Hyland, K. (2015). Corpora and written academic English. In The Cambridge Handbook of English Corpus Linguistics, In D. Biber & R. Reppen. Cambridge University Press. pp 292-308.

Corpora and written academic English

Ken Hyland

The impact of corpora in the study of written academic English over the past twenty years has been enormous, transforming how we understand, study and teach this key area of language use. Corpora provide language data which represent a speaker's experience of language in a particular domain and so therefore offer evidence of typical patterning of academic texts. It is a method which focuses on community practices and the ways members of particular disciplines understand and talk about the world. Bringing an empirical dimension to the study of academic writing allows us not only to support intuitions, strengthen interpretations and generally to talk about academic genres with greater confidence, but it contrasts markedly with impressionistic methods of text analysis which tend to produce partial and prescriptive findings, and with observation methods such as keystroke recording, which seek to document what writers do when they write. It also differs from methods which employ elicitation methods such as questionnaires and interviews or introspection methods like think aloud protocols to understand the perspectives of writers or readers on how they use texts.

Perhaps most significantly, corpus approaches to academic writing provide insights into disciplinary practices which helps explain the mechanisms by which knowledge is socially constructed through language. Together, this research explicitly contradicts the view that Corpus Linguistics takes an impoverished, decontextualised view of texts and replaces it with a detailed picture of how students and academics write in different genres and disciplines. In this chapter I discuss some of the key studies and ideas which contribute to our understanding of academic writing in English. Section 1 offers an overview of published studies while Section 2 describes a study which illustrates how corpus research can inform our understanding of academic writing.

Section 1: Research into academic writing in English

This section discusses previous research, identifies a number of key studies, and provides an overview of the research methodologies that have been employed.

a. A brief survey of research

The textual data for studying academic writing includes all the ways of using language in the academy. This is a range of genres which enact complex social activities like educating students, demonstrating learning, disseminating ideas, evaluating research and constructing knowledge, and almost all have been collected and analysed as corpora. Studies of these corpora reveal that all academic texts are, in one way or another, designed to persuade readers of something. In most cases this is the efficacy of an idea or piece of research, so that claims are encoded, warrants employed, arguments framed and appropriate attitude to readers conveyed in ways that a potential audience will find most convincing. Thus, the ways academics represent themselves in bios, webpages and prize applications, for example, seek to persuade readers of their competence and expertise as disciplinary insiders by drawing on attributes and experiences which relate the individual closely to what is valued in their community (e.g. Hyland, 2012).

More specifically, the comparison of features in a corpus of 240 research articles and 56 textbooks (Table 1) shops how arguments are constructed to persuade different audiences in the two genres.

Table 1: Selected features in Research articles and textbooks

per 1,000 words	Hedges	Self-mention	Citation	Transitions
Research Articles	15.1	3.9	6.9	12.8
University Textbooks	8.1	1.6	1.7	24.9

The greater use of *hedging*, for example, underlines the need for caution and opening up arguments in the research papers compared with the authorized certainties of the textbook, while the removal of *citation* in textbooks shows how statements are presented as facts rather than claims grounded in the literature. The greater use of *self-mention* in articles points to the personal stake that writers invest in their arguments and their desire to gain credit for claims. The higher frequency of transitions, which are conjunctions and other linking signals, in the textbooks is a result of the fact that writers need to make connections far more explicit for readers with less topic knowledge. Thus, to achieve their persuasive purposes academics draw on the same repertoire of linguistic resources again and again. This is, in part, because writers try to anticipate their readers' background knowledge, processing needs, and rhetorical expectations through use of familiar rhetorical features. It is these patterns of repetition which corpus analyses seek to uncover.

Corpus analyses have, for example, been productive in identifying the structural regularities of a range of genres, describing moves in grant proposals (Connor & Upton, 2004), in dissertation acknowledgements (Hyland, 2003), in application statements for medical and dental school (Ding, 2007) and in PhD theses (Bunton, 1999). They have also described the patterns within moves, with research article introductions (e.g. Ozturk, 2007) and results sections (e.g. Bruce, 2009) receiving the considerable attention from analysts. Beyond moves, corpus analysts have also identified a range of key features which have previously gone largely unnoticed, such as 'attended and unattended *this'* (Wulff, et al, 2012), *evaluative that* (Hyland & Tse, 2005), code glosses (Hyland, 2007) and the use of *this* and *these* as pronouns (Gray & Cortes, 2011). Several studies have also shown that academic writing in English is composed far more of fixed phrases than was previously supposed (Biber, et al, 2004).

