CrossMark

Lang. Teach. (2017), **50.1**, 90–106 © Cambridge University Press 2014
doi:10.1017/S0261444814000378 First published online 22 December 2014

Plenary Speech

Corpus-based approaches to language description for specialized academic writing

John Flowerdew English Department, City University of Hong Kong, Tat Chee Avenue, Kowloon Tong, Hong Kong
enjohnf@cityu.edu.hk

Language description is a fundamental requirement for second language (L2) syllabus design. The greatest advances in language description in recent decades have been done with the help of electronic corpora. Such language description is the theme of this article. The article first introduces some basic concepts and principles in corpus research. It then reviews some recent corpus studies of relevance to the teaching of English for specific purposes (ESP) writing. The speech focuses on four different types of corpora: expert professional corpora, expert student corpora, L2 learner corpora, and lingua franca corpora. It also discusses application of the corpus approach, including indirect applications and direct applications. In a final section, it presents some caveats of the corpus approach for future language teaching research.

1. Introduction

Language description is a fundamental requirement for L2 syllabus design. Much work has been done on language description in the context of disciplinary academic writing, the theme of this conference. The greatest advances in language description in recent decades have been achieved with the help of electronic corpora. Such language description is what I'll be talking about today.

First of all, I'd like to talk about some principles of the corpus approach to academic language description. I'm going to talk about expert professional corpora, expert student corpora, learner corpora, and lingua franca corpora. I'll explain each of these terms when I come to it. I'll also talk about some of the applications of this type of research. I'll furthermore discuss some caveats you need to consider in terms of application. I should say that I am going to assume, for the purposes of this talk, which is in the context of a conference on *Writing for academic disciplines*, that we want to focus on disciplinary specificity. Of course, you could argue that, given that learners' needs are far from fully predictable, due to shifts in career, developments in the discipline, and individual ideologies, etc., we should focus on increasing awareness of intra- and inter-disciplinary variation. Nevertheless, even if such a

Revised version of a plenary address given at the *11th Annual Technology and Teaching Writing for Academic Disciplines Conference*, at Iowa State University, Iowa, USA, 19 October, 2013.

https://doi.org/10.1017/S0261444814000378 Published online by Cambridge University Press

stance is taken, the starting point needs to be a description of the individual disciplines that we want to compare. So the focus here will be on discipline-specificity.

2. Principles of the corpus approach

2.1 What is a corpus?

This is a definition I quite like by Sinclair (1991: 171).

> A collection of naturally occurring language text in electronic form, selected according to external criteria to represent as far, as possible, a language or language variety as a source of data for linguistic research.

General corpora include the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA), which try to give us as wide a sample of the language as possible. There are more specific corpora including, for academic English, the British Academic Spoken English Corpus (BASE), the British Academic Written English Corpus (BAWE), the Michigan Corpus of Academic Spoken English (MICASE), and the Michigan Corpus of Upper-Level Student Papers (MICUSP). As an example for business, there is the Hong Kong Polytechnic University Corpus of Business Correspondence. Another way of categorizing corpora is if they contain English of so-called native speakers, or, in the case of the International Corpus of Contemporary English (ICE), different varieties of English from outside the American and British 'standards'. There are also English as a lingua franca (ELF) corpora, such as the Vienna Oxford International Corpus of English (VOICE), or the English as a Lingua Franca in Academic Settings Corpus (ELFA). Furthermore, we have learner corpora, that is, corpora created from English produced by learners of the language. Most corpora are synchronic; but we also have diachronic corpora to see how features of language evolve over time, although such corpora are not of interest to us here today.

2.2 A fundamental principle: frequency

As Leech (2011: 14) put it, applying this principle to language teaching, 'the principle *more frequent = more important to learn* can scarcely be gainsaid as a general principle.' So frequency is fundamental and is a major contribution of corpus linguistics to language teaching (although, of course, in language teaching we shouldn't forget that there are other criteria for teaching, such as learner need, learner interest, teachability, classroom use, and systematicity).

2.3 Levels of analysis

In corpus linguistics, there are three main levels of analysis which have been used: single words, multi-word units, and rhetorical moves. Coxhead's well-known academic word list is a corpus-based study of the most frequent academic vocabulary, consisting of 650 word families (Coxhead 2000). (Coxhead created her own academic corpus on which to base the

https://doi.org/10.1017/S0261444814000378 Published online by Cambridge University Press

study.) Multi-word units are also referred to as 'collocations', 'lexical bundles', and 'n-grams'. Hyland defines them as 'extended collocations that appear more frequently than expected by chance' (Hyland 2012a: 150). The bundle approach is based on Sinclair's idea that 'a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments' (Sinclair 1991: 110). Perhaps less talked about here, but very interesting to me, is the level of move. You are probably familiar with the Swalesian notion of move in genres: if you are writing an introduction to a research article, first of all, establish the territory, then identify a gap in the field, and then say how you are going to fill the gap. These are the three moves you would typically have in a research article introduction (Swales 1990) and we can use corpus techniques to investigate such moves.

