers help students develop a more restricted, discipline-based lexical repertoire" (Hyland & Tse, 2007: 235).

[1] This threshold level of 95% is, of course, an oversimplification of a complicated picture, where reading comprehension depends on many factors including importance of a particular lexical item for comprehension, its position in the text and 'guessability' (Ward, 1999: 309).

[2] As Ward (2009) points out though, word list coverage figures are usually based on *word families* rather than all the possible inflections and derivations of headwords, assuming that learners will automatically recognize any derived forms if they know the base form. The combined GSL and AWL expand from 2,570 words to about 11,000 words when all the family members are included, so we may be underestimating students' vocabulary learning loads.