Developing a Specialized Corpus of Civil Engineering Research Articles (SCCERA)

Abstract

A corpus is a large, principled collection of naturally occurring text, stored electronically and used in the descriptive analysis of a language. Whilst 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 analysis of the language used in specific academic or professional contexts. This report describes the development of the Specialized Corpus of Civil Engineering Research Articles (SCCERA), and the potential role it can play in helping students and staff in civil engineering departments (where members are non-native speakers of English) to identify key vocabulary or language patterns in their field, and to write up their research in a natural, discipline-specific manner. It also offers a useful framework for other academics keen to develop their own specialized corpora.

Keywords: Corpus design; corpus analysis; specialized corpora; civil engineering research articles; materials design; English for Specific Purposes

1. Introduction

Language corpora have been available to the linguistics community since the mid-1960s when the one million-word Brown Corpus of American English was originally constructed. The first 'corpus-informed' dictionary (the American Heritage Dictionary) quickly followed, and today, all of the dictionaries produced by the major publishers, as well as many grammar reference books (e.g. Sinclair 1990; Carter & McCarthy 2006), are based on large general corpora (Kennedy 1998). The pedagogical value of corpora lies in their ability to show us how language is really used in specific discourse communities; traditionally, educational materials for language teaching have tended to rely heavily on native-speaker intuitions, which are notoriously unreliable and therefore run the risk of providing us with a distorted view of the target language (Wolfson 1989; Biber, Conrad & Reppen 1998).

Whilst 'mega-corpora', like the BNC or COCA[i], 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 analysis of the language used in specific academic or professional contexts. Large variability has been found to

exist between different disciplines in terms of word frequencies, collocational patterns[ii] and rhetorical moves. For example, Hyland (2008), comparing 4-word lexical bundles[iii] from the fields of Biology, Electrical Engineering, Applied Linguistics and Business Studies, calculated that over 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. He points out that it is the use of this kind of genre-specific language that identify writers as expert members of their own particular discourse community. Given the wide discrepancies in the linguistic characteristics of different academic disciplines, it would seem sensible to use specialized corpora as the starting point in the design of English for Specific Purposes (ESP) materials. ESP teachers are in particular need of support since they are often neglected by international publishers, who tend to focus their attention on more generic language learning materials where the financial rewards are higher (Boulton 2012).

This paper describes the development of the Specialized Corpus of Civil Engineering Research Articles (SCCERA), and discusses the potential insights it can offer into the rhetorical practices of professional civil engineers around the world. It is hoped that it can also provide a framework for other ESP specialists wishing to develop or exploit specialized corpora in their own fields.

2. Method

SCCERA was designed and built over a period of two years at the University of Tokyo, supported by funding from the Japanese Society for Promotion of Science (JSPS). There were four main phases to the project: planning, construction, preliminary analysis, and consideration of pedagogic applications.

2.1 Phase 1: Planning a balanced & representative corpus

The initial phase of the project involved designing a specialized corpus that would be seen as both *balanced* and *representative* by the target users: 'balance' here means inclusion of all of the various sub-disciplines of civil engineering in roughly equal

proportions, while 'representative' means that the final corpus is a fair reflection of the genre it claims to represent – in this case, civil engineering research articles. Corpus linguists and academic staff in the Department of Civil Engineering at the University of Tokyo were consulted on the make-up of SCCERA and it was decided that the research articles (RAs) selected for inclusion would be:

i. Peer-reviewed papers from influential journals, preferably cited in the Science Citation Index Expanded (SCI®) or Social Sciences Citation Index (SSCI®).
ii. Taken from journals widely read and respected by academic staff in the Department of Civil Engineering and considered to be 'key' or 'desired outlets for academic work'.
iii. Representative of the 11 main sub-disciplines in the Department of Civil Engineering at the University of Tokyo.
iv. 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.
v. Articles listed as 'most cited' or 'most viewed' by the publishers (where information was available on the publisher's website).
vi. Sufficient in number to create a final corpus of at least one million words - the minimum size recommended for specialized corpora (Kennedy 1998; Pearson 1998; Rea Rizzo 2010).

2.2 Phase 2: Construction of SCCERA

One hundred articles from the key journals identified for each of the 11 sub-disciplines of civil engineering were selected for inclusion in the corpus and downloaded as PDF files in order to provide samples of the original articles for reference purposes. Corpus analysis software normally requires plain text files ('.txt' extension) for processing so all the data was copy-pasted into MS-Word files and saved initially as Word documents (for the cleaning up stage) and then plain text files, allowing line breaks and character substitution. Where available online, HTML versions of the articles were used for copy-pasting in preference to the original PDF files, since this simplifies the time-consuming process of cleaning up the data. All extraneous information (such as references, tables and figures, mathematical equations and HTML fragments) was removed and the remaining text cleaned up in

preparation for analysis. This work was carried out by a trained research assistant over a period of 32 days and took a total of 138 hours for the complete corpus of 1,100 articles, with an average processing time of around 7.5 minutes per article. As Figure 1 below shows, the research assistant was able to quickly speed up this process over the first week as he became more familiar with the tools and methodology, although there were still variations in the time necessary, depending on whether he was working with HTML or PDF documents.