Now, what is the argument for specialist corpora? Here is another point made by Leech (2011: 13): '[F]requency counts are least useful when they are based on a general corpus covering a range of the language; they are more useful if they are differentiated for region and register.' The basic idea is that a given lexical or grammatical item must be related to its particular rhetorical purpose. If you are teaching writing, you need to think about what sort of writing you are going to teach. You need to base your frequency on a particular variety of language. So the principle here is that a given lexical or grammatical item must be related to its particular rhetorical purpose, and these purposes vary according to register. This approach can inform writing for specific purposes. Bloch (2008: 182–183) suggests that

> Although developing explicit knowledge of grammar can be helpful, the universalist approach to grammar as found in grammar textbooks does not work well, because the emphasis has shifted to studying the 'local knowledge' of language.

So the idea here is that language varies according to the local contexts. I think Bloch is talking about the shift from the more cognitive Chomskyan focus on universal grammar to the more social view of the language which corpus researchers find more interesting. Language varies according to the local situation. Friginal (2013: 208) adds that

> Intuitively, knowledge gleaned from corpus-based research, which identifies features and systematic associations of language characteristic of writers in a particular field, could aid the teaching of writing for specific purposes and writing across various disciplines.

For Friginal, then, corpus techniques are extremely useful for specific purpose writing.

3. Expert professional corpora

This brings me to the first type of corpora I am going to talk about today: expert professional corpora. By professional corpora I mean corpora made up of data that has been written or spoken by professional academics. Most of the professional corpora to date have been based on research articles, lectures and textbooks. I'll review a few key studies to show the scope of this work.

https://doi.org/10.1017/S0261444814000378 Published online by Cambridge University Press

3.1 Biber's study of university language

Biber's study of university language covers topics including vocabulary use, lexico-grammatical variation, stance, lexical bundles, and multidimensional analysis (Biber 2006). The last-mentioned approach is a statistics-based analysis of clusters of linguistic features that express a particular function of language – whether the language is informative, not informative, personal or impersonal, etc. Biber looked at ten university registers: class sessions, classroom management (spoken), labs and other in-class groups, office hours, study groups, service encounters, textbooks, course packs, course management (written), and institutional writing. He did not cover relevant genres for academic writing, however, so this study is not of any value for the teaching of discipline-specific writing (although the methodology is nevertheless of interest and I will return to it later).

3.2 Hyland's work on social interactions in academic writing

Hyland's work is more relevant here. Hyland (2000) started his research into disciplinary language by asking questions like 'why do engineers "report" while philosophers "argue" and biologists "describe"'? In other words, why do different disciplines use language in different ways? He is interested in disciplinary variation, the relationships between the cultures of academic communities and their unique discourses. Hyland uses a multi-method approach, with traditional discourse analysis, interviews with insiders and discourse-based interviews (interviews where interviewees are asked about their texts), as well as corpus linguistics. Interestingly, Hyland's work combines corpus work with more ethnographic types of investigation, e.g. interviews. Table 1 shows the variation of the frequency of the use of citations across disciplines from Hyland's (2000) study.

Table 1 shows that in sociology there are, on average, 104 citations per paper, while in physics there are only about 24 citations per paper. So if you are designing a writing course for sociology students, you probably want to prioritize how to cite. If you are writing a course for physics students, you might well give less attention to citation practices (although you wouldn't want to ignore the skill altogether).

In the next table from Hyland (see Table 2), he is looking at different verbs that different disciplines favour when they cite. You can see that philosophers use *say, suggest,* and *argue,* while mechanical engineers use *describe, show* and *report.* Again this is very useful information if you are designing specific purposes writing courses.

3.3 Hyland's work on disciplinary identities

In his later book, Hyland (2012b) looks at individual identities as well as disciplinary identity. He argues that 'authors convey aspects of their identities within the constraints placed upon them by their disciplines' rhetorical conventions' (Hyland 2012b: reverse cover blurb). He claims that corpus methods are important tools in identity research. He uses keyword and

https://doi.org/10.1017/S0261444814000378 Published online by Cambridge University Press

Table 1 Rank order of citations by discipline (adapted from Hyland (2000: 24).

Rank	Discipline	Av. per paper	Per 1,000 words	Total citations
1	Sociology	104.0	12.5	1040
2	Marketing	94.9	10.1	949
3	Philosophy	85.2	10.8	852
4	Biology	82.7	15.5	827
5	Applied linguistics	75.3	10.8	753
6	Electronic engineering	42.8	8.4	428
7	Mechanical engineering	27.5	7.3	275
8	Physics	24.8	7.4	248
Total				5372

Table 2 Reporting verbs in citations (adapted from Hyland (2000: 27)

Discipline	Reporting structures		Most frequent forms
	Per paper	% of citations	
Philosophy	57.1	67.0	say, suggest, argue, claim, point out, propose, think
Sociology	43.6	42.0	argue, suggest, describe, note, analyse, discuss
Applied linguistics	33.4	44.4	suggest, argue, show, explain, find, point out
Marketing	32.7	34.5	suggest, argue, demonstrate, propose, show
Biology	26.2	31.7	describe, find, report, show, suggest, observe
Electronic engineering	17.4	40.6	propose, use, describe, show, publish
Mechanical engineering	11.7	42.5	describe, show, report, discuss
Physics	6.6	27.0	develop, report, study
Overall			suggest, argue, find, show, describe, propose, report

collocation analysis to highlight norms of a particular genre and an author's idiosyncratic choices. (A keyword is a word in a given corpus that occurs significantly more frequently in a corpus than a reference, or general, corpus. For example, the word *lecture* in a corpus of

https://doi.org/10.1017/S0261444814000378 Published online by Cambridge University Press

lecture data may be a keyword because it occurs more frequently in the specialist corpus than the general reference corpus that it is compared with).

