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

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.

The availability of corpus tools enables researchers and teachers to identify discipline specific core lexical items and how they are used in context. For example, Wang, Liang, and Ge (2008) report on the compilation of medicine specific academic wordlist based on research articles, while Yang (2015) follows a similar methodology to produce an academic wordlist for nursing. Based on frequency and excluding words found in the GSL and technical terms, these lists represent discipline specific versions of the AWL. They provide marginal gains in text coverage over the average 10% reported by Coxhead (2000) – 12.2% and 13.8%, respectively. For the sub-discipline of midwifery, Chiba, Millar, and Budgell (2010) adopt a statistical approach (corpus keywords) to the identification of discipline specific core vocabulary. Their analysis focuses on keywords that do not occur in established word lists (which they term 'off-list words'). They show that while the AWL and GSL combined provide 85% coverage of texts in the discipline, the addition of the top 1000 'off-list words', many of which constitute technical and semi-technical terms, raises the text coverage to 94.8%, bringing the coverage in line with estimates for fluid reading (Laufer, 1992). A similar approach to Chiba et al. (2010) is adopted in this study of the language of civil engineering research articles.

1.3. *Fixed expressions/word bundles in corpora*

In addition to knowledge of vocabulary, language learners require an understanding of how words are combined into meaningful expressions (Nation, 2013). Corpus linguistics has demonstrated that natural language often takes the form of recurrent clusters of words (Kjellmer, 1994; Altenberg, 1993). For example, Biber et al. (1999) estimate that 3–4 word 'lexical-bundles' (i.e. recurring multiword sequences; *n-grams*) make up 28% of conversational text and 20% of academic texts. As Martinez and Schmitt (2012) note, formulaic language is prevalent because it is used to encode a wide range of important referential, communicative, and textual functions. For example, in academic texts we might find recurrent fixed or semi-fixed phrases present for expressing cause–effect relationships (e.g. *it can be concluded X*), the writer's stance (e.g. *it is important to note X*), and indexing (*X is shown in Fig #*). There is evidence that, for learners, knowledge of formulaic language can lead to faster processing (Conklin & Schmitt, 2008), while the ability to use it appropriately can lead to gains in perceived fluency (Boers, Eyckmans, Kappel, Stengers, & Demecheleer, 2006) and comprehensibility (Millar, 2011).

As reported by Biber and Barbieri (2007), lexical bundles occurring in a given register tend to be (i) extremely common, (ii) structurally incomplete, (iii) non-idiomatic and (iv) lacking in perceptual salience. In terms of frequency, fixed expressions can occur as often as individual lexical items – the 4-word bundle *on the other hand*, for example, crops up in SCCERA to a similar extent as words such as *absorption*, *corrosion*, *normalized*, or *evaluate*. It might seem strange therefore that, as Martinez and Schmitt (2012) point out, lexical bundles are rarely targeted for explicit attention or noticing in language textbooks. This probably relates to the fact that because they lack perceptual salience, they have tended to go largely unnoticed by researchers, materials designers and language teachers until relatively recently. By providing us with objective frequency data, corpus linguistics has allowed us to see more clearly what was always there, in front of our eyes, and affords the opportunity for a more principled basis for selection of target language items. As Römer (2010) suggests, the shift away from a focus on individual words to also include key fixed expressions can provide deeper insights into, what she terms, the *phraseological profile* of a genre. In fact, lexical bundles typically complement the function of individual lexical items by either bridging or preceding new propositional information in the developing discourse and therefore acting as '*discourse frames*' (Biber & Barbieri, 2007; Biber, Conrad, & Cortes, 2004):

i. The compressive strength *as a function of* cumulative heat release (bridging propositional content)
ii. *It can be seen that* the wave height is dampened over the vegetation region (preceding propositional content).

Chen and Baker (2010) point out that the growing evidence from corpus analysis for the importance of lexical bundles in academic writing is not yet reflected in the curricula and language learning materials produced by ELT publishers. They conclude that 'after careful selection and editing, the frequency-driven formulaic expressions found in native expert writing can be of great help to learner writers to achieve a more native-like style of academic writing, and should thus be integrated into ESL/EFL curricula' (Chen & Baker, 2010: 44). We agree with their position and also adopt a frequency-driven approach here to identify key lexical bundles relevant to civil engineering students, categorized into functional groups for pedagogic rather than research purposes.

1.4. *Goals of the present research*

This paper describes a corpus-based investigation of the Specialized Corpus of Civil Engineering Research Articles (SCCERA), developed at the University of Tokyo between 2013 and 2014. A keyword analysis was first performed in order to identify words which are associated with the field of Civil Engineering, and therefore of potential pedagogic value. The keywords, i.e. those words overrepresented in comparison to a general English reference corpus, were then categorized according to two established external wordlists – the New General Service List (NGSL) and the New Academic Word List (NAWL). This allowed identification of keywords (i) commonly occurring in general English; (ii) commonly occurring in academic English, and (iii) not occurring in either the NGSL or NAWL, and, therefore, likely to be particular to this field. In a separate analysis, 3-, 4-, 5- and 6-word lexical bundles were extracted in order to identify phrases typical of the field.

The investigation aims to address 2 broad research questions:

1. What can an analysis of keywords and lexical bundles in SCCERA tell us about the language of civil engineering?
2. What are the potential pedagogic benefits of creating specialized corpora such as SCCERA for ESP instructors?

2. Data and methods

2.1. The Specialised Corpus of Civil Engineering Research Articles

The Specialised Corpus of Civil Engineering Research Articles (SCCERA) contains 1100 civil engineering articles published in 45 international journals between 1989 and 2014, with contributions from 3807 authors from 1598 institutions in 80 countries. At approximately 8 million words, the corpus exceeds the minimum size of one million recommended for specialized corpora (Kennedy, 1998; Pearson, 1998; Rea Rizzo, 2010). SCCERA is designed to be representative of the type of articles that researchers in the Department of Civil Engineering at the University of Tokyo seek to publish. The following criteria for inclusion of articles in the corpus were identified through consultation with staff from the Department of Civil Engineering:

i. Articles are peer-reviewed and published in influential journals cited in the Science Citation Index Expanded (SCI®) or Social Sciences Citation Index (SSCI®). Exceptions are made for specific journals considered to be ‘key’ or ‘desired outlets for academic work’ in the Department of Civil Engineering.
ii. Preference is given to articles listed as ‘most cited’ or ‘most viewed’ by the publishers (where information was available on the publisher’s website).
iii. Articles are representative of the 11 main sub-disciplines in the Department of Civil Engineering at the University of Tokyo.
iv. Articles are representative of variety within the field, in terms of research topic, author, language characteristics (based on geographical location: native-speakers and non-native speakers) and publishers.

Table 1 shows the size and lexical variety of the corpus and its sub-components. Type/token ratio (TTR) is calculated by dividing the total number of different words (types) by the total word count (tokens). Standardized type/token (STTR) ratio is calculated by taking the average type/token ratio based on analysis of consecutive 1000-word chunks of text. While both provide a measure of lexical variety within a text, STTR is less sensitive to variations in text length (Scott, 2012). The eleven sub-corpora are broadly comparable in size, ranging between 888,880 words for Hydrology & Water Resources Engineering and 578,839 words for Infra-structure Development. Lexical variation in the corpus (STTR value of 35.05%) is situated between that of general corpora of written English (e.g. 45.53% in the Freiburg-LOB corpus[3]) and corpora of spoken English (32.96% in the spoken sub-corpus of the BNC) (Baker, 2006: 52).

2.2. Identification and classification of keywords in the corpus

The term *keywords* is used in corpus linguistics to describe lexical items that occur more frequently in the target corpus than would be expected by chance, when compared to a larger reference corpus (Baker, Hardie, & McEnery, 2006). The degree to which a type is over-represented in a target corpus in comparison to a reference corpus, the ’keyness’ of a word, is typically measured by log-likelihood – a statistic which, similar to chi-square, compares observed and expected values for two data

Table 1
Lexical characteristics of SCCERA.