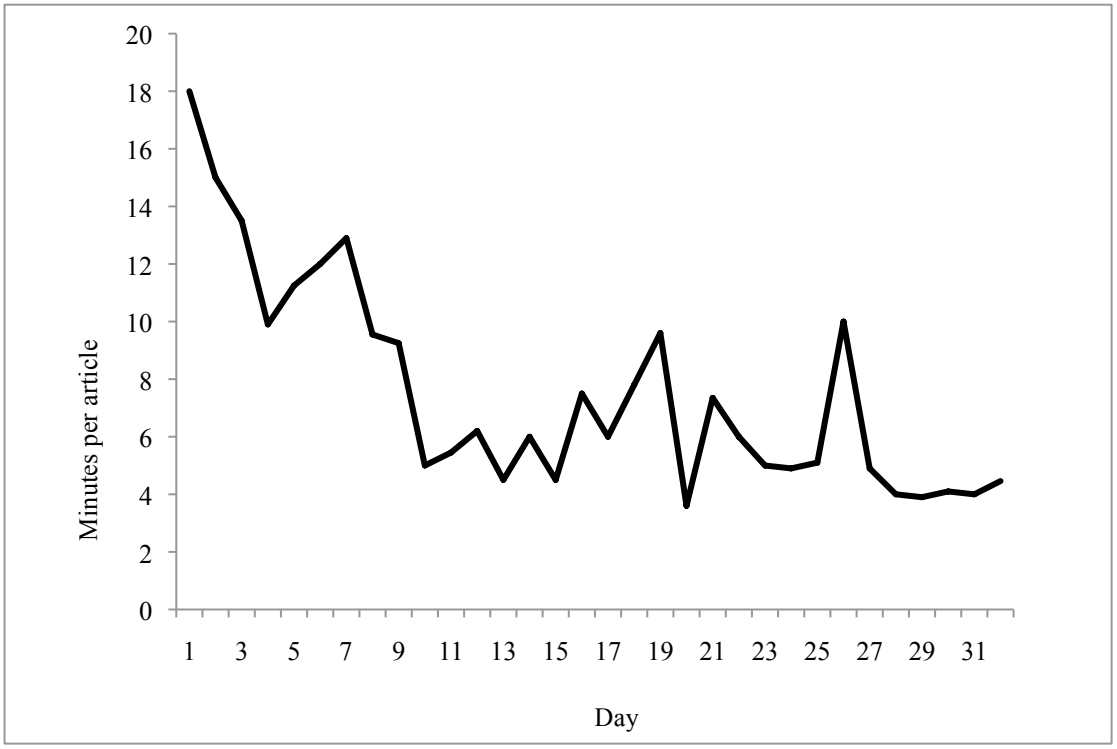
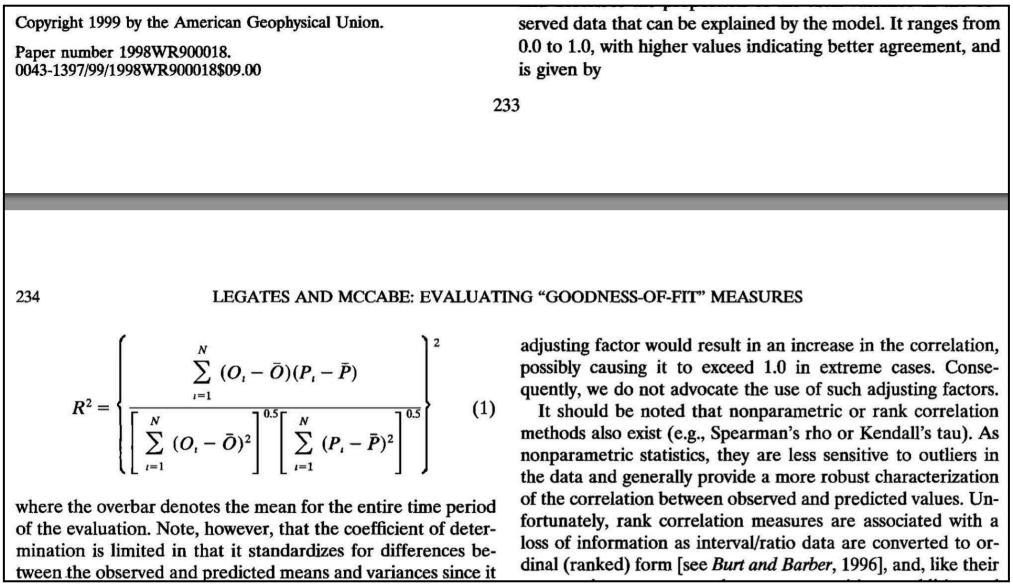


Figure 1: Processing time (minutes per article) for the cleaning-up process

Time-consuming work in the cleaning-up process included:

i. Removing hyphens from words split at the ends of lines in the original article
ii. Deleting page numbers and footnotes breaking up the flow of the main text
iii. Rearranging blocks of text where double columns in the original PDF document have not been recognized by Word
iv. Deleting mathematical symbols not recognized by Word and replacing them with a tag (Equation 1, etc.)
v. Checking the final text for spelling and grammar errors produced in the copy-paste process (underlined in red or green)

Figure 2 illustrates one example of this process, where the resulting Word text produced from a journal article (Legates & McCabe 1999) requires considerable revision. Here, from the perspective of corpus analysis, a valuable expression for describing mathematical equations ([…] *and is given by (Equation1) where x denotes*…) has been lost in the copy-paste process because of the presence of footnotes, headers, page numbers, mathematical symbols and double columns. This illustrates the importance of this stage for the quality of the resulting corpus and the usefulness of the data that can be extracted from it.