Hyland compares two famous linguists' writing styles: John Swales and Deborah Cameron. He shows that these two scholars write in very different ways and project different identities. He concludes that Deborah Cameron constructs an identity of a 'radical linguist', while John Swales constructs a rather different identity of an 'inquiring colleague'. To take just one example of the differences between the two academics, significant keywords in the discourse of Deborah Cameron are *not*, *but*, and *though*; as keywords these words are used much more frequently than is the norm in academic discourse and are indicative of Deborah Cameron's forcefulness in argumentation, Hyland claims. An important keyword used by John Swales, on the other hand, is *I*, and this word is frequently used with verbs such as *think*, *believe*, *suspect*, *hope*, *try*, and *guess*; this is indicative, Hyland argues, of John Swales's more reflective approach than that of Deborah Cameron. Of course you could say that these people are outliers. More average academics like me or you probably write in a pretty similar way. These two are very individual personalities within the same discipline. This is the point Hyland wants to make. In corpus work, we usually look at generalities, so it is quite interesting that Hyland looks at idiosyncrasy as well as generalization. The moral here might be that the extent of variation within a discipline-specific corpus may be as important as the general norms.

3.4 Flowerdew and Forest's work on signalling nouns

I now move to some work done by myself with my collaborator, Richard Forest. First, I am going to talk about an article I worked on with Forest in 2009 (Flowerdew & Forest 2009), the idea of which was to see if there is any systematic relation between rhetorical moves and lexico-grammatical realization of those moves.

In this study we started out by looking at some key keywords in a corpus of Ph.D. thesis literature review chapters. (I'm using a bit of poetic licence here by including this study under the 'expert professional corpora' heading, even though the data is produced by students, albeit very advanced ones.) A key keyword is one which occurs consistently across the individual texts in a corpus and is not just frequent in a relatively small number of those texts. It helps you avoid looking at words that might be more frequent in just one or more files in the corpus. In Table 3, you can see that some of the words, such as *study, analysis, research, example, approach, discussion,* and *strategy* – all very typical in academic discourse – are abstract nouns.

Based on the frequency of these key keywords, we decided we would investigate one of them in greater detail – *research*. We wanted to focus on this word and see how it is distributed across the moves in the literature review sections of research articles. Figure 1 shows the distribution of *research* across the moves. You can break moves down into component steps or sub-moves. So here we have a fine-grained analysis of the moves and sub-moves of literature review thesis chapters (based on Kwan 2006).

The noun *research* occurs in the gap where the authors state 'research has been done on this and that; however, etc., etc., etc., . . . ' – that is the gap. So the word *research* occurs very frequently in that move. When you look at the concordance lines (see Fig. 2), it typically collocates with *little*: 'and little research has been carried out'; 'there has been little research

https://doi.org/10.1017/S0261444814000378 Published online by Cambridge University Press

Table 3 Key keywords in the literature review corpus (adapted from Flowerdew & Forest (2009)).

Rank	Key keyword	Texts	% of Texts	Overall word frequency	Standardized frequency per 1000
1	#	20	100	15,840	25.7
2	language	19	95	1,512	2.5
3	of	17	85	16,994	27.6
4	non	16	80	274	0.4
5	context	15	75	547	0.9
6	studies	15	75	535	0.9
7	analysis	14	70	604	1.0
8	linguistic	14	70	330	0.5
9	research	14	70	900	1.5
10	study	14	70	978	1.6
11	text	14	70	1,128	1.8
12	texts	14	70	680	1.1
13	chapter	13	65	193	0.3
14	communicative	13	65	182	0.3
15	discourse	13	65	617	1.0

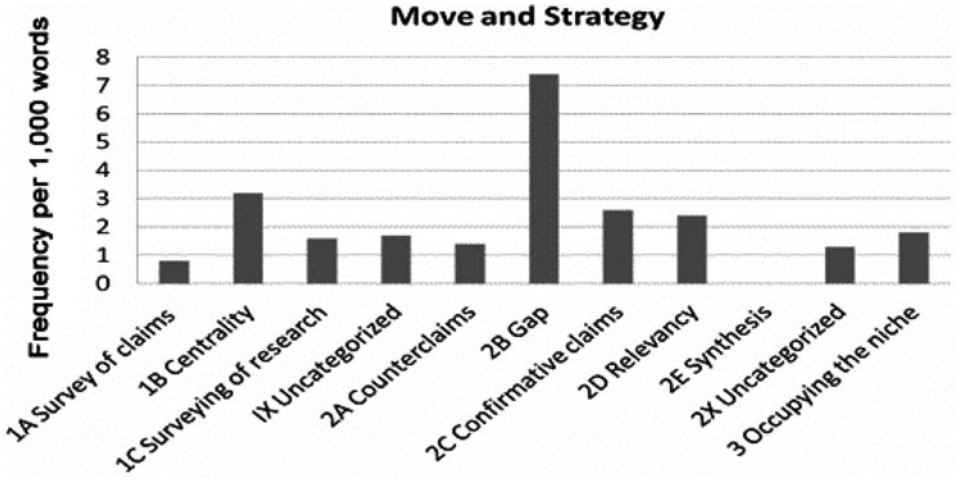


Figure 1 Key keyword frequencies: *research* (adapted from Flowerdew & Forest (2009)).