Component of SCCERA	Tokens	TTR	STTR
Whole corpus	**7,806,431**	**2.59%**	**35.05%**
1. Coastal Engineering	758,567	2.81%	35.33%
2. Infra-structure Development	578,839	2.84%	35.85%
3. Concrete Engineering	679,709	2.55%	35.22%
4. Regional Planning, Surveying, Remote Sensing	657,245	2.70%	35.13%
5. River & Environmental Engineering	698,282	2.88%	35.49%
6. Hydrology & Water Resources Engineering	888,880	2.25%	34.30%
7. Geotechnical Engineering	635,579	2.39%	32.91%
8. International Projects	866,121	2.39%	37.18%
9. Transportation Research	734,853	2.54%	35.60%
10. Mechanics & Structures	707,325	2.28%	33.30%
11. Earthquake & Disaster Mitigation	601,031	2.85%	35.21%

[3] The Freiburg-Lancaster-Oslo-Bergen Corpus (Freiburg-LOB corpus) is a general English corpus of approximately 1 million words, composed of 500 texts distributed across 15 text categories.

sets but does not make assumptions of normal distribution (see Dunning, 1993). Keywords thus represent what makes a corpus unique or different, and they can often provide a clear indication of what a set of texts is about. As Scott and Tribble (2006: 55/6) put it, keywords are "[w]hat the text "boils down to" … once we have steamed off the verbiage, the adornment, the blah blah blah". For materials design in ESP contexts, keywords can be more informative than raw word frequency lists, helping to guide us towards core topics or vocabulary that should be included in an ESP syllabus (e.g. Paquot, 2007; Watson Todd, 2017).

The keyness values for all words in SCCERA were calculated using WordSmith Tools, Version 6.0 (Scott, 2012), with appropriate subsections of the Corpus of Contemporary American English (COCA[4]) as the reference corpus. Subsections included were written fiction, magazine and newspaper texts (1990 – 2014). Total size of the COCA reference corpus was 290.4 million tokens. In addition to keyness, a dispersion metric for each word was calculated by counting the number of research articles each word occurs in (out of a total of 1,100). This was done because any given word may be over-represented in a target corpus due to highly frequent use in only one or several articles – e.g. a word, such as a place name, that is unique to the topic of a single article. Such a word may not be of generalisable importance in the civil engineering domain. The identification of keywords of core importance in SCCERA was thus based on two metrics: (i) *keyness* – the degree to which given word is over-represented compared to a benchmark; and (ii) *dispersion* – the range of texts in which that word appears. Although a minimum threshold of ten occurrences was applied, this was effectively superseded by the dispersion threshold (i.e. occurrence in at least 5% of all texts), which meant that that the actual minimum occurrence of any keyword in the corpus is 66.

An additional measure, primarily of pedagogical interest, was applied by comparing keywords to established external word lists. Each word was identified as occurring in (i) the New General Service List (NGSL) (Browne et al., 2013) (the 2801 most frequently used lemmas in the English language), (ii) the New Academic Word List (NAWL) (Browne et al., 2013) (963 lemmas that occur frequently in a range of academic disciplines but not on the NGSL) or (iii) neither of these lists ('off-list'). Manual sorting of keywords into word lemmas (e.g. *model*, *models*, *modelled*, *modelling*) was carried out and total frequencies for the 'set' calculated. Finally, the off-list keywords found across the 11 sub-disciplines of civil engineering were compared and hierarchical agglomerative cluster analysis was performed in order to investigate the degree of homogeneity existing within this discipline. Hierarchical agglomerative cluster analysis is a statistical method for dividing "a set of elements into clusters, or groups, such that the members of one group are very similar to each other and at the same time very dissimilar to members of other groups" (Gries, 2013: 306–7). Percentage coverage figures provided by the NGSL, NAWL and the top 650 off-list keywords were also calculated for each sub-discipline.

2.3. Identification of important lexical bundles in the corpus

Lexical bundles (also known as clusters, formulaic sequences, fixed expressions, or N-grams), such as '*it should be noted that* ~', can be defined as the most frequently occurring lexical sequences in a particular register (Biber et al., 2004). They are a prevalent feature of scientific writing and as such play an important role in the production of texts that conforms to the rhetorical norms of a specific research field (Salazar, 2014). Biber and Barbieri (2007: 269) identify the following five common characteristics of lexical bundles:

i. They are extremely common
ii. They are normally non-idiomatic in meaning
iii. They are not perceptually salient
iv. They tend to be incomplete structural units, preceding or bridging 2 phrases
v. They function as 'discourse frames' for introducing new information, but do not express new propositional meaning themselves

For this investigation, the most frequent 3- to 6-word sequences in the corpus were identified using WordSmith Tools, Version 6.0 (Scott, 2012), and the complete list was narrowed down for more detailed analysis using the commonly accepted cut-off rates of 20–40 times per million words (e.g. Biber & Barbieri, 2007). As clusters often overlap with other sequences, the list was manually sorted to identify 'related clusters' (Scott, 2012) and cleaned up to produce a final list of lexical bundles with potential pedagogic applications. This process is described in more detail in section 3.3.

3. Results & discussion

3.1. Keywords in SCCERA

Comparison of SCCERA against selected subsections of the Corpus of Contemporary American English (COCA) yielded a list of 2967 statistically significant keywords (i.e. log-likelihood value > 15.31, $p > 0.0001$), occurring in at least 5% of all articles and at least three sub-corpora (for a justification of the significance threshold level, see Rayson, Berridge, & Francis, 2004).

[4] Full-text corpus data for COCA is available at: http://corpus.byu.edu/full-text/purchase.asp.

This list was then organised according to externally established word lists within which each word occurs – 1792 (60.4%) words falling within the NGSL, 500 (16.9%) within the NAWL and 675 (22.8%) in neither the NGSL or NAWL, and therefore likely to be particular to the specific civil engineering domains represented in SCCERA.

Table 2 shows the top 50 keyword lemmas, by category (NGSL, NAWL or 'off-list'), ordered according to frequency.[5] The keywords meet our expectations of the 'aboutness' of civil engineering research articles, with a focus on materials – examining or measuring their physical properties and modelling their behaviour in time or space. The top hit in the off-list, et al., reflects the high incidence of co-authored papers in scientific research and also underlines the importance of citations to support claims. Words such as *model*, *fig/figure*, *table*, and *eq/equation* highlight the importance of 'multimodality' in civil engineering reports, where text is combined with other modes (photographs, diagrams, mathematical equations, tables, charts) in order to convey information precisely. This is something which, in our view, distinguishes engineering discourse from other types of professional writing and should be a focus in materials design.

Only a few grammar words occur in the keyword lists, and these provide some indication of the style of writing preferred by civil engineers. The presence of *the* as a keyword suggests a high frequency of nouns in the corpus, as does *of* which is commonly used for post-modification of noun phrases (e.g. the process *of* mixing). The occurrence of *is/are* in the list, along with just a handful of past participle verb forms (e.g. *based*, *observed*, *measured*, *shown*, *used*, *calculated*, *obtained*, *estimated*, *computed*, *simulated*, *defined*) reflect the common use of passive forms in research articles to shift the focus from the actor to the action itself. Finally, the high degree of nominalisation, where noun forms are used in preference to verb forms (e.g. *analysis*, *behaviour*, *distribution*, *measurements*, *displacement*, *simulation*, *response*, *observations*, *deviation*) is another common characteristic of academic writing which allows for increased informational density, as well as shifting the focus from the agent (i.e. the 'doer') to the action itself, in a similar way to passive forms.

3.2. Pedagogical implications

Corpus data of this kind is, of course, useful for ESP material designers in identifying appropriate texts or task types to include in teaching materials and providing optimal conditions for learners to encounter and learn the important language or rhetorical norms of their field. Online searches of Google or Google Scholar using keywords (from the NAWL or 'off-list' categories in particular) allow a convenient shortcut to relevant literature with the desired characteristics. For example, a short extract (of approximately 1000 words) on soil erosion processes from a PhD thesis (Saavedra, 2005: Section 2.2.1) identified in this way was found to include 47 (31.3%) of the keywords shown in Table 2 and also had the multimodal characteristics typical of civil engineering discourse, with a high density of schematic diagrams, photographs and mathematical equations. Task design is usually strongly influenced by the nature of the selected text itself and the discourse features present which can be 'noticed' by learners (e.g. Schmidt, 1990), but it can also be guided by corpus data that is suggestive of the kind of 'work' typically done in the genre analysed. For civil engineers, for example, tasks could usefully focus on measuring and describing the properties of different types of materials, integrating diagrams with text and predicting behaviours under different conditions. Specialised corpora can also be useful for retrospective analysis of a syllabus – texts chosen for inclusion in a course can be checked against frequency lists for coverage of key language and any items not arising naturally added as supplementary materials (see Willis, 2003: 223 for more on this notion of a 'pedagogic corpus').

Direct as well as indirect learning can play a role – research suggests that courses that encourage explicit attention to target lexis, in addition to incidental learning, lead to better results (e.g. Ellis, 2008: 451; Norris & Ortega, 2000: 500). An example of a pedagogical application from keyword analysis is the generation of 'word clouds'[6] to highlight for students the relative importance of particular vocabulary in their discipline. Figure 1 shows a word cloud for 'off-list' keywords in SCCERA, with the size of each word reflecting its 'keyness' value.