Copyright 1999 by the American Geophysical Union.

Paper number 1998WR900018.
0043-1397/99/1998WR900018$09.00

served data that can be explained by the model. It ranges from 0.0 to 1.0, with higher values indicating better agreement, and is given by

233

234 LEGATES AND MCCABE: EVALUATING "GOODNESS-OF-FIT" MEASURES

$$R^2 = \left\{ \frac{\sum_{i=1}^{N} (O_i - \bar{O})(P_i - \bar{P})}{\left[\sum_{i=1}^{N} (O_i - \bar{O})^2\right]^{0.5} \left[\sum_{i=1}^{N} (P_i - \bar{P})^2\right]^{0.5}} \right\}^2 \quad (1)$$

where the overbar denotes the mean for the entire time period of the evaluation. Note, however, that the coefficient of determination is limited in that it standardizes for differences between the observed and predicted means and variances since it

adjusting factor would result in an increase in the correlation, possibly causing it to exceed 1.0 in extreme cases. Consequently, we do not advocate the use of such adjusting factors.

It should be noted that nonparametric or rank correlation methods also exist (e.g., Spearman's rho or Kendall's tau). As nonparametric statistics, they are less sensitive to outliers in the data and generally provide a more robust characterization of the correlation between observed and predicted values. Unfortunately, rank correlation measures are associated with a loss of information as interval/ratio data are converted to ordinal (ranked) form [see *Burt and Barber*, 1996], and, like their

and describes the proportion of the total variance in the ob- served data that can be explained by the model. It ranges from 0.0 to 1.0, with higher values indicating better agreement, and is given by

233

)

234

LEGATES AND MCCABE: EVALUATING "GOODNESS-OF-FIT" MEASURES

2 R=

2 (o,- o)(p,-p) 1=1|

N

'

adjusting factor would result in an increase in the correlation, possibly causing it to exceed 1.0 in extreme cases. Conse- quently, we do not advocate the use of such adjusting factors.

It should be noted that nonparametric or rank correlation methods also exist (e.g., Spearman's rho or Kendall's tau). As nonparametric statistics, they are less sensitive to outliers in the data and generally provide a more robust characterization of the correlation between observed and predicted values. Un- fortunately, rank correlation measures are associated with a loss of information as interval/ratio data are converted to or- dinal (ranked) form [see *Burt and Barber,* 1996], and, like their parametric counterparts, they are not sensitive to additive and proportional differences between the observed and model- simulated values.

2.2. Coefficient of Efficiency *E*

The coefficient of efficiency *E* has been widely used to eval- uate the performance of hydrologic models [e.g., *Leavesley et al,* 1983; *Wdcox et al,* 1990]. *Nash and Sutcliffe* [1970] denned the coefficient of efficiency which ranges from minus infinity to 1.0, with higher values indicating better agreement, as

Figure 2: Example of the cleaning-up process (original PDF article above converted to Word document below)

The final versions of the plain text files were compared with the original PDF documents to ensure consistency and part-of-speech annotated ('POS tagged') using CLAWS 4 software[iv]. Finally, the annotated files were saved in folders according to journal name and sub-discipline, in order to facilitate storage and retrieval in the completed corpus.

2.3 Phase 3: Preliminary analysis of SCCERA

Preliminary analysis of SCCERA was carried out using WordSmith Tools, Version 6.0 (Scott 2012), with a range of lexico-grammatical features investigated across the different sub-sections of the corpus:

i. Frequency lists: Provide a rank ordering of all the vocabulary occurring in a corpus, in order of frequency. This is pedagogically useful because it helps us to identify the core vocabulary used in civil engineering research articles, which should be familiar to students.
ii. Keywords: Keyword analysis highlights words "whose frequency is unusually high in comparison with some norm" (Scott 2012: 176). The study corpus is compared to a larger 'reference corpus' (here, the British National Corpus), which helps us to characterize the genre and, in this case, identify what civil engineering texts are usually 'about'.
iii. Cluster analysis: 'Chunks' of language, or 'lexical bundles', can be just as important as individual words and, as mentioned in section 1, have been shown to be quite discipline-specific in the research literature. For this reason, an investigation of the most common 3-, 4-, 5- or 6-word combinations in SCCERA can also help civil engineers to write up their research appropriately.
iv. Concordance lines: Concordance lines are samples of text recovered from the corpus, showing the 'lexical or grammatical environment' around a particular query item. By searching for words of interest in the corpus, we can see how they are commonly used in civil engineering RAs.
v. Part-of-speech: POS tagging of SCCERA allows the relative proportions of different parts-of-speech to be calculated and compared against other reference corpora in order to evaluate the characteristics of POS classes in civil engineering writing.

2.4 Phase 4: Investigation of potential pedagogic applications for SCCERA

Based on the preliminary analysis of SCCERA, potential pedagogic applications of the corpus were evaluated in collaboration with faculty members from the Department of Civil Engineering at the University of Tokyo. As an academic resource for non-native speakers (NNSs) of English, two possible approaches were considered:

i. An indirect, 'corpus-informed' approach, facilitating production of language learning materials designed specifically for civil engineers.
ii. A more direct approach, involving training students or staff to query the corpus themselves in order to find answers to specific questions they have connected to their academic writing in English.