Corpus research has also enabled researchers to make comparisons across different corpora. Thus Biber's (2006) work, for instance, confirms differences in spoken and written texts, such as the fact that lectures contain many features of conversation and comprise a series of relatively short clauses, while "university textbooks rely heavily on complex phrasal syntax rather than clausal syntax" (Biber, 2006:

5). Considerable work has also sought to distinguish written genres from each other so, in research articles, abstracts differ from introductions (Samraj, 2005) and features such as hedges, self-mention and transition signals all differ considerably between articles and textbooks (Hyland, 2008a). Comparisons have also been made between the ways men and women write academic texts (e.g. Tse & Hyland, 2008), and how experts and novices write them, revealing for example, differences in the use of reader engagement, (Hyland, 2006) and bundles (Hyland, 2008b).

Perhaps comparison has been explored most extensively in the effects that culture and/or first language has on writing in English. Culture, seen as an historically transmitted and systematic network of meanings, is inextricably bound up with language and so influences writers' expectations about appropriacy, audience, ways of organizing ideas and of structuring arguments. Corpus research has broadly supported the view that the schemata of L2 and L1 writers differ and influence how they write in English (e.g. Loi, 2010; Moreno & Suárez, 2008). Much of this work has focused on student genres and has identified a range of different features in first and second language writing in English (e.g. Hinkel, 2002). More recent research, however, has focused on different perceptions of interpersonal appropriacey and self-representation in academic writing in different languages so that first person pronouns, for instance, have been found to be far more frequent in English research articles than those written in Italian (Molino, 2010) and Spanish (Dueñas, 2007).

Yet corpus studies show discipline to be a decisive factor in the construction of academic genres. Individuals use language as members of social groups and they write essays, theses and articles by framing problems and understanding issues in ways specific to their disciplines (Hyland, 2004). While discipline is something of a contested term, corpus studies allow us to say something of how they are created and maintained, along with the knowledge they establish, through the routine rhetorical preferences of their members. They show, for example, that there is considerable variation in writers' choices of academic lexis. Hyland & Tse (2007) found that the so-called 'universal' sub-

technical items from the Academic Word List (Coxhead 2000) vary enormously across disciplines in terms of range, frequency, collocation, and meaning. The frequency and use of citations also differ, being about twice as frequent in the soft fields where the literature is more dispersed and the readership more heterogeneous than in the hard sciences, so writers cannot presuppose a shared context but have to build one far more through citation (Hyland, 2004).

Finally, it is worth mentioning the support corpus studies have provided to the view that academic writing is permeated by social interaction and intersubjectivity. These concepts have become central to language studies in recent years as we have come to realize that academics do not simply produce texts that talk about the world, but use language to acknowledge, construct and negotiate social relations. Corpora have helped illuminate the range of features writers use to construct an appropriate authorial self. So, the considerable use of self-mention in research articles (Hyland, 2001), abstracts (Bondi & Silver, 2004) and undergraduate theses (Hyland, 2002), for example, suggests that academic writing is not a self-evidently objective and impersonal discourse. Genres such as textbooks (Bondi, 2012), student essays (Matsuda & Jeffery, 2012) and book reviews (Hyland & Diani, 2009) have been explored to identify how writers seek to construct and negotiate participant relationships. This has also had an impact on the development of interpretive approaches to interaction such as metadiscourse, stance, and appraisal.

Despite considerable work on genre structure, however, the ways that moves are signaled by writers and identified by readers has been far less studied. Typically, researchers have relied on changes in discourse function, or what particular stretches of a text are contributing to the overall purpose of the discourse. Often these are explicitly signaled, so that in this example from a research article abstract in biology we see a purpose statement announced by a to + infinitive clause and a method move indicated by a switch to past tense active verbs:

> To study the expression of ALDHs in plants we isolated and characterized a cDNA coding for a putative mitochondrial ALDH (TobAldh2A).

More generally, it is likely that particular rhetorical features cluster within particular moves to perform the specific functions of those moves, but more work needs to be done to identify these frequently occurring signals.

Another criticism of academic corpus studies is, until fairly recently, these have been largely text-focused so that features seem rather abstract and disembodied from real users. Studies are needed that do not just analyze text corpora but which involve the authors or the readers of the texts in the analysis by also collecting interview data. Finally, there are still many gaps as researchers have been tempted to build and analyze corpora of the most publically prominent and easily accessible genres, typically published texts and student work. This tends, however, to skew research towards a narrow area of the academy and neglects more 'occluded' genres (Swales, 1996). These are genres which, unlike published texts, are less public and accessible, such as those which sit behind the publication process such as reviewers comments on submissions to journals, or applications for promotions or prizes.

b. Some key studies

The studies selected here represent both significant contributions to a particular area of academic writing and important moments in the evolution of corpus research in academic discourse.