3.3. Distribution of off-list keywords across the sub-corpora of SCCERA

In order to assess the degree of homogeneity existing within the different sub-disciplines of civil engineering, the distribution of off-list keywords across the sub-corpora of SCCERA was investigated. Using the specified subsections of COCA as a reference corpus, keywords were identified for each of the eleven sub-disciplines ($p < 0.0001$). Table 3 shows the top twenty off-list keywords for each area ranked by keyness value, with types appearing in more than one sub-discipline highlighted in bold.

Given that civil engineering is widely recognized as a broad field, it is perhaps not surprising to see considerable variation in the off-list keywords between sub-disciplines, with only 35.9% of word families reoccurring in two or more areas. This raises certain questions for ESP instructors in terms of just *how specialized* a corpus of civil engineering research articles needs to be in order to maximize its pedagogic value. For undergraduate students, who require a more general understanding of the discipline, it might make sense to look at keywords from the complete corpus. However, at the post-graduate level, where students typically specialize in a particular area of study, it may be more appropriate for them to focus in on specific parts of

[5] The complete list of keywords from SCCERA is available online at: www.dropbox.com/s/56ksuh9uhpichi8/Keyword%20lemmas%20in%20SCCERA.pdf?dl=0.

[6] A wide variety of word cloud generators are available online. See, for example: http://www.edudemic.com/word-cloud-generators/.

Table 2
Top 50 key word lemmas by list (NGSL, NAWL and ‘off-list’).

NGSL	Freq.	NAWL	Freq.	Off-list	Freq.
the	600,933	parameter	8557	et al	27,743
of	313,669	distribution	7269	fig	20,185
and	256,039	obtain	6765	shear	4763
be	240,015	impact	6105	earthquake	4239
in	188,579	coefficient	5388	eq	4060
for	95,189	velocity	5052	cement	3048
this	63,555	particle	4894	seismic	2615
with	58,380	simulation	4432	coastal	2531
as	57,731	spatial	3993	ash	2040
by	53,633	displacement	3778	stiffness	1979
from	41,595	specimen	3739	groundwater	1779
use	39,271	non	3705	pore	1674
model	34,695	emission	3517	spectral	1595
which	23,197	vertical	3469	atmospheric	1536
can	22,691	dynamic	3448	infrastructure	1512
data	22,444	sediment	3097	hazard	1482
result	20,817	linear	3095	hydraulic	1280
show	19,111	correlation	3057	seasonal	1279
value	18,823	precipitation	2735	drought	1272
between	17,826	magnitude	2709	vibration	1266
time	17,676	algorithm	2696	calibration	1240
high	17,252	prediction	2586	deformation	1231
than	16,107	regression	2573	compressive	1197
also	15,669	numerical	2526	satellite	1177
equation	15,225	simulate	2490	moisture	1161
water	15,157	scenario	2389	analytical	1159
increase	15,128	matrix	2349	hurricane	1129
study	14,866	beam	2314	width	1123
base	14,428	rainfall	2288	sensor	1108
effect	13,603	acceleration	2212	modulus	1091
level	13,275	spectrum	2209	roughness	1067
table	12,585	variability	2175	reinforcement	1061
test	12,581	damp	2126	compression	1059
system	12,528	horizontal	2103	tensile	1054
figure	12,521	empirical	2082	median	1043
large	12,225	vegetation	2027	corrosion	1028
area	12,035	estimation	2020	normalized	1025
change	11,882	accuracy	2005	tunnel	1024
low	11,742	induce	2003	axial	1022
such	11,595	deviation	1969	runoff	984
each	11,484	indicator	1879	nm	982
analysis	11,393	basin	1867	sensors	972
measure	11,303	classification	1819	validation	969
different	11,202	domain	1811	respondents	968
may	11,127	flux	1792	hydration	943
where	11,038	elevation	1773	saturation	923
case	10,986	aggregate	1754	dam	908
project	10,925	intensity	1752	ensemble	905
method	10,884	clay	1703	robust	902
estimate	10,822	optimal	1649	downstream	901

the corpus. Since the research articles making up SCCERA are stored according to sub-discipline, it is easy for users to vary their search range depending on their own needs or interests – the question is where exactly to draw the line? It seems likely that the language of certain sub-disciplines within SCCERA is more similar than others; for example, we might assume considerable overlap between Hydrology, River and Environmental Engineering and Coastal Engineering.

In order to identify meaningful groupings of sub-disciplines we conducted a cluster analysis based on the overlap of keywords. First we extracted all keywords occurring in three or more sub-disciplines and tabulated their occurrence/absence in the sub-disciplines in a binary matrix measuring 2608 × 11. We submitted this matrix to hierarchical agglomerative cluster analysis[7] where the similarity of the sub-disciplines (in the columns) was calculated by the Euclidean distance measure and the clusters were amalgamated using Ward’s method. The resulting dendrogram is shown in Figure 2.

The *y*-axis (height) provides a measure of the degree of closeness (or dissimilarity) between individual data points or clusters, where zero would indicate identical samples. Cutting the dendrogram at level 45, we can see 3 major clusters in

[7] Hierarchical agglomerative cluster analysis builds a ‘hierarchy of clusters’, represented in a dendrogram (or clustering tree), with the vertical axis calibrating the level of clustering. It uses a measure of dissimilarity between sets of observations to decide which clusters are most closely related and should be combined (e.g. Yim & Ramdeen, 2015).

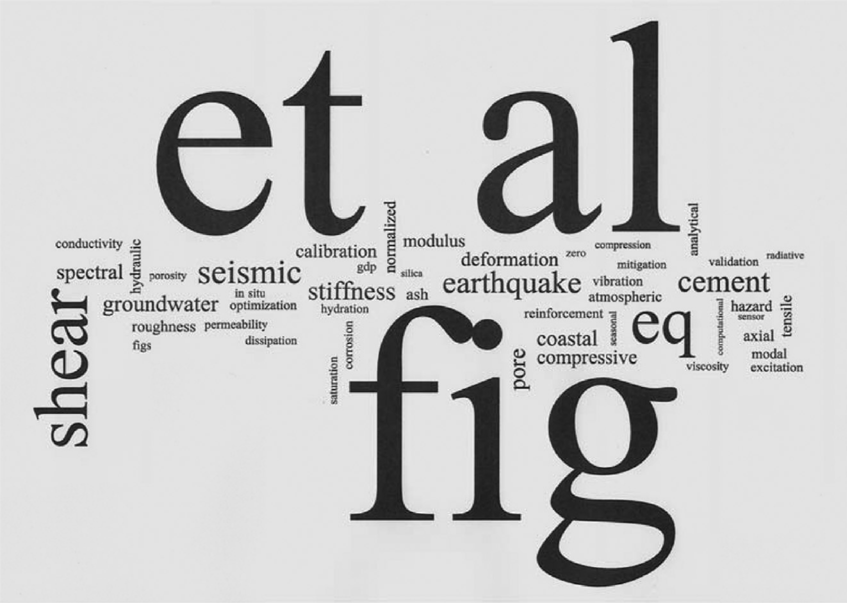


Figure 1. Example word cloud for ‘off-list’ keywords in SCCERA (generated using Wordle).

SCCERA, corresponding roughly to ‘transportation and trade’, ‘water resources’ and ‘structures’. Alternatively, cutting the tree at level 55 gives 2 major clusters, corresponding to the ‘softer’ or ‘harder’ areas of civil engineering. Thus, while the language of civil engineering research articles displays a certain degree of heterogeneity, cluster analysis of keywords in SCCERA provides corpus users with a principled method for varying the specificity of their queries.

As we saw in section 1.2, percentage lexical coverage figures are often calculated from corpus wordlists in order to estimate the vocabulary size necessary for acceptable levels of reading comprehension or incidental learning of vocabulary in the target discourse. The coverage needed for comprehension may vary according to the particular genre under investigation, but is generally estimated to be between 95% and 98% (Hu & Nation, 2000; Laufer, 1989; Nation, 2006; Webb & Rodgers, 2009). Table 4 shows the percentage coverage of civil engineering research articles in the various sub-disciplines provided by the NGSL (2,801 lemmas), the NAWL (963 lemmas), and the top 650 off-list keywords respectively.

As can be seen, percentage coverage rates vary somewhat according to sub-discipline, with the top 650 off-list words providing fewer gains for the ‘softer’ subject areas such as Infrastructure and International Projects. This probably reflects the broader nature of these sub-disciplines, where the topics dealt with by researchers show greater diversity and the specialized vocabulary necessary is therefore less predictable. It should be noted, however, that the ‘softer’ civil engineering areas have better coverage in the NGSL, resulting in similar total percentage coverage figures.