into'. This is a positive finding for our research question: 'is there a relationship between move and lexico-grammar?' Certainly there is here with *research*. If you are thinking about how we can use this information to teach people to write literature reviews, you need to say to learners that they should write something like 'there is little research in . . . '; 'little research has been done.' We can see this in the concordance lines in Figure 2.

This is one way to indicate the gap. Another way to do it, as indicated by other concordance lines (not shown here for reasons of space) is '[f]urther research is called for.'

https://doi.org/10.1017/S0261444814000378 Published online by Cambridge University Press

- And little research has been carried out with 'SUBJECT POPULATION'.
- However… there has been little research into 'DOMAIN/ TOPIC'.
- Aside from 'PROPER NAME'S' work, there has been little research into 'DOMAIN/ TOPIC'.
- While there has been little research with 'SUBJECT POPULATION 1', there has been even less, if any, with 'SUBJECT POPULATION 2'.
- NP is clearly important but there is little research in 'NAME OF APPROACH' to…
- There is little research on 'DOMAIN/ TOPIC' however, and…
- There is little research on 'DOMAIN/ TOPIC' .
- Relatively little research has been done in 'NAME OF DISCIPLINE'.
- and so it is perhaps surprising that little research has been conducted into 'DOMAIN/ TOPIC.
- There is very little research on 'DOMAIN/ TOPIC.
- There has been minimal research into 'DOMAIN/ TOPIC.
- But minimal empirical research… has been carried out to date within 'NAME OF DISCIPLINE'.

Figure 2 Concordance lines with *research* preceded by *little* (and *minimal*)
Note. NP = noun phrase

Another word from our key keyword list interacts with *research* – instead of saying *research*, you can say *studies*. That is why *studies* (in the plural form) is interesting, not *study* (in the singular). This is, incidentally, an argument for not only looking at lemmas. Here, the plural form is more interesting than the singular. Of course, *studies* can also be a verb, but such instances were not included in the counts.

So *research* and *study* are examples of abstract nouns which are frequently used in academic discourse. I have thought a lot about this type of noun over many years and my recent monograph, again with Forest (Flowerdew & Forest 2015), is a comprehensive study of how these nouns function in academic discourse. I call them SIGNALLING NOUNS (SN). Some other examples of these nouns are *advantage, case, evaluation, possibility, procedure, situation,* and *solution*. The term 'signalling noun' applies to abstract nouns which refer to a general area of meaning of which the specific meaning is found elsewhere in the clause or text. We can say that there is identity of reference between the SN and its specifics. These nouns can also be viewed as nominalized processes. They all have some process in their semantic content. Many of them are, in fact, derived from verbs. SN is a functional category in that these abstract nouns are not always SNs. They are SNs only when their specific meaning may be found elsewhere in the text. If these words do not have their meaning specified in the text, but the reader or listener has to go outside the text to interpret their full meaning, they are not SNs. Here is an example of how SNs occur in text.

- the **realization** that *it was possible to simulate a prebiotic milieu in the laboratory* ushered *in a new era in origin-of-life studies*

In this example, 'realization' is the SN and the underlined part in italics is the specifics. Here is another example:

- their **role** is *to carry out the depolarizing phase of an action potential*

https://doi.org/10.1017/S0261444814000378 Published online by Cambridge University Press

You can see from these examples that the specifics occur within the same clause as the SN. But the specifics may also be found across clauses. Here is an example:

- . . . after the earth was formed, it was subjected to a period of heavy bombardment with large (100 km diameter) comets and meteorites. During this **time** . . .

So the SN 'time' is anaphoric here; it is referring back to the underlined part. In the next example, the SN refers forward.

- Resources are not unlimited. Shortages, temporary or permanent, can result from several **causes**. Brisk demand may bring in orders that exceed manufacturing capacity or outpace the response time required to gear up a production line . . .

There are several 'causes' and the underlined parts begin to tell you what these causes are. This is how these nouns function in discourse.

There are some interesting findings if you compare the disciplines for the occurrences of SNs. The percentage of SNs varies a lot across disciplines (Table 4).

Table 4 Relative frequency of signalling nouns across disciplines

Discipline	Tokens (Norm. Freq)	% of SNs
Biology	1328	5.10%
Chemistry	1257	4.80%
Ecology	1477	5.60%
Engineering	1997	7.60%
Physics	2000	7.60%
Natural sciences	8059	30.70%
Business	2926	11.20%
Economics	3059	11.70%
Law	3969	15.80%
Political science	4143	15.20%
Sociology	4018	15.40%
Social sciences	18115	69.30%
Total	26174	100.00%

Notes. a. 'Tokens' refers to the total number of SNs in the given sub-corpus.

b. 'Norm. Freq' refers to the fact that these numbers are normalized per million words.