Due to space limitations, only the first approach will be discussed here.

3. Results & Discussion

The principal characteristics of SCCERA can be summarized as follows:

- Total size: approx. 8 million words
- 11 sub-corpora (representing the different sub-disciplines of civil engineering at the University of Tokyo)
- Sourced from 45 international journals[v], considered 'key' by members of the Department of Civil Engineering (see Appendix 1)
- 1,100 research articles (most cited/downloaded) published between 1989 and 2014
- 3,807 contributing authors from 1,598 institutions in 80 countries:

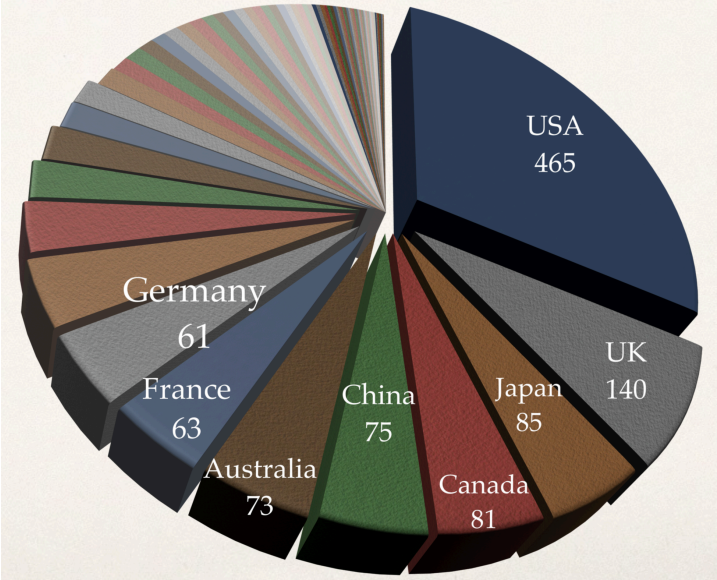


Figure 3: Contributing authors by institution country

3.1 Word frequencies in SCCERA

A rank ordering of the vocabulary in the corpus using WordSmith Tools (Version 6.0) reveals that a total of 78,329 word varieties occur in SCCERA. However, a large proportion of these (75.49%) are infrequent, with only 1-10 occurrences, and are less salient from a pedagogical perspective. The 50 most frequent words in SCCERA are listed below in Table 1:

1	the	26	data
2	of	27	model
3	and	28	not
4	in	29	can
5	to	30	fig
6	a	31	have
7	is	32	these
8	for	33	we
9	that	34	between
10	are	35	time
11	with	36	than
12	as	37	used

13	by	38	also
14	on	39	more
15	be	40	has
16	this	41	water
17	from	42	results
18	at	43	equation
19	was	44	using
20	et al	45	all
21	an	46	two
22	or	47	based
23	it	48	been
24	which	49	table
25	were	50	other

Table 1: The 50 most frequent words in SCCERA

Some of these words are typical of written academic genres, for example:

i. *A* and *the* indicate a high frequency of noun phrases in the corpus (e.g. *a* given density; *the* material)
ii. *Of* is commonly used for post-modification of noun phrases (e.g. the process *of* mixing)
iii. *That* is often used as a subordinator after report verbs (e.g. Evans (2014) suggests *that*…), or as a relative pronoun in relative clauses (e.g. contaminants *that* can be found in recycled concrete)
iv. Prepositions such as *in*, *to* and *for* are often used in prepositional phrases (e.g. increasing *in* the last few years; provides a good fit *to* the plot), which serve to add precision and detail to scientific writing.

However, other words in the list are more specific to science or civil engineering and can help ESP material writers to design their courses in a more principled manner. For example:

i. *Et al.* (the 20th most frequent expression in the corpus) reflects the high occurrence of references to collaborative work and co-authored papers in scientific research. This suggests that NNS learners are likely to need help with learning a variety of ways to cite others' work in order to avoid repetition in their writing.
ii. The high frequency of words such as *data*, *model*, *fig*, *equation*, and *table* all demonstrate the prevalence of 'multimodality' in civil engineering reports. In other words, texts are often mixed with other kinds of 'modes' (e.g. photographs, diagrams, mathematical equations, tables & charts) which writers then refer to (e.g. *Fig. 3 presents*…; *as in eqn. (1); see Table 2*). ESP writers in civil engineering could therefore usefully focus on language used to mediate between different modes in research reports.

Word frequency lists are therefore a useful starting point for materials design and, in collaboration with civil engineering faculty, can provide a summary of core vocabulary civil engineering students or fledgling academics should know.