3.4. Pedagogical implications

The NGSL and NAWL combined to provide only an average coverage of just 84.6%. Research on text comprehension would suggest that this level of vocabulary knowledge is not sufficient for students’ fluid reading of research articles in their field. By adding in the top 650 keywords in SCCERA, the average coverage is increased to 92.4% – much closer to the figure of 95% estimated as necessary for comfortable reading comprehension or incidental vocabulary acquisition to occur. Corpus data can therefore be seen to provide ESP instructors with a principled method for the selection of core vocabulary to include in a language syllabus.

3.5. Identification of important lexical bundles in SCCERA

A total of 142,709 3- to 6-word lexical bundles with 5 or more occurrences were identified in SCCERA. This quantity of target phrases would obviously overwhelm students in the language classroom and it was therefore felt necessary to narrow down the list to a more manageable size. Using the established cut-off rates of 20 or 40 occurrences per million words from the literature for analysis of fixed expressions (e.g. Biber et al., 2004) allowed us to reduce the list to between 366 and 1138 phrases, as shown in Figure 3.[8]

[8] Given that types do not follow a linear distribution, as a corpus grows in size it will generate progressively fewer lexical bundles types than would be expected if the growth rate were linear. Biber and Barbieri (2007) discuss how this can cause problems when comparing the number of lexical bundle types across differently sized corpora, but the present study does not involve such cross-corpora comparison of frequency.

Table 3
Top 20 off-list keywords by sub-discipline.

Concrete	Coastal	Disaster	Geotechnical	Hydrology	Infrastructure	Int. Proj.	Mechanics	Reg. Planning	River/Enviro.	Transport
fig	**et al**	**earthquake**	**fig**	**et al**	**et al**	remittances	**fig**	**et al**	**et al**	**et al**
cement	**fig**	**seismic**	**shear**	**fig**	**fig**	**et al**	**et al**	**fig**	**fig**	**fig**
hydration	coastal	**tsunami**	**et al**	ash	pavement	per capita	**shear**	landsat	**eq**	transit
compressive	**eq**	**fig**	**eq**	**groundwater**	**respondents**	openness	**stiffness**	**spectral**	hydraulic	**eq**
silica	tidal	evacuation	**pore**	atmospheric	optimization	**eq**	damper	azimuth	**groundwater**	congestion
corrosion	erosion	**hazard**	**liquefaction**	radiative	**infrastructure**	microfmance	vibration	pixel	**roughness**	accessibility
slag	shoreline	**et al**	suction	runoff	**eq**	dummy	**eq**	segmentation	**shear**	hub
pore	salinity	**shear**	undrained	seasonal	overlay	spillovers	excitation	pixels	floodplain	logit
shrinkage	**shear**	mitigation	triaxial	streamflow	organisational	causality	dampers	reflectance	conductivity	commuting
tensile	dissipation	inundation	compression	**hydrological**	rework	determinants	**axial**	validation	downstream	**respondents**
ferroelectric	runup	**modal**	tunnel	ensemble	stakeholder	macroeconomic	**seismic**	**calibration**	turbulence	freight
mortar	**tsunami**	vulnerability	**deformation**	evaporation	**sustainability**	endogenous	ductility	dataset	upstream	inland
porosity	breakwater	**stiffness**	consolidation	**hydrologie**	contractors	**fig**	**modal**	**wavelet**	**hydrologie**	**infrastructure**
et al	bathymetry	hurricane	**modulus**	**calibration**	organizational	exogenous	**modulus**	classifier	dam	commuters
modulus	arctic	**liquefaction**	**axial**	albedo	**hazard**	estimator	**deformation**	sensor	**calibration**	connectivity
reinforcement	swash	hazus	viscosity	evapotranspiration	**roughness**	**sustainability**	**compressive**	doppler	catchment	**modal**
clinker	tide	**drought**	volumetric	**drought**	deterioration	liberalization	**wavelet**	interpolation	scour	stochastic
ceramics	dune	resilience	figs	anomalies	demolition	cointegration	**earthquake**	suitability	assimilation	congested
dielectric	meltwater	prefecture	permeability	aquifer	contractor	enrollment	**tensile**	imagery	**hydrological**	logistics
sulfate	hydrodynamic	**spectral**	compaction	lidar	rebar	governance	**spectral**	biomass	normalized	queue

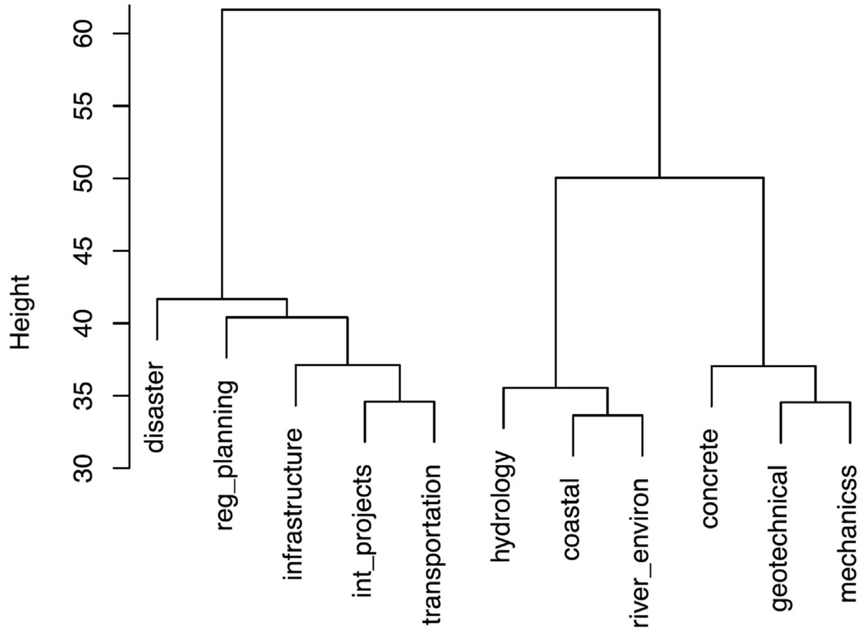


Figure 2. Dendrogram of eleven sub-disciplines in SCCERA clustered according to overlap of keywords.

Table 4
Percentage coverage provided by NGSL, NAWL and the top 650 off-list keywords for each sub-discipline.

Sub-discipline	% coverage			Total % coverage
	NGSL	NAWL	Top 650 off-list keywords	
Coastal	76.4	5.1	9.0	90.5
Concrete	75.7	5.5	10.6	91.8
Disaster	81.2	4.0	8.0	93.2
Geotechnical	77.6	5.4	10.1	93.1
Hydrology	76.1	5.5	8.8	90.4
Infrastructure	85.4	3.7	4.7	93.8
Int. projects	84.8	3.9	4.9	93.6
Mechanics	78.1	6.3	9.0	93.4
Regional planning	80.2	5.0	7.0	92.2
River	76.9	6.0	8.1	91.0
Transport	83.5	4.1	5.7	93.3

As can be seen above, 3-word phrases predominate in the list and it was therefore decided to use the more conservative cut-off rate of 40 occurrences per million words for 3-word bundles, the less strict cut-off rate of 20 occurrences per million for 4- or 5-word bundles, and no cut-off limit for 6-word bundles, reducing the list to 472 items.

A second stage of selection was then necessary in order to exclude the large number of repetitions caused by ‘fragments’ of longer bundles re-occurring in the list. For example, the 6-word bundle *it should be noted that the* generated related 3-word clusters such as *be noted that* and *should be noted* which had to be manually identified and removed. This is a time-consuming process involving individual decisions on each fragment based on KWIC (Key Word in Context) searches of SCCERA using AntConc, in addition to pedagogical considerations. To illustrate, a search for the sequence *should be noted* generated 374 hits in the corpus (Figure 4).

This collocated with *that* to the right in 90.6% of concordance lines and with *it* to the left in 97.3% of cases and it was therefore concluded that the 3-word bundle *should be noted* was typically a fragment of a 5 or 6-word bundle, *it should be noted that (the)* and could be validly removed from the list. However, a similar search in the corpus for the phrase *in the form*, revealed that although it was collocated to the right with *of* in 84.8% of cases, a secondary significant pattern, *in the form + equation (#)*,[9] also existed (Figure 5).

[9] All mathematical equations occurring in articles from SCCERA were replaced with the annotation *equation (#)* since Wordsmith Tools is unable to recognise the symbols used.