The numbers indicate that social scientists use these nouns more than twice as often as natural scientists, which is interesting. In Flowerdew & Forest (2015), we argue that the reason for

https://doi.org/10.1017/S0261444814000378 Published online by Cambridge University Press

Table 5 High frequency SNs in the natural and social sciences

Natural sciences		Social sciences	
equation	87.6%	right	100.0%
solution	87.0%	policy	99.1%
method	73.8%	issue	96.6%
problem	52.9%	theory	95.1%
model	52.2%	argument	93.3%
result	51.9%		
thing	50.3%		

Table 6 Frequency of SNs across genres

Genre	SN tokens	% of SNs in genre
Journals	6088	37.9%
Textbooks	4903	30.5%
Lectures	5067	31.6%
Total	**16058**	**100.0%**

this is that natural scientists use a lot more technical terms than do social scientists and so have less need for SNs. Some of the nouns are predominantly within one of these domains of social and natural science (Table 5). Thus *equation* occurs nearly 88% of the time in natural science texts, for example, with *solution,* and *method* also very frequent. Social scientists, on the other hand, prefer the SNs *right*, *policy*, *issue*, and *theory,* all of which occur over 90% of the time in their disciplines. That's just to give you a flavour of some of the variation.

The corpora I worked on are from three genres: research articles, textbooks, and lectures. The frequency of SNs varies a lot across these genres (Table 6). The table shows that SNs are the most frequent in the journals, second most frequent in the lectures, and third most

https://doi.org/10.1017/S0261444814000378 Published online by Cambridge University Press

frequent in the textbooks. Based on what we know about written texts from people like Biber, we would expect that journals would have the highest frequency because they are very nominal; and SNs are, of course, nominal. We would probably think lectures would have the least frequency of SNs, because they employ spoken language. Actually, the least frequent occurrence is in the textbooks but very close, in terms of frequency, to lectures. I think the reason for that is that textbooks and lectures are both didactic. The person who produces these didactic texts is at pains to make things clear, to explain in simple terms. So these texts are not too dense and do not have so many SNs, whereas journals do not need to worry about simplicity.

4. Expert student corpora

If you are designing a writing course for students, they may not need to write research articles or textbooks, but they do need to write essays. That is where 'expert' student corpora come in – the corpora of language written by grade 'A' students. There are a number of corpora of this type available and I referred to two of them earlier. One of these is MICUSP (http://micusp.elicorpora.info/), and if you want to know how really good students in Michigan write, this is the corpus you want to look at. Then there is the British counterpart of MICUSP – BAWE, where masters-level papers, as well as those of undergraduates, are included. These corpora have quite nice interfaces and are fairly easy to use. Students can interact directly with them to improve their writing. These 'expert' corpora, by the way, are not limited to native speakers, but include any writing achieving the required standard, as evaluated by the students' content teachers.

Basing their research findings on the BAWE corpus, Nesi & Gardner (2012) discovered that in British universities there are 13 different genres that students typically write in. We tend to think that students have to write essays; however, writing may include, for example, case study; critique; design specification; empathy writing; methodology recount; and narrative recount. Empathy writing (writing for non-academic audiences) is interesting because we usually think academic writing is very impersonal, and just to provide information. But in some genres, students have to empathize; they have to express their own subjective attitudes towards what they are writing about. As in the work of Hyland (2000, 2012a, 2012b), Nesi & Gardner's research on the BAWE corpus was triangulated with (in their case) interviews with teachers and students. In this project, they also collaborated with Biber to do multidimensional register analysis, plotting the genres according to Biber's multidimensional analysis, as referred to earlier. Figure 3 shows you the involved versus informational dimension spectrum (the involved dimension is characterised by more interpersonal linguistic items and the informational dimension has more impersonal features). You can see from the figure that 'empathy writing' is the most involved of the registers, while literature survey is the most informational.

In Table 7, we see the typical verbs used in the different disciplines studied by Nesi & Gardner. There is quite a lot of disciplinary variation among these verbs. If you take a look at 'law' you can see lawyers use the following verbs: they *afford*, they *assert*, they *bind*, and they *commit*. Now take a look at what student writers do in 'classics': they *attack*, they *defend*, they

https://doi.org/10.1017/S0261444814000378 Published online by Cambridge University Press

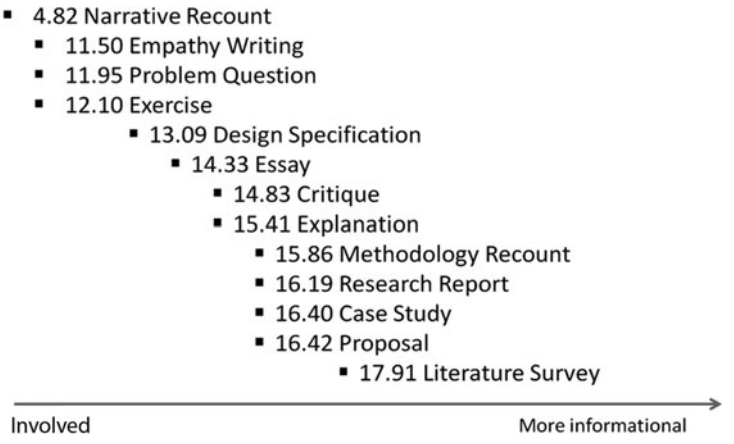


Figure 3 The involved versus informational dimension (adapted from Nesi & Gardner (2012: 44)).