Genres across the disciplines (Hilary Nesi and Sheena Gardner, 2012)

This is the first detailed description of the kinds of assessed writing students do in different disciplines and in different years of their studies in UK universities. Based on the 6.5 million word *British Academic Written English* (BAWE) corpus, the study develops a genre classification to identify and describe thirteen major types of assignment according to their purpose, stages, genre networks and characteristic language features. Each chapter in the book discusses a 'family' of genres, each with a particular social function. These are 'demonstrating knowledge and understanding' (e.g. explanations and exercises); 'developing powers of informed and independent reasoning' (e.g. critiques and essays); 'developing

research skills' (e.g. research reports and literature surveys); 'preparing for professional practice' (e.g. proposals and design specifications); and writing for oneself and others (e.g. empathy writing and narrative recounts).

The study is also a good example of how corpus techniques can be used to map patterns in a large collection of texts, identifying moves in different genres, comparing frequencies of various language features and offering detailed description of how individual words and phrases are used. The study is especially important as it provides a detailed account of undergraduate writing and descriptions of several previously disregarded genres. More importantly, it underlines the significance of disciplinary variation, showing how some genres are found almost exclusively in certain fields and how genres change as students progress through their course of study. This information is not only useful to discourse analysts, but also to teachers and those involved in syllabus and materials design for students in higher education.

University language: a corpus-based study of spoken and written registers. (Doug Biber, 2006).
This book discusses the genres which confront students at US universities, both inside and outside the classroom, addressing their linguistic characteristics to provide a more representative basis for constructing and validating the TOEFL test. The study draws on the TOEFL 2000 Spoken and Written Academic Language (T2K-SWAL) corpus of 2.7 million words and comprises both spoken genres (such as study groups, class sessions and service encounters) as well as 1 million words of written texts. The written genres are textbooks, course packs (lecture notes, study guides, readings, etc.) and course management materials collected from six disciplines and three levels, together with a range of varied institutional texts such as brochures, programme descriptions and student handbooks. The corpora were grammatically annotated using an automatic tagging programme which identified part-of-speech categories for individual words and multi-word grammatical units such as *that is* and *for example*.

The descriptions offer important characterizations of academic writing so, for example, Biber found that over 50% of all nouns in the written corpus had abstract/process meanings that refer to intangible concepts or processes (*system, factor, difficulty*). The study also confirms the variation between written and spoken texts, with textbooks containing twice as many different words as classroom teaching, despite their broadly similar instructional purposes, largely due to their use of specialized lexis. The book is also interesting for its methodological approaches which includes Multi-Dimensional analysis to identify sets of linguistic features that commonly co-occur with markedly high frequencies in texts. Thus institutional writing and course management texts contain high frequencies of necessity and prediction modals, 2nd person pronouns and conditional adverbial clauses which push them towards the procedural end of a continuum with more content-focused genres such as textbooks and course packs at the other.

Disciplinary Discourses: social interactions in academic writing (Ken Hyland, 2004)

This work presents a series of studies focusing on eight disciplines and a variety of professional, rather than student genres, examining interactional features in corpora of research articles, book reviews, abstracts, scientific letters and undergraduate textbooks. Together, these studies explore how academics use language to both create knowledge and define their academic allegiances. They show how writers present their topics, signal disciplinary membership, and stake their claims through careful negotiations with, and considerations of, their colleagues. The book illuminates how disciplinary constraints on discourse are both restrictive and authorizing, allowing academics to construct credibility and agreement through features such as citation, hedges and boosters, claims for novelty and significance and metadiscourse. Methodologically the studies use the simple techniques of frequency and collocation, and combines these with interviews with academics to gain an understanding of how insiders view their literacy practices and see their participation in their disciplines. This data was collected through unstructured interviews and more focused 'discourse based interviews' about particular pieces of writing. The approach is therefore designed to focus on *discourse*, a process of social interaction, rather

than just *texts*, by giving explicit attention to user perspectives and the social institutions within which they work.

Learner English on computer (Edited by Sylviane Granger, 1998)

This was the first book to introduce readers to the field of corpus-based research into written learner language. Edited by the founder and co-ordinator of the International Corpus of Learner English (ICLE), the book's chapters offer a comprehensive overview of all aspects of corpus compilation, design and analysis, providing readers with both an understanding of methodological approaches to the field and analytical insights into aspects of academic writing by L2 students writing in English. The opening chapters review the software tools available for analysing learner language and give examples of how they