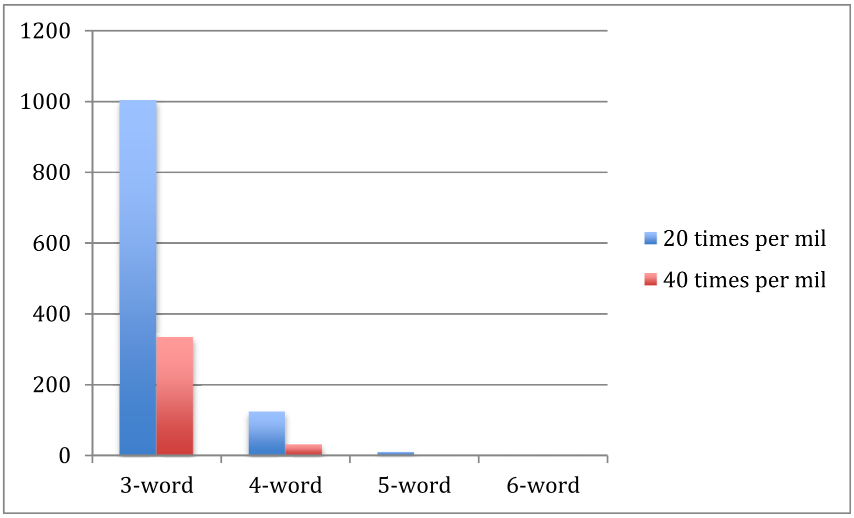
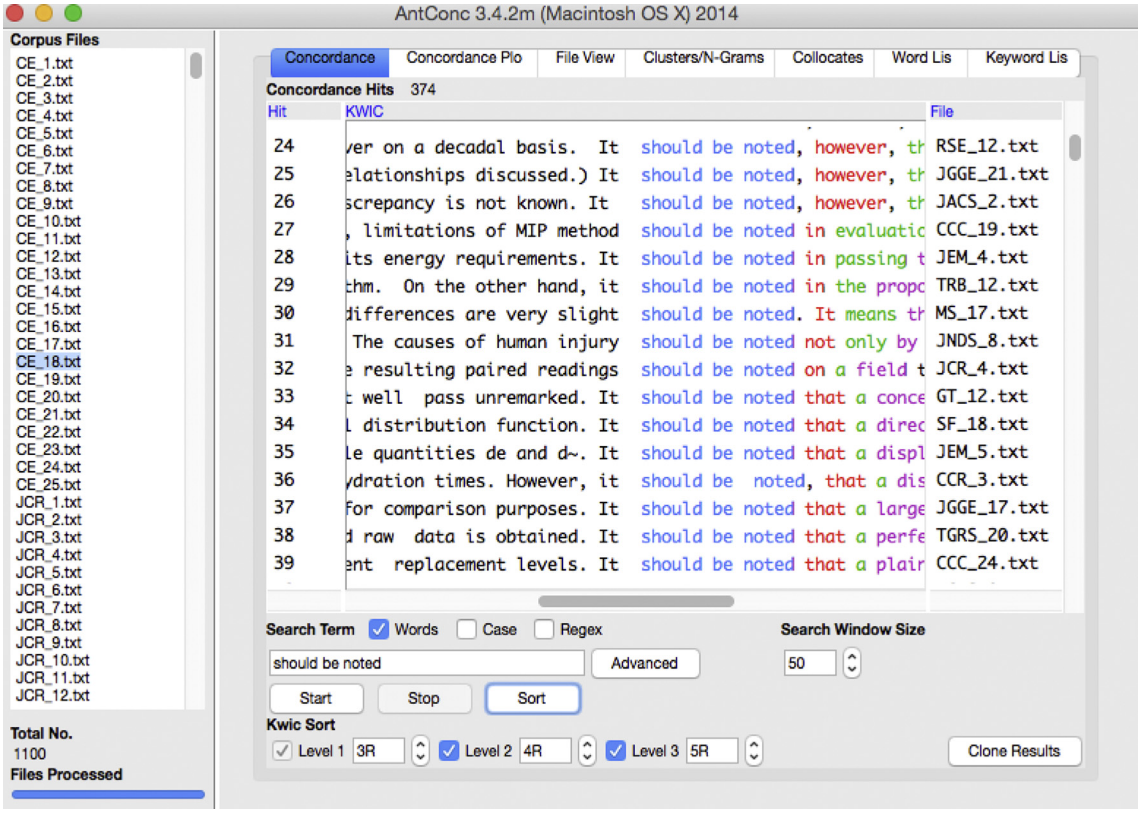


Figure 3. Occurrences of 3- to 6-word lexical bundles in SCCERA with cut-off rates of 20 or 40 times per million words.



```
Corpus Files
CE_1.txt
CE_2.txt
CE_3.txt
CE_4.txt
CE_5.txt
CE_6.txt
CE_7.txt
CE_8.txt
CE_9.txt
CE_10.txt
CE_11.txt
CE_12.txt
CE_13.txt
CE_14.txt
CE_15.txt
CE_16.txt
CE_17.txt
CE_18.txt
CE_19.txt
CE_20.txt
CE_21.txt
CE_22.txt
CE_23.txt
CE_24.txt
CE_25.txt
JCR_1.txt
JCR_2.txt
JCR_3.txt
JCR_4.txt
JCR_5.txt
JCR_6.txt
JCR_7.txt
JCR_8.txt
JCR_9.txt
JCR_10.txt
JCR_11.txt
JCR_12.txt

Total No.
1100
Files Processed
```

```
Concordance Hits   374
Hit   KWIC                                                                        File
24    ver on a decadal basis.  It   should be noted, however, th   RSE_12.txt
25    elationships discussed.) It   should be noted, however, th   JGGE_21.txt
26    screpancy is not known. It    should be noted, however, th   JACS_2.txt
27    , limitations of MIP method   should be noted in evaluatic   CCC_19.txt
28    its energy requirements. It   should be noted in passing t   JEM_4.txt
29    thm.  On the other hand, it   should be noted in the propc   TRB_12.txt
30    differences are very slight   should be noted. It means th   MS_17.txt
31     The causes of human injury   should be noted not only by    JNDS_8.txt
32    e resulting paired readings   should be noted on a field t   JCR_4.txt
33    t well  pass unremarked. It   should be noted that a conce   GT_12.txt
34    l distribution function. It   should be noted that a direc   SF_18.txt
35    le quantities de and d~. It   should be noted that a displ   JEM_5.txt
36    ydration times. However, it   should be  noted, that a dis   CCR_3.txt
37    for comparison purposes. It   should be noted that a large   JGGE_17.txt
38    d raw  data is obtained. It   should be noted that a perfe   TGRS_20.txt
39    ent  replacement levels. It   should be noted that a plair   CCC_24.txt
```

Search Term ☑ Words □ Case □ Regex

should be noted Advanced

Search Window Size 50

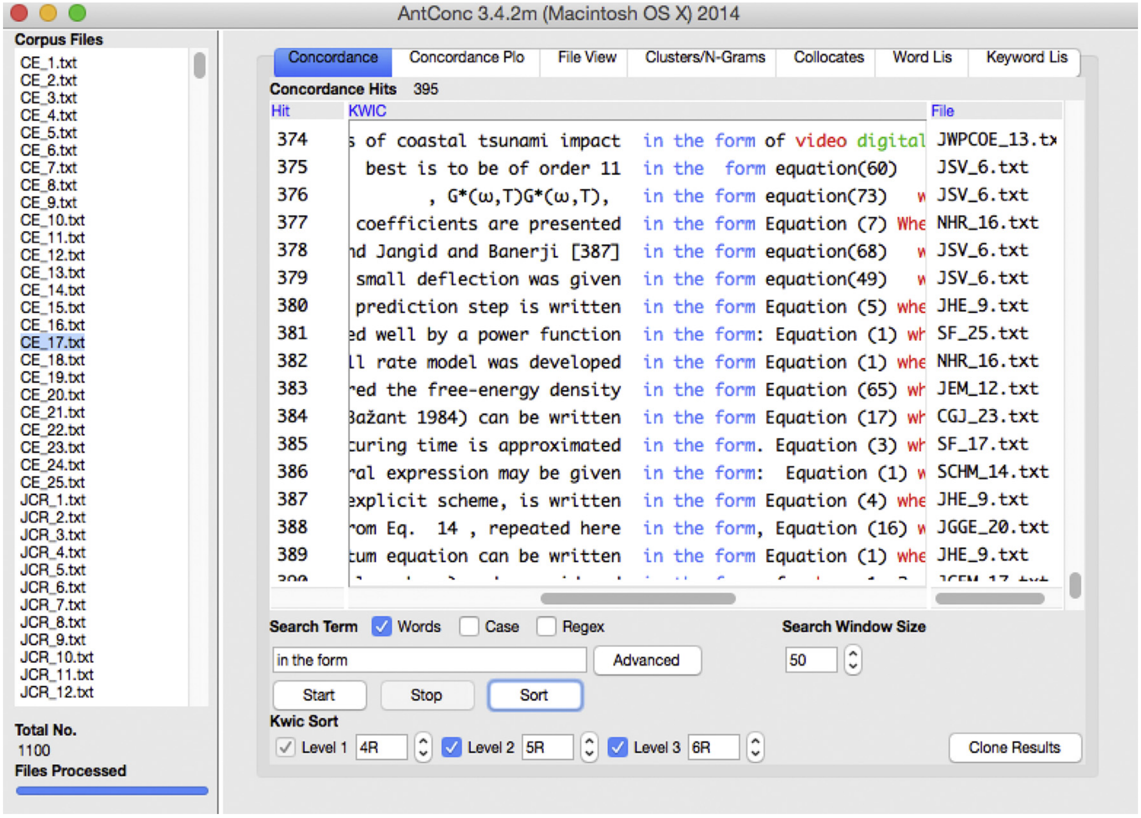
Kwic Sort

☑ Level 1 3R ☑ Level 2 4R ☑ Level 3 5R

Figure 4. Screen shot of right-sorted search for *should be noted* in SCCERA using AntConc.