Table 7 Verbs across the disciplines (adapted from Nesi & Gardner (2012: 128)).

Law	Philosophy	Sociology	English	Classics
Verbs				
Acknowledge	Accept	Challenge	Capture	Attack
Afford	Admit	Cite	convey	Born
Assert	Appeal	Constitute	Depict	Defeat
Bind	Claim	Criticise	Desire	Defend
Commit	Commit	Dominate	Die	Depict
Deem	Conceive	Embed	Emphasise	Desire
Deter	Concede	Embody	Escape	Die
Enforce	Criticise	Embrace	Evoke	Fight
Impose	Deny	Emerge	Explore	Kill
Intervene	Distinguish	Legitimise	Force	Let
Justify	Doubt	Oppress	Hear	Love
Prohibit	entail	Police	Imagine	Marry
Protect	Imagine	Portray	Imply	Portray
Refuse	Justify	Question	Love	Remind
rely	Perceive	Redefine	Portray	Rule
	Possess	Reinforce	Read	Send
	Postulate	Reproduce	Refuse	Tell
	Reason	Retain	Speak	Translate
	Refute	Shape	Witness	Win
	shall	transcend	write	worship

die, they *fight*, and they *kill*. Much more primordial types of words are used in classics than in law. This would be very useful data if you were teaching academic writing to classics students; you might not realize you have to teach these words in an academic writing course. This is not your typical academic word list here because it is more empathy-type writing.

I have focused on individual words, and moves, but Nesi & Gardner (2012) also look at n-grams. Table 8 shows the 4-gram or 4-word clusters used in the sub-corpus referred

https://doi.org/10.1017/S0261444814000378 Published online by Cambridge University Press

Table 8 4-grams in the professional development and planning subcorpus (adapted from Nesi & Gardner (2012: 241)).

	Cluster	Frequency	Texts
1	To be able to	21	10
2	I was able to	18	13
3	The rest of the	18	14
4	Health and social care	12	5
5	The upper level pitch	11	5
6	At the same time	10	7
7	I would like to	10	7
8	Will be able to	10	8
9	I will be able	9	7
10	I feel that I	8	7
11	I feel that this	8	6
12	I will try to	8	6
13	In order to improve	8	5
14	Member of the group	8	5
15	The end of the	8	8

to as 'professional development planning', a type of language with a strong interpersonal dimension, as you can see from some of these n-grams.

5. Learner corpora

With learner corpora, the idea is that you get a corpus of language created by non-native speaking learners of English. The objective is to see how the learners write compared to the native speakers, or the experts (which can, again, include non-native speakers). Following on from this analysis, you are in a position to design some pedagogical intervention to bring the learners closer to the expert model. The best known learner corpus is called the International Corpus of Learner English (ICLE), created by Université Catholique de Louvain (http://www.uclouvain.be/en-cecl-icle.html). The corpus includes argumentative essays written by higher intermediate to advanced learners of English from many different language backgrounds. An interesting study based on this corpus by Paquot (2010) used ICLE as the learner corpus and created a reference corpus out of various sources containing expert academic language. Based on this, Paquot came up with a set of 930 academic key keywords based on keyness, range, and evenness of distribution. I mentioned Coxhead (2000) earlier and she excluded very common English words. But this list includes all the 2,000 frequent words. Based on her analysis, Paquot came up with a range of difficulties learners have in argumentative essay writing, including a limited lexical repertoire, a lack of register

https://doi.org/10.1017/S0261444814000378 Published online by Cambridge University Press

awareness, use of infelicitous word combinations, misuse of words semantically, sentence-initial, positioning of adverbs, and transfer effects. Some people may say this is a deficit view of language learning and we should focus more on the positive aspects of learner language. But this can nevertheless be a useful approach, depending on your point of view concerning my next topic: lingua franca corpora.

6. Lingua franca corpora

There is a lot of interest in lingua franca English these days: some people are saying you need to follow standard English; others are saying we can use lingua franca English. There are two well-known corpora for lingua franca English: the VOICE corpus in Vienna (https://www.univie.ac.at/voice/) and the ELFA corpus in Helsinki: (http://www.helsinki.fi/englanti/elfa/elfacorpus). These are both spoken corpora and so are not highly relevant to today's talk. However, based on these corpora, research is showing that people using English as a lingua franca (ELF) are perfectly able to communicate effectively with each other in 'non-standard' English. If this is the case, the question arises as to whether or not ELF can be used in written genres, many of which are much more conservative than spoken ones, especially research articles, which might be thought to be as conservative a genre as any. Rozycki & Johnson (2013) took a collection of articles from the top engineering journals which had all been given a prize for being the best articles of the year. A corpus study of these articles showed the presence of lingua franca grammar, including the dropping of articles or insertion of articles and a lack of concord in number marking between subject and predicate. The researchers claim that this may indicate that the gate-keeper role in engineering now reflects the predominance of non-native speakers in the field, not Anglophones from the so-called inner circle countries (Kachru 1992). The standard is now determined by these international reviewers and not by the 'native speakers' in America or the UK or Australia. It's the English of the world; it's ELF.