3.2 Keywords in SCCERA

In some ways, keywords are more helpful for materials design than raw word frequency lists because they can help to reveal the 'aboutness' of a text or genre (e.g. Scott & Tribble 2006); in this case, words which typically characterize civil engineering RAs. Table 2 below shows the top 50 key words in SCCERA (using the BNC as the reference corpus):

1	et al	26	temperature
2	fig	27	measured
3	model	28	behavior
4	data	29	coefficient
5	equation	30	ratio
6	results	31	spatial
7	values	32	variables
8	models	33	distribution
9	flow	34	strain

10	concrete	35	method
11	table	36	parameter
12	shear	37	measurements
13	using	38	shown
14	wave	39	earthquake
15	figure	40	value
16	surface	41	density
17	parameters	42	average
18	water	43	respectively
19	analysis	44	precipitation
20	eq	45	displacement
21	soil	46	effects
22	based	47	cement
23	stress	48	climate
24	observed	49	maximum
25	velocity	50	project

Table 2: The 50 most frequent key words in SCCERA (compared to the BNC reference corpus)

Some of these words also occurred in the frequency list above (Table 1), but key words analysis is better at showing us how the language of a particular specialty differs from general English. Not surprisingly for a corpus of civil engineering texts, keywords in SCCERA seem to focus on materials – examining their physical properties and modeling or describing their behavior. Notice as well the low number of verbs appearing in the key word list (*observed*, *measured*, *shown*) - all past participles, suggesting a high frequency of passive forms in civil engineering RAs, as we would expect. This is also indicative of a large degree of nominalization, where noun forms are used in preference to verb forms (*analysis*, *behavior*, *distribution*, *measurements*, *precipitation*, *displacement*), which is another common feature of modern scientific writing. As Biber (2003: 170) writes, the 'informational explosion' of the 20th century has put writers of expository texts under more and more pressure 'to communicate information as efficiently and economically as possible, resulting in compressed styles that depend heavily on tightly integrated noun-phrase constructions'.

Key words analysis can also be useful for examining individual sub-disciplines within civil engineering. For example, Table 3 below compares the top 30 key words from the 11 sub-corpora of SCCERA, in descending order of 'keyness':

	Coastal Engineering	**International Projects**	**Earthquake & Disaster Mitigation**	**Mechanics & Structural Eng.**	**Transportation**
1	wave(s)	country/ies	earthquake(s)	damper(s)	travel
2	sea	growth	building(s)	beam(s)	vehicle(s)
3	coastal	income	disaster(s)	damping	car(s)
4	ice	poverty	tsunami	bridge(s)	transport
5	beach(es)	financial	seismic	response	time(s)
6	shelf/ves	firm(s)	damage	structural	trip(s)
7	ocean	we	evacuation	stiffness	traffic
8	breaking	capital	hazard(s)	control	link(s)
9	wind(s)	trade	ground	vibration	passenger(s)
10	coast(s)	GDP	stor(e)y	frequency/ies	activity/ies
11	island(s)	foreign	motion(s)	structure(s)	route(s)
12	water(s)	household(s)	loss(es)	plate(s)	network
13	storm	remittances	fire(s)	load	choice
14	shoreline	economic	risk	steel	port(s)
15	erosion	world	Japan	force	bus
16	tidal	development	response	displacement	rail
17	tide	FDI	emergency	strain	transit
18	currents	market	mitigation	equation	cost(s)
19	depth	bank	recovery	damage	congestion
20	arctic	labor	warning	the	hub
21	numerical	sector	hurricane	excitation	public
22	height	inequality	city	CFRP	cycling
23	bed	that	roof	nonlinear	utility
24	shore	political	vulnerability	dynamic	transportation
25	offshore	capita	shaking	system	demand
26	dune	institutions	residents	FRP	service
27	runup	poor	inundation	elastic	accessibility
28	reef	our	community	records	bike
29	salinity	aid	figure	concrete	commuting
30	Chl	investment	collapse	fatigue	freight

	Geotechnical	**Hydrology**	**River & Environmental Eng.**	**Regional Planning, Surveying**	**Concrete**	**Infrastucture**
1	soil(s)	climate	flow(s)	image(s)	cement(s)	project(s)
2	stress(es)	et al	river(s)	land	concrete	construction
3	test(s)	forcing	water	pixel(s)	material(s)	management
4	strain	precipitation	velocity/ies	accuracy/ies	strength	cost(s)
5	clay(s)	cloud(s)	bed	forest	phase(s)	risk(s)
6	sand(s)	ash	sediment	classification	hydration	manager(s)
7	shear	emission(s)	channel	spatial	fiber(s)	success
8	pile(s)	aerosol	vegetation	data	mixture(s)	life
9	tunnel	concentrations	hydraulic	area(s)	properties	team(s)
10	Fig	snow	depth	modis	mortar(s)	pavement
11	slope	change(s)	stream(s)	landsat	silica	contractor(s)
12	pore	temperature	habitat	urban	temperature	LCA
13	liquefaction	atmospheric	discharge	band(s)	aggregate(s)	research

14	pressure	rainfall	downstream	cover	corrosion	safety
15	curve(s)	runoff	turbulence	scene(s)	chloride	leadership
16	suction	groundwater	dam	azimuth	compressive	performance
17	specimen(s)	annual	roughness	object(s)	slag	organis/zational
18	KPA	warming	fish	SAR	calcium	process
19	loading	radiative	floodplain	GIS	content	environmental
20	cyclic	SST	upstream	feature(s)	ceramics	cycle
21	settlement	ice	Al	resolution	paste	maintenance
22	undrained	seasonal	groundwater	classes	shrinkage	respondents
23	consolidation	simulations	aquatic	algorithm	samples	overlay
24	failure	yr	ENKF	segmentation	gel	practices
25	strength	river	scale	EVI	chemical	design
26	effective	basin	figure	percent	glass	schedule
27	compression	ensemble	flood	LST	ferroelectric	knowledge
28	triaxial	dust	flux	map	reaction	infrastructure
29	drained	atmosphere	pressure	spectral	nano	PM
30	saturated	global	turbulent	tree	fly	information

Table 3: The 30 most frequent key words in the 11 sub-corpora of SCCERA

As can be seen in Table 3, the different sub-disciplines have quite distinct characteristics, with only 16 words re-occurring in more than one list (*bed*; *concrete*; *cost(s)*; *damage*; *depth*; *figure*; *groundwater*; *ice*; *pressure*; *response*; *risk*; *river(s)*; *strain*; *strength*; *temperature*; *water(s)*). This reflects, as Paxton et al. (2008, 115) note, the broad range of 'discourses that define the nature and practice of engineering that exist in some tension with each other […] management, economics, sociology, politics and development', and means that the kind of field-specific vocabulary students in each sub-discipline need to learn varies significantly. Of course, a quantitative analysis like this is only the first step in any materials design process – qualitative, pedagogic decisions then need to be taken in terms of which words from the lists warrant further investigation in the classroom. High frequency words such as *wave* might seem trivial for engineering students, however it should be remembered that receptive and productive knowledge of lexis are quite different and they might not be aware of ways that this word is used naturally in civil engineering contexts. In SCCERA, there are over 80 common collocates of *wave*, many of which may be unfamiliar to learners:

Wave collocates	**Hits in SCCERA**
~ height	813
~ velocity	279
~ period	263
~ break	249

~ energy	174
~ model	148
~ condition	147
~ propagation	145
~ runup	115

Table 4: Top collocates for *wave* in SCCERA

3.3 Cluster analysis in SCCERA

Lexical bundles (also known as clusters, formulaic sequences, prefabricated expressions, or N-grams), such as '*at the end of the ~*', are groups of words that commonly occur together in a particular register. They make up a large part of any discourse, often in excess of 50% (Schmitt 2004), and as such play an important role in the production of writing that conforms to the rhetorical norms of a particular research field. As Hyland (2004: 90) points out, 'persuasion is not simply accomplished with language, but with language that demonstrates legitimacy as writers draw on institutional practices which appeal to readers from within the boundaries of their discipline.' Significant lexical bundles can be identified in a corpus using a frequency-driven approach, although there is still some debate in the research literature over where the cut-off for inclusion should lie, with accepted recurrence rates usually ranging between 20 and 40 times per million words (e.g. Biber & Barbieri, 2007). However, selecting appropriate formulaic sequences for pedagogical applications goes beyond quantitative considerations and more difficult subjective decisions have to be made on questions such as the following[vi]:

i. Since shorter 2- or 3-word lexical bundles occur much more frequently than longer examples in any corpus, should the cut-off level for significance vary according to length?
ii. How can we best determine when to view a lexical bundle as an independent item or a fragment of a longer sequence?
iii. Should lower frequency bundles also be included in an ESP curriculum if they are shown to be salient to members of a particular discourse community?

The top 3- to 6-word lexical bundles in SCCERA are shown below in Table 5 (optional add-ons are in parentheses and variations in vocabulary are indicated with diagonal slashes):

Bundle size	Lexical bundle	Approx. # of hits in SCCERA
6-word	it should be noted that (the)	462
	it can be seen that (the)	387
	it is important to note that	94
5-word	in the case of (the)	1,326
	(as) a function of (the)	1,315
	on the other hand (the)	1,266
	on the basis of (the)	854
	at the end of (the)	659
	as a result (of) (the)	626
4-word	is/are shown in fig/figure/table	6,796
	as well as (the)	3,628
	is/are based on a/the	3,505
	in terms of (the)	2,998
	as shown in fig/figure/table	2,511
	with respect to (the)	2,174
	can be used (to)	1,775
	the results of (the)	1,744
	the effect of the	1,730
	(the) size of the	1,092
3-word	as well as	2,904
	in order to	2,498
	part(s) of the	1,705
	a (large) number of	1,338
	such as the	1,101
	used in the/this	1,643
	according to the	1,064
	most of the	1,023
	because of the	984
	the impact of	975

Table 5: Top 3- to 6-word lexical bundles in SCCERA

Predictably, the variation in bundle types in the corpus falls away rapidly with increasing length (335 3-word bundles; 124 4-word bundles; 10 5-word bundles; 3 6-word bundles) so, again, pedagogical judgments come into play when selecting which sequences to include in an engineering syllabus. 4-word bundles and above are perhaps the most relevant because they constitute a manageable number of items for overt instruction. The complete list of 472 bundles is, however, useful for ensuring that language-learning materials for civil engineers represent the most frequent naturally occurring collocation patterns in this genre, and for providing a more objective measure of the 'naturalness' of texts.

The types of bundles occurring in SCCERA reflect the kinds of 'work' civil engineers need to do in research articles, with at least 5 common varieties identifiable:

i. Language showing cause-effect relationships (*the effect of*; *the result of*).
ii. Language of comparison and contrast (*on the other hand*; *as well as*).
iii. Language for quantifying (*part(s) of the*; *the size of the*).
iv. Referential language, inside and outside the text (*is shown in fig/table*; *it can be seen that*).
v. Language showing the writer's stance (*it should be noted that*; *it is necessary to*).

These broad categories can also be useful in guiding materials development and ensuring that civil engineering students develop the language skills necessary for their future careers. For example, the high occurrence of referential expressions (see iv. above) in the corpus suggest that engineering RAs are extremely multimodal, with writers making regular use of diagrams, photographs, graphs, tables, mathematical equations, and so on to support their message. This was a feature also noted by Hyland (2008) who found similar characteristics in his Electrical Engineering corpus.