Given its important deictical role (‘pointing’ to mathematical equations in the text), it was felt that this 3-word bundle performs a useful function, mediating between modes in engineering writing, and it was therefore included in the final list. This illustrates how ultimate decisions on exactly what to include or exclude need to be based on both qualitative and quantitative considerations.

The cleaning up process resulted in a final list of 257 lexical bundles considered to have pedagogic value (6-word bundles = 3; 5-word bundles = 6; 4-word bundles = 90; 3-word bundles = 158). The top hits are shown below in Table 5



Concordance Hits 395

```
Hit  KWIC                                                                     File
374  s of coastal tsunami impact  in the form of video digital                JWPCOE_13.tx
375    best is to be of order 11  in the  form equation(60)                   JSV_6.txt
376          , G*(ω,T)G*(ω,T),    in the form equation(73)   w                JSV_6.txt
377   coefficients are presented  in the form Equation (7) Whe                NHR_16.txt
378  nd Jangid and Banerji [387]  in the form equation(68)   w                JSV_6.txt
379   small deflection was given  in the form equation(49)   w                JSV_6.txt
380   prediction step is written  in the form Equation (5) whe                JHE_9.txt
381  ed well by a power function  in the form: Equation (1) wh                SF_25.txt
382  ll rate model was developed  in the form Equation (1) whe                NHR_16.txt
383  red the free-energy density  in the form Equation (65) wh                JEM_12.txt
384  Bažant 1984) can be written  in the form Equation (17) wh                CGJ_23.txt
385  curing time is approximated  in the form. Equation (3) wh                SF_17.txt
386  ral expression may be given  in the form:  Equation (1) w                SCHM_14.txt
387  explicit scheme, is written  in the form Equation (4) whe                JHE_9.txt
388  rom Eq.  14 , repeated here  in the form, Equation (16) w                JGGE_20.txt
389  tum equation can be written  in the form Equation (1) whe                JHE_9.txt
```

Search Term ☑ Words □ Case □ Regex

in the form

Search Window Size 50

Kwic Sort

☑ Level 1 4R ☑ Level 2 5R ☑ Level 3 6R

Total No. 1100

Files Processed

Figure 5. Screen shot of a right-sorted search for *in the form* in SCCERA using AntConc.

(parentheses denote weaker collocates; slashes denote optional lexical choices). If time is limited, a focus on 4-word bundles in the corpus would seem most efficient: 5- and 6-word bundles are rare while the high frequency of 3-word bundles makes analysis problematic. In addition, they tend to be more opaque in nature and therefore of less pedagogic value.

Frameworks for analysing the functions of lexical bundles vary in the literature, depending on the particular interests or research focus of investigators. Biber's classification (Biber, 2006; Biber et al., 2004), widely adopted by other researchers, identifies three main types: (i) stance expressions; (ii) discourse organizers and (iii) referential expressions. Hyland (2008), on the other hand, prefers to categorize bundles according to whether they are: (i) research-oriented – bundles referring to real-world activities and experiences such as location and quantification (e.g. *in the present study*, *a wide range of*); (ii) text-oriented – bundles managing the organization, meaning and message of the text (e.g. *on the other hand*, *as shown in figure*); or (iii) participant-oriented – bundles focussing on the writer or reader of the text (e.g. *are likely to be*, *it should be noted*).

Hyland's taxonomy is probably of more relevance here since his corpus was generated solely from written academic discourse, whereas Biber's analysis was based on a broader corpus composed of conversation, classroom teaching, textbooks and academic prose (see Biber et al., 2004: 397). The distribution of functions for the 4-word lexical bundles identified in SCCERA are shown below, contrasted with those found by Hyland in his study (2008: 14).

As can be seen in Table 6, a large proportion of the bundles in SCCERA (57.8%) were research-oriented, reflecting the 'real-world' focus of civil engineering research and its descriptions of physical objects, materials, contexts, processes and quantities. Hyland (2008) also found higher levels of research-oriented bundles in his science/technology corpora (see above), although none reached the proportions seen in our corpus. This could be due to a number of reasons: (i) genuine differences between the disciplines; (ii) the use of a narrowed down list in our data affecting distributions; or (iii) inconsistencies in categorization decisions. In practice, categorizing lexical bundles is not always an exact science and there is the possibility of them taking on dual functions or varying in function with context[10]. For example, in SCCERA, the 3-word bundle '*in the same*' can sometimes be classified according to Hyland's taxonomy as 'research-oriented/location' (e.g. *in the same region/department*) and at other times as 'text-oriented/structuring signals' (e.g. *in the same figure/table*), depending on the precise context in which it occurs.

[10] Hyland (2008: 13/4) actually categorises *in the present study* as both a research-oriented and text-oriented lexical bundle.

Table 5
Top hits for 3-, 4-, 5- and 6-word bundles in SCCERA.

Bundle type	Examples	Bundle type	Examples
6-words	*it can be seen that (the)*	3-words	*in order to*
	it should be noted that (the)		*part(s) of the*
	it is important to note that		*a (large) number of*
5-words	*(as) a function of (the)*		*such as the*
	at the end of the		*used in the/this*
	on the other hand (the)		*according to the*
	in the case of (the)		*most of the*
	on the basis of (the)		*because of the*
	(as a) result(s) (of) (the)		*the impact of*
	as shown in (the) fig/figure/table		*the development of*
4-words	*as well as (the)*		*the amount of*
	can be used (to)		*related to (the)*
	with respect to (the)		*associated with (the)*
	in terms of (the)		*some of the*
	the results of (the)		*compared to/with the*
	is/are shown in fig/figure/table		*an/the increase in*
	the size of (the)		*there is no*
	is/are based on a/the		*in which the*
	the effect(s) on/of the		*of/in the model*
	at the same (time)		*change(s) in the*
	in the context of		
	is assumed to be		
	the fact that (the)		
	in the form (of)		
	it is possible to		
	in this paper (we)		
	in addition (to) the		
	for each of (the)		
	the difference between (the)		

Table 6
Distribution of 4-word bundle functions (%): SCCERA compared with disciplines analysed in Hyland (2008).

Discipline	Research-oriented	Text-oriented	Participant-oriented
Civil engineering (SCCERA)	57.8	32.2	10
Biology	48.1	43.5	8.4
Electrical engineering	49.4	40.4	9.2
Applied linguistics	31.2	49.5	18.6
Business studies	36.0	48.4	16.6

3.6. Pedagogical implications

For language learners, taxonomies such as those outlined above are not particularly meaningful or helpful and we therefore sought to organize the 257 key lexical bundles identified in SCCERA into categories that would make more sense for ESP materials designers and civil engineering students. Five types of language function were commonly observed in our list:

i. Language showing **cause–effect relationships** such as:

due to the	*the effect of*	*on the basis of*
as a result of	*as a function of*	*it can be concluded that*

ii. Language of **comparison and contrast** such as:

as well as	*compared to the*	*on the other hand*
at the same time	*the difference between*	*is consistent with*

iii. Language for **quantifying** such as:

part of	*the amount of*	*in the range of*
the magnitude of	*the value of the*	*most of the*

iv. **Deictic language** used to reference time, place or text (often pointing to other ‘modes’; photographs, diagrams, mathematical equations, tables & charts) such as:

in this paper	*the presence of*	*is shown in fig*
is given by equation	*at the end of the*	*can be found in*

v. Language showing the **writer's stance** (the attitude of the writer to the topic or message) such as:

the fact that
it is possible to
according to the
it is important to note
is assumed to be
play an important role in

As Simpson-Vlach and Ellis (2010: 510) point out, this kind of reframing of lexical bundles according to their discourse/pragmatic functions can increase their pedagogic relevance and help bridge the gap between research and practice: "[…] functional linguistic classification and the organization of constructions according to academic needs and purposes is essential in turning a list into something that might usefully inform curriculum or language testing materials".