For those of us who are English teachers, the non-standard forms found by Rozycki & Johnson are the things that we spend a lot of time telling students not to do. ELF could be bad news for us, because English teachers might become redundant if we allow for it! I mentioned Rozycki & Johnson's paper in a conference in Taiwan and one delegate got very upset. 'I've spent 30 years trying to stop my students from writing like this and now you are saying it is OK . . . ' I am not saying it is OK, I am just telling you what is happening. At any rate, I like the conclusion of the study, as the authors cite Wood (2001) from a book that I edited more than a decade ago. Wood (2001: 82) stated,

> If ISE [International Scientific English] is not the property of the native speaker but of scientists of any language background, it can become the property of student scientists as well, who will integrate the language into their developing practice. . . in a way which would be impossible if English is seen as the property of the English-speaking world. . .

https://doi.org/10.1017/S0261444814000378 Published online by Cambridge University Press

Now Wood was looking forward, back in 2001, when he wrote this, and he was predicting correctly, as in leading engineering journals, at least, this is what is happening. The scientists from different nationalities are deciding what is acceptable and what is not.

7. Applications of corpus approaches

7.1 Indirect and direct applications

Finally, I am going to talk very briefly about the indirect applications and direct applications (Flowerdew 2009) of the corpus approach in more general terms. One of the main areas of indirect application is the compilation of dictionaries. Most dictionaries now are corpus-based, the earliest example being *Collins COBUILD English Language dictionary*, but there are now many others. Corpora have also been used in the creation of reference grammar books. *Collins COBUILD English grammar*, was the first one here, but perhaps more used now is the *Longman grammar of spoken and written English* (Biber et al. 1999) (See also Biber & Conrad 2009 and Biber, Conrad & Leech 2011). One problem in this area of potential application is the demise of the English for science and technology (EST) textbook. It used to be very popular in the 1980s but now has died out. I don't know of any corpus-based or corpus-informed commercially available EST or ESP course books, which is a pity.

Turning now to direct application, this is where students are encouraged to use the corpus tools themselves. Bloch (2008) suggests 'Concordance programmes have numerous features that allow students and teachers to explore various aspects of language use' (p. 178). This is often referred to as 'Data-driven learning' (DDL) and was initiated by Tim Johns at Birmingham University, UK, with the idea that students may become 'language detectives' (Johns 1997: 101). Learners identify and analyse recurrent patterns in concordance lines and make their own generalizations. As Bloch again pointed out, '[t]he idea behind this approach to language is not to rely on often unhelpful rules but to construct a possible answer based on the available evidence, test hypotheses and make generalizations' (Bloch 2008: 186). One problem with this DDL approach is the danger of plagiarism, as students may reuse large chunks of texts from the corpora they are working with. This is an issue I have explored in an article with Yongyan Li (Flowerdew & Li 2007). On the other hand, during the activities, students may also be alerted to this danger and learn to distinguish bona fide language re-use from plagiaristic activities.

7.2 Some caveats

John Swales is the director of the MICUSP and MICASE projects but he is a bit hesitant about applications of corpus findings. Back in 2002, in a book I edited, he said, 'there are the relatively meagre mature fruits of an investigative harvest in which the great majority of investigations have died on the vine' (Swales 2002: 164). Just before I submitted the manuscript to the publisher, Swales wrote to me and said his colleagues in the MICASE

https://doi.org/10.1017/S0261444814000378 Published online by Cambridge University Press

project were not happy with what he had written and asked if he could make it less negative. And I said, 'no, no . . . I think it's important to be a bit humble and not to say we can conquer the world with corpora'. In another (published) interview, Swales also talked about the value of the old approach to discourse analysis: '[T]he very act of eyeballing the texts produces various kinds of insights not so easily available via a few quick clicks of a mouse' (Swales 2011: 222). So I think, in terms of language description for specific purposes writing, we need to be careful not to throw the baby out with the bathwater and abandon previous approaches.

8. Conclusion

The sort of corpus findings I have presented today can be very useful for disciplinary consciousness-raising. Nesi & Gardner (2012) suggest that 'writing classes need to teach students about genres of writing, and writing teachers need to be able to discern what the key features of these genres are' (p. 4). The findings from corpora can also be directly incorporated into teaching materials. You can, furthermore, use corpora and let the students work on corpora themselves, i.e. DDL.