3.4 Parts-of-speech in SCCERA

POS tagging of SCCERA using CLAWS 4 (Lancaster University UCREL) indicated the following relative proportions of parts-of-speech in SCCERA, compared with the Brown Corpus (A general corpus of American English):

Part of speech	SCCERA	Brown Corpus
Nouns	32.2%	23.1%
Verbs	13.4%	15.5%
Prepositions	13.4%	-
Adjectives	10.2%	6.9%

Table 6: Parts-of-speech in SCCERA compared to the Brown Corpus

As can be seen, SCCERA contains a higher proportion of nouns and adjectives than the general corpus and this is in line with the typical characteristics of academic texts, where information-dense, noun-phrase constructions tend to be prevalent. An example from the corpus (Durdu, Mendoza & Terrones 2009: 208) illustrates how complex noun phrases with post-modification are often built up, one on top of another, in academic writing:

We also found that the adjustments in foreign assets and key macroeconomic aggregates triggered by financial globalization and Sudden Stop risk follow a gradual process with persistent current account surpluses and undervalued real exchange rates.

EAP students obviously find both reading and reproducing these dense texts in their own writing extremely challenging and need guided practice in constructing or deconstructing noun phrases.

The POS data from SCCERA also highlights the vocabulary-learning burden L2 engineering researchers face while learning to write in a discipline-specific manner. For example, the high percentage of adjectives in the corpus reflects the wide range of adjectival choices needed by academic writers for the modification of nouns. For instance, a search for general adjectives preceding the noun *aggregate(s)* in the tagged corpus reveals at least 24 choices that writers can select from:

annual
bottom-only
coarse
common
concrete
crushed
dry/dried
fine
future
given
high-grade
interacting
joint
lightweight
natural
particular
reactive
recycled
resistant
secondary
standard
suitable
virgin
wetted

aggregate(s)

Learners can also be trained how to query a specialized corpus such as SCCERA themselves, using commonly available freeware such as AntConc (available at http://www.laurenceanthony.net/software/antconc/). Many linguistic issues arising during the drafting of research articles can't be predicted in advance so this 'direct approach' to corpus use is an important adjunct to the corpus-informed approach to materials design described here.

4. Conclusion

This paper has described in some detail the development of the Specialized Corpus of Civil Engineering Research Articles (SCCERA) and the potential insights it can provide into the rhetorical practices of civil engineers writing for their peers in academic journals. As Feak & Swales (2010: 282) say, 'the era of specialized corpora in ESP contexts is upon us' and it is hoped that the methodology outlined here can provide a framework for other ESP specialists wishing to develop or exploit specialized corpora in their own fields. The construction of a balanced, representative corpus is only the starting point in a long journey of textual exploration for researchers or materials designers, however – corpora do not reveal their secrets easily and the identification of pedagogically useful patterns in the data depends on the right questions being asked. As we have seen in this paper, the goal of corpus

queries can vary widely, focusing on features such as lexical frequency, keyword analysis, cluster analysis and parts-of-speech. These provide different, but complementary, perspectives on the data which, together, can usefully inform the development of effective ESP materials.

Although corpus analysts are undoubtedly convinced of the potential of corpus linguistics to help improve language pedagogy, they are often criticized for not making sufficient efforts to communicate these benefits to practitioners. Methodological descriptions such as those outlined in this paper are important if we are to overcome the disciplinary barriers that often disrupt the free flow of relevant information between different academic fields.

References

Biber, Douglas. 2003. "Compressed noun-phrase structure in newspaper discourse: The competing demands of popularization vs. economy." In Jean Aitchison and Diana M. Lewis, eds., *New media language*, 169–181. London and New York: Routledge.

Biber, Douglas, Susan Conrad and Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use.* Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad, and Viviana Cortes. 2004. "*If you look at …*: Lexical bundles in university teaching and textbooks". *Applied Linguistics* 25 (3): 371-405. *doi: 10.1093/applin/25.3.371*

Biber, Douglas and Federica Barbieri. 2007. "Lexical bundles in university spoken and written registers." *English for Specific Purposes* 26: 263-286.

Boulton, Alex. 2012. "Corpus consultation for ESP". In Alex Boulton, Shirley Carter-Thomas and Elizabeth Rowley-Jolivet, eds., *Corpus-informed research and learning in ESP: Issues and applications*. Amsterdam: John Benjamins, 261-291.

Carter, Ronald and Michael McCarthy. 2006. *Cambridge grammar of English.* Cambridge: Cambridge University Press.

Durdu, Ceyhun B., Enrique G. Mendoza and Marco E. Terrones. 2009. 'Precautionary demand for foreign assets in Sudden Stop economies: An assessment of the New Mercantilism'. *Journal of Development Economics* 89: 194 – 209.

Feak, Christine B. and John Swales. 2010. "Writing for publication: Corpus-informed materials for postdoctoral fellows in perinatology". In Nigel Harwood, ed., *English Language Teaching Materials*. Cambridge: Cambridge University Press, 279-300.

Hyland, Ken. 2004. "A convincing argument: Corpus analysis and academic persuasion". In Ulla Connor and Thomas A. Upton, eds., *Discourse in the professions: Perspectives from corpus linguistics*. Amsterdam: John Benjamins.

Hyland, K. 2008. “As can be seen: Lexical bundles and disciplinary variation”. *English for Specific Purposes* 27: 4-21.
Jurafsky, Daniel and Martin, James H. 2009. *Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing* (2nd Edition). Upper Saddle River, NJ: Prentice Hall.
Kennedy, Graeme. 1998. *An introduction to corpus linguistics*. New York: Longman.
Legates, David and Gregory J. McCabe Jr. 1999. “Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation”. *Water Resources Research* 35 (1): 233-241.
McCarthy, Michael. 1990. *Vocabulary*. Oxford: Oxford University Press.
Paxton, Moragh, Ermien van Pletzen, Arlene Archer, Moeain Arend and Clement Chihota. 2008. “Writer’s stance in disciplinary discourses: A developmental view.” *Southern African Linguistics and Applied Language Studies* 26 (1): 107-118.
Pearson, Jennifer.1998. *Terms in context.* Amsterdam: John Benjamins.
Rea Rizzo, Camino. 2010. “Getting on with corpus compilation: From theory to practice”. *ESP World* 1(9): 1-22.
Schmitt, Norbert, ed. 2004. *Formulaic sequences*. Amsterdam: John Benjamins.
Scott, Mike. 2012. *WordSmith Tools Version 6*, Stroud: Lexical Analysis Software.
Scott, Mike and Christopher Tribble. 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.
Sinclair, John M. 1990. *Collins COBUILD English grammar.* London: Collins.
Wolfson, Nessa. 1989. *Perspectives.* Boston MA: Heinle & Heinle Publishers.
Wray, Alison. 2000. “Formulaic sequences in second language teaching: Principle and practice.” *Applied Linguistics* 21 (4): 463-489.

Appendix 1: Summary of journals used in SCCERA

Department	Journal name
Coastal Engineering	J. of Geophysical Research: Oceans
	J. of Coastal Research
	Coastal Engineering
	J. of Waterway Port Coastal & Ocean Engineering
Infra-structure Development	J. of Construction Engineering & Management
	J. of Infrastructure Systems (ASCE)
	Int. J. of Project Management (Social Sciences C.I.)

Concrete Engineering	Cement & Concrete Research
	Cement & Concrete Composites
	J. of American Ceramic Society
	Material & Structures
	ISPRS J. of Photogrammetry & Remote Sensing
Regional Planning, Surveying, Remote Sensing	Remote Sensing of Environment
	IEEE Transactions on Geoscience and Remote Sensing
	Int. J. of Geographical Information Science
	J. of Regional Science (Social Sciences C.I.)
	ASCE J. of Hydraulic Engineering
River & Environmental Engineering	Water Resources Research
	J. of Hydraulic Research
	River Research and Applications
Hydrology & Water Resources Engineering	Journal of Geophysical Research: Atmospheres
	Journal of Hydrology
	Journal of Climate
	Hydrological Processes
Geotechnical Engineering	Canadian Geotechnical Journal
	ASCE J. of Geotechnical and Geoenvironmental Engineering
	Soils and Foundations (Japanese Journal)
	World Development
	Geotechnique
International Projects	Journal of Development Economics
	Transportation Research (Part B - Methodological)

Transportation Research	Journal of Transport Geography
	Transportation Research (Part A - Policy & Practice)
	Journal of Engineering Mechanics (ASCE)
Mechanics & Structures	Journal of Structural Engineering (ASCE)
	Journal of Sound and Vibration
	Journal of Earthquake Engineering
	Engineering Structures
	Structural Control and Health Monitoring
	Journal of Bridge Engineering
	Journal of Disaster Research (Japanese Journal)
Earthquake & Disaster Mitigation	Journal of Natural Disaster Science (Japanese Journal)
	Natural Hazard Review
	Earthquake Engineering and Structural Dynamics
	Earthquake Spectra

Endnotes

[i] 'Mega corpora' are large general corpora made up of hundreds of millions or billions of words. The British National Corpus (BNC) and Corpus of Contemporary American English (COCA) are given as examples here because they are both freely available for consultation online: http://corpus.byu.edu/

[ii] Collocation describes the way words do or do not tend to co-occur, for example the table below shows how 4 adjectives denoting size collocate with different nouns:

	problem	*amount*	*shame*	*man*
large	?	√	X	√
great	√	√	√	√
big	√	√	X	√
major	√	?	X	X

√ = collocates; ? = questionable; X = does not collocate

Table 1: Collocation matrix for adjectives denoting size (McCarthy 1990: 12)

[iii] 'Lexical bundles', also commonly known as 'formulaic sequences' (Wray 2000), 'multiword lexical units/chunks' (Lewis 1993), or 'N-grams' (Jurafsky & Martin 2009) are defined here as 'the most frequently recurring lexical sequences in a register' (Biber, Conrad & Cortes 2004)

[iv] Part-of-speech tagging is the most common form of corpus annotation used in corpus linguistics. CLAWS 4, developed at the University of Lancaster, UK, tags each word according to its part of speech (e.g. noun, verb, adjective), with 96-97% accuracy. There are a total of 137 tag types in the C7 tag set: http://ucrel.lancs.ac.uk/claws7tags.html

[v] Of which 43 are cited in Science Citation Index-Expanded (SCI) or Social Science Citation Index (SSCI).

[vi] These issues will be discussed in greater detail in a forthcoming paper.