In terms of teaching formulaic language, Nation (2013: 497) suggests following similar principles to those for isolated words, 'across the four strands of meaning focused input, meaning-focused output, language-focused learning, and fluency development'. Given their prevalence, lexical bundles are likely to be picked up incidentally through reading, but noticing (and therefore acquisition) of formulaic language can be enhanced in texts with the use of underlining, bold font, italics, colour, or glossing. The learning burden is normally quite light since the constituent parts of the phrases are often familiar to students – it is the selection of natural collocation patterns for a particular genre which presents the challenge. Language focused tasks could include gap-fill exercises to elicit formulaic sequences commonly occurring in ESP texts, grouping activities to categorize lexical bundles according to pragmatic function, or examination of concordance lines containing lexical bundles along with guided tasks (Hatami, 2015; Jones & Haywood, 2004; Nation, 2013). Outside of the classroom, lexical bundles can be a useful reference resource for students, particularly when writing up their research for publication. This is the kind of approach adopted by the Academic Phrasebank at Manchester University, where functional taxonomies such as 'being cautious', 'describing quantities' and 'explaining causality' allow writers to quickly identify phrases relevant to their needs (http://www.phrasebank.manchester.ac.uk/), and also by Simpson-Vlach and Ellis (2010) in their classification of the Academic Formulas List (AFL) according to discourse/pragmatic function.

In summary, information on important lexical bundles, combined with keywords lists, can help guide the production of discipline-specific language learning materials that effectively address the needs of civil engineering students. They allow a more principled approach to ESP course design than was possible in the past, when authors tended to rely predominantly on their own intuitions in selecting language content.

3.7. Limitations of the study

By including an analysis of both keywords and fixed expressions in our corpus investigations, we argue that corpus tools make it possible to produce a far more comprehensive and pedagogically useful description of a particular discipline. As a methodology for ESP material writers, this approach is however, limited by the tools. For example, lexical profiling tools such as *Range* (Heatley, Nation, & Coxhead, 2002) currently only analyze text for individual words and could be further improved by taking into consideration word bundles, as suggested by Martinez and Schmitt (2012). ESP course designers could then systematically test their materials to ensure that all the core language for a specialized field of study had been covered.

The methodology employed in this study generated a list of 675 off-list keywords of potential value to civil engineering students, which still represents a considerable learning load. This list of target items could be further narrowed down by identifying words with opaque meanings for special attention in the class, as suggested by Watson Todd (2017). The lack of perceptual salience of lexical bundles also presents a challenge for ESP teachers since this characteristic makes them extremely difficult to learn (Wood & Appel, 2014: 2).

4. Conclusion

This paper has demonstrated how a corpus-based approach to the discourse of civil engineering research articles can provide useful insights into the language patterns typically used by civil engineers. These patterns are important because they help to identify writers as expert members of their discourse community, but without access to corpus data they tend to go unnoticed. Keywords analysis and cluster analysis are complementary in many ways – keyword lists highlight the propositional content which typifies civil engineering texts, while word bundles 'frame' that content: expressing the writer's stance, clarifying the discourse organization or performing a deictic role (making time, place or text references).

Although space here precludes any detailed discussion of more *direct* uses of specialized corpora in the language classroom, approaches such as data-driven learning (e.g. Johns, 1991) which encourage learners to discover language patterns in concordance lines for themselves, inductively, also have great potential. At the University of Tokyo, civil engineering students are taught how to test their own hypotheses and discover rules independently, using SCCERA with the free corpus analysis software, AntConc – tools which can be particularly useful when they are writing up their own research for publication.

Given the paucity of discipline-specific materials in ESP, specialized corpora offer ESP teachers a principled methodology for the design of language syllabuses that meet the needs of their learners. We hope that this work can provide a framework for other ESP specialists wishing to exploit specialized corpora in their own particular contexts.

Acknowledgement

This work was supported by the Japan Society for the Promotion of Science under Grant No. 25370619.

Appendix 1. Summary of composition of SCCERA.

Department	Journal name (Impact Factor, 2013)	No. articles	Tokens	Types
Coastal Engineering	J. of Geophysical Research: Oceans (3.174)	25	233,115	9,822
	J. of Coastal Research (0.496)	25	178,947	12,895
	Coastal Engineering (2.239)	25	187,553	7,529
	J. of Waterway Port Coastal & Ocean Engineering (1.0)	25	158,952	7,044
Sub-total		100	758,567	21,287
Infra-structure Development	J. of Construction Engineering & Management (0.876)	34	186,040	8,888
	J. of Infrastructure Systems (ASCE) (0.983)	33	199,329	9,544
	Int. J. of Project Management (1.686)	33	193,470	9,038
Sub-total		100	578,839	16,463
Concrete Engineering	Cement & Concrete Research (3.112)	25	189,005	8,559
	Cement & Concrete Composites (2.523)	25	133,371	6,849
	J. of American Ceramic Society (2.107)	25	234,668	10,976
	Material & Structures (1.184)	25	122,665	6,296
Sub-total		100	679,709	17,311
Regional Planning, Surveying, Remote Sensing	ISPRS J. of Photogrammetry & Remote Sensing (3.313)	20	126,906	7,501
	Remote Sensing of Environment (5.103)	20	148,900	7,563
	IEEE Transactions on Geoscience and Remote Sensing (3.467)	20	109,421	5,470
	Int. J. of Geographical Information Science (1.613)	20	131,077	7,585
	J. of Regional Science (2.279)	20	140,941	7,953
Sub-total		100	657,245	17,729
River & Environmental Engineering	ASCE J. of Hydraulic Engineering (1.276)	25	148,063	7,066
	Water Resources Research (3.149)	25	213,346	9,322
	J. of Hydraulic Research (1.037)	25	160,699	9,103
	River Research and Applications (2.425)	25	176,174	10,571
Sub-total		100	698,282	20,082
Hydrology & Water Resources Engineering	Journal of Geophysical Research: Atmospheres (3.174)	25	358,298	10,974
	Journal of Hydrology (2.964)	25	185,736	9,108
	Journal of Climate (4.362)	25	179,240	7,186
	Hydrological Processes (2.497)	25	165,606	9,321
Sub-total		100	888,880	19,977
Geotechnical Engineering	Canadian Geotechnical Journal (0.811)	25	154,240	7,522
	ASCE J. of Geotechnical and Geoenvironmental Engineering (1.156)	25	175,391	8,069
	Soils and Foundations (0.413)	25	163,970	6,649
	Geotechnique (1.481)	25	141,978	7,035
Sub-total		100	635,579	15,182
International Projects	World Development (1.527)	50	428,527	16,552
	Journal of Development Economics (2.353)	50	437,594	12,115
Sub-total		100	866,121	20,687
Transportation Research	Transportation Research (Part B – Methodological) (2.944)	34	271,318	9,944
	Journal of Transport Geography (1.942)	33	220,017	10,448
	Transportation Research (Part A – Policy & Practice) (2.725)	33	243,518	10,834
Sub-total		100	734,853	18,634
Mechanics & Structures	Journal of Engineering Mechanics (ASCE) (1.116)	14	147,752	8,066
	Journal of Structural Engineering (ASCE) (1.206)	14	67,296	4,671
	Journal of Sound and Vibration (1.613)	14	120,636	6,474
	Journal of Earthquake Engineering (0.661)	15	128,223	5,979
	Engineering Structures (1.713)	15	81,999	4,664
	Structural Control and Health Monitoring (1.544)	14	81,720	4,887
	Journal of Bridge Engineering (0.793)	14	79,699	5,135
Sub-total		100	707,325	16,092
Earthquake & Disaster Mitigation	Journal of Disaster Research (no IF)	20	102,092	6,965
	Journal of Natural Disaster Science (no IF)	20	98,756	7,134
	Natural Hazard Review (0.78)	20	130,249	7,784
	Earthquake Engineering and Structural Dynamics (1.898)	20	142,741	6,412
	Earthquake Spectra (1.079)	20	127,193	8,125
Sub-total		100	601,031	17,134
Total corpus		1,100	7,806,431	78,322[a]

[a] The total number of types for the complete corpus is less than the sum of the figures for the 45 journals since many words reoccur throughout the texts sampled.

References

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics, 24*(4), 425-438. https://doi.org/10.1093/applin/24.4.425.

Altenberg, B. (1993). Recurrent verb-complement constructions in the London-Lund Corpus. In *English language corpora: Design, analysis and exploitation* (pp. 227-245). Amsterdam: Rodopi.

Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.

Baker, P., Hardie, A., & McEnery, T. (2006). *A glossary of corpus linguistics*. Edinburgh: Edinburgh University Press.

Bennett, G. (2010). *Using corpora in the language learning classroom: Corpus linguistics for teachers*. Ann Arbor MI: University of Michigan Press.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*, 263-286. https://doi.org/10.1016/j.esp.2006.08.003.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at …: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371-405. https://doi.org/10.1093/applin/25.3.371.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman grammar of spoken and written English* (Vol. 2). Harlow, Essex: Pearson.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research, 10*(3), 245-261. https://doi.org/10.1191/1362168806lr195oa.

Boulton, A. (2012). Corpus consultation for ESP: A review of empirical research. In A. Boulton, S. Carter-Thomas, & E. Rowley-Jolivet (Eds.), *Corpus-informed research and learning in ESP* (pp. 261-291). Amsterdam: John Benjamins.

Brezina, V., & Gablasova, D. (2013). Is there a core general vocabulary? Introducing the new general Service list. *Applied Linguistics, 36*(1), 1-22. https://doi.org/10.1093/applin/amt018.

Browne, C., Culligan, B., & Phillips, J. (2013). *The new general service list*. http://www.newgeneralservicelist.org. (Accessed 29 January 2018).

Chen, Y., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology, 14*(2), 30-49.

Chiba, Y., Millar, N., & Budgell, B. (2010). The language of midwifery and perinatal care. *Journal of Japan Academy of Midwifery, 24*(1), 1-10. https://doi.org/10.3418/jjam.24.74.

Chung, T. M., & Nation, P. (2004). Identifying technical vocabulary. *System, 32*(2), 251-263. https://doi.org/10.1016/j.system.2003.11.008.

Civil engineering project.(2017). http://www.cewriting.org/558328b9e4b0c27daf07faba/. (Accessed 29 January 2018).

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics, 29*(1), 72-89. https://doi.org/10.1093/applin/amm022.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238. https://doi.org/10.2307/3587951.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*(1), 61-74.

Ellis, R. (2008). *The study of second language acquisition*. Oxford: Oxford University Press.

Gavioli, L. (2006). *Exploring corpora for ESP learning*. Amsterdam: John Benjamins.

Gilmore, A. (2015). Research into practice: The influence of discourse studies on language descriptions and task design in published ELT materials. *Language Teaching, 48*(4), 506-530. https://doi.org/10.1017/S0261444815000269.

Gries, S. T. (2013). *Statistics for linguistics with R: A practical introduction*. Walter de Gruyter.

Hatami, S. (2015). Teaching formulaic sequences in the ESL classroom. *TESOL Journal, 6*(1), 112-129. https://doi.org/10.1002/tesj.143.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). *Range [computer software]*. http://www.victoria.ac.nz/lals/resources/range. (Accessed 29 January 2018).

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.

Hsu, W. (2014). Measuring the vocabulary load of engineering textbooks for EFL undergraduates. *English for Specific Purposes, 33*, 54-65. https://doi.org/10.1016/j.esp.2013.07.001.

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403-430.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*, 4-21. https://doi.org/10.1016/j.esp.2007.06.001.

Hyland, K., & Tse, P. (2007). Is there an “academic vocabulary”? *TESOL Quarterly, 41*(2), 235-253. https://doi.org/10.1002/j.1545-7249.2007.tb00058.x.

Johns, T. (1991). Should you be persuaded – two samples of data-driven learning materials. In T. Johns, & P. King (Eds.), *‘Classroom concordancing’. ELR Journal* (vol. 4, pp. 1-16).

Jones, M., & Haywood, S. (2004). Facilitating the acquisition of formulaic sequences: An exploratory study in an EAP context. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 269-300). Philadelphia, PA: John Benjamins.

Kennedy, G. (1998). *An introduction to corpus linguistics*. New York: Longman.

Kjellmer, G. (1994). *A dictionary of English collocations: Based on the Brown Corpus*. Oxford: Clarendon Press.

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren, & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316-323). Clevedon, UK: Multilingual Matters.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. J. L. Arnaud (Ed.), *Vocabulary and applied linguistics* (pp. 126-132). London: Palgrave Macmillan.

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics, 33*(3), 299-320. https://doi.org/10.1093/applin/ams010.

Millar, N. (2011). The processing of malformed formulaic language. *Applied Linguistics, 32*(2), 129-148. https://doi.org/10.1093/applin/amq035.

Mudraya, O. (2006). Engineering English: A lexical frequency instructional model. *English for Specific Purposes, 25*, 235-256. https://doi.org/10.1016/j.esp.2005.05.002.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*(1), 59-82. https://doi.org/10.3138/cmlr.63.1.59.

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge: Cambridge University Press.

Norris, J., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning, 50*(3), 417-528. https://doi.org/10.1111/0023-8333.00136.

Paquot, M. (2007). Towards a productively-oriented academic word list. In J. Walinski, K. Kredens, & S. Gozdz-Roszkowski (Eds.), *Practical applications in language and computers 2005* (pp. 127-140). Frankfurt: Peter Lang.

Pearson, J. (1998). *Terms in context*. Amsterdam: John Benjamins.

Rayson, P., Berridge, D., & Francis, B. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. In *Paper presented at the 7th International Conference on Statistical Analysis of Textual Data, 10–12 March 2004, Louvain-la-Neuve*.

Rea Rizzo, C. (2010). Getting on with corpus compilation: From theory to practice. *ESP World, 1*(9), 1-22.

Römer, U. (2010). Using general and specialized corpora in English language teaching: Past, present and future. In M. C. Campoy-Cubillo, B. Bellés-Fortuño, & M. L. Gea-Valor (Eds.), *Corpus- based approaches to English language teaching* (pp. 18-35). London: Continuum.

Rundell, M., & Stock, P. (1992). The corpus revolution. 3-part article. *English Today, 8*(2). https://doi.org/10.1017/S0266078400006520, 8(3); 8(4), 9–17; 31, 21-38; 32, 45-51.

Saavedra, C. (2005). *Estimating spatial patterns of soil erosion and deposition in the Andean region using geo-information techniques: A case study in Cochabamba, Bolivia* (Unpublished doctoral dissertation). Wageningen, the Netherlands: Wageningen University.

Salazar, D. (2014). *Lexical bundles in native and non-native scientific writing*. Amsterdam: John Benjamins.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129-152. https://doi.org/10.1093/applin/11.2.129.

Schonell, F. J., Meddleton, I. G., & Shaw, B. A. (1956). *A study of the oral vocabulary of adults*. Brisbane: University of Queensland Press.

Scott, M. (2012). *WordSmith tools version 6, Stroud: Lexical analysis software*. http://www.lexically.net/wordsmith/version6/. (Accessed 29 January 2018).

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*(4), 487-512. https://doi.org/10.1093/applin/amp058.

Wang, J., Liang, S. L., & Ge, G. C. (2008). Establishment of a medical academic word list. *English for Specific Purposes, 27*(4), 442-458. https://doi.org/10.1016/j.esp.2008.05.003.

Ward, J. (1999). How large a vocabulary do EAP engineering students need? *Reading in a Foreign Language, 12*(2), 309-324.

Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes, 28*(3), 170-182. https://doi.org/10.1016/j.esp.2009.04.001.

Watson Todd, R. (2017). An opaque engineering word list: Which words should a teacher focus on? *English for Specific Purposes, 45*, 31-39. https://doi.org/10.1016/j.esp.2016.08.003.

Webb, S., & Rodgers, M. P. H. (2009). Vocabulary demands of television programs. *Language Learning, 59*(2), 335-366. https://doi.org/10.1111/j.1467-9922.2009.00509.x.

West, M. P. (1953). *A general service list of English words: With semantic frequencies and a supplementary word-list for the writing of popular science and technology*. London: Longman Green.

Willis, D. (2003). *Rules, patterns and words*. Cambridge: Cambridge University Press.

Wood, D. C., & Appel, R. (2014). Multiword constructions in first year business and engineering university textbooks and EAP textbooks. *Journal of English for Academic Purposes, 15*, 1-13. https://doi.org/10.1016/j.jeap.2014.03.002.

Yang, M. N. (2015). A nursing academic word list. *English for Specific Purposes, 37*, 27-38. https://doi.org/10.1016/j.esp.2014.05.003.

Yim, O., & Ramdeen, K. T. (2015). Hierarchical cluster analysis: Comparison of three linkage measures and application to psychological data. *The Quantitative Methods for Psychology, 11*(1), 8-21. https://doi.org/10.20982/tqmp.11.1.p008.

Alex Gilmore has been teaching English as a Foreign Language for over 20 years in England, Spain, Mexico, Saudi Arabia and Japan. He has a MA in English Language Teaching and PhD in Applied Linguistics from the University of Nottingham, UK, and has published widely in the fields of language pedagogy and applied linguistics.

Neil Millar teaches technical writing at the University of Tsukuba in Japan. His research focuses on the language learning and the practical applications of corpus linguistics.