Also, in this talk I tried to revisit the issue of specificity. Specificity is a big issue in ESP. Some people say we need to be very specific. Others say we need to have a wide angle. The corpus approach would lead you in the direction of the more specific the better, other things being equal; an academic corpus is better than a general one; a discipline-specific corpus is better than a broader one; and a genre-specific corpus is better than one covering various genres. This is not to discount, however, the possibility of comparing corpora from an inter- or intra-disciplinary perspective, the caveat that I mentioned at the beginning of my talk.

Acknowledgements

I would like to thank Professor Carol Chapelle of Iowa State University and her team for inviting me to give this plenary address and for hosting me at their university. I would also like to thank Simon Ho Wang for his careful transcription of the recording of the speech and subsequent assistance with editing it. Finally, I thank the three anonymous reviewers who gave me useful feedback on an earlier version of this transcript.

References

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.
Biber, D. & S. Conrad (2009). *Real grammar: A corpus-based approach to English*. Harlow: Pearson/Longman.
Biber, D., S. Conrad & G. N. Leech (2011). *Longman student grammar of spoken and written English. [1]*. Harlow, Essex: Longman.
Biber, D., S. Johansson, G. Leech, S. Conrad, E. Finegan & R. Quirk (1999). *Longman grammar of spoken and written English*. London: Longman.
Bloch, J. (2008). *Technology in the L2 composition classroom*. Ann Arbor: University of Michigan Press.
Coxhead, A. (2000). A new academic word list. *TESOL Quarterly* 34.2, 213–238.

https://doi.org/10.1017/S0261444814000378 Published online by Cambridge University Press

Flowerdew, J. (2009). Corpora in language teaching. In M. H. Long & C. J. Doughty (eds.), *The handbook of language teaching*. Oxford: Wiley-Blackwell, 327–335.
Flowerdew, J. & R. Forest (2009). Schematic structure and lexico-grammatical realization in corpus-based genre analysis: the case of Research in the Ph.D. literature review. In M. Charles, D. Pecorari & S. Hunston (eds.), *Academic writing: At the interface of corpus and discourse*. London: Equinox, 15–36.
Flowerdew, J. & R. Forest (2015). *Signalling nouns in English: A corpus-based discourse approach*. Cambridge: Cambridge University Press.
Flowerdew, J. & Y. Li (2007). Language re-use among Chinese apprentice scientists writing for publication. *Applied Linguistics* 28.3, 440–465.
Friginal, E. (2013). Developing research report writing skills using corpora. *English for Specific Purposes* 32, 208–220.
Hyland, K. (2000). *Disciplinary discourses: Social interactions in academic writing*. Harlow: Longman.
Hyland, K. (2012a). Bundles in academic discourse. *Annual Review of Applied Linguistics* 32, 150–169.
Hyland, K. (2012b). *Disciplinary identities: Individuality and community in academic discourse*. Cambridge University Press.
Johns, T. (1997). Contexts: the background, development and trialling of a concordance-based CALL program. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (eds.), *Teaching and language corpora*. London: Longman,100–115.
Kachru, B. (1992). *The other tongue: English across cultures*. University of Illinois Press.
Kwan, B. (2006). The schematic structure of literature reviews in doctoral theses of applied linguistics. *English for Specific Purposes* 25.1, 30–55.
Leech, G. (2011). Frequency, corpora and language learning. In G. G. & F. Meunier, S. De Cock, G. Gilquin & M. Paquot. (eds.), *A taste for corpora. In honour of Sylviane Granger*. Amsterdam/Philadelphia: Benjamins, 7–32.
Nesi, H. & S. Gardner (2012). *Genres across the disciplines: Student writing in higher education*. Cambridge: Cambridge University Press.
Paquot, M. (2010). *Academic vocabulary in learner writing*. London and New York: Continuum.
Rozycki, W. & N. H. Johnson (2013). Non-canonical grammar in Best Paper award winners in engineering. *English for Specific Purposes* 32.3, 157–169.
Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
Swales, J. M. (2002). Integrated and fragmented worlds: EAP materials and corpus linguistics. In Flowerdew, J. (ed.), *Academic discourse*. London: Longman, 150–164.
Swales, J. M. (2011). A critical view on the use of corpora. In V. Viana, S. Zyngier & G. Barnbrook (eds.), *Perspectives on corpus linguistics*. Amsterdam/Philadelphia: John Benjamins, 221–227.
Wood, A. (2001). International scientific English. In J. Flowerdew & M. Peacock (eds.), *Research perspectives on English for academic purposes*. Cambridge: Cambridge University Press, 71–83.

John Flowerdew is a professor in the Department of English, City University of Hong Kong. He has published widely in the leading applied linguistics, language teaching and discourse analysis journals, focusing on academic, educational and political discourse. His books include *Second language listening: Theory and practice* (with Lindsay Miller) (Cambridge, 2005); *Advances in discourse studies* (with Vijay Bhatia and Rodney Jones) (Routledge, 2008); *Critical discourse analysis in historiography* (Palgrave Macmillan, 2012); *Discourse in English language education* (Routledge, 2013); and *Discourse in context* (Bloomsbury, 2014). His latest book (2015) is *Signalling nouns in English: A corpus-based discourse approach* (with Richard W. Forest) (Cambridge University Press).

https://doi.org/10.1017/S0261444814000378 Published online by Cambridge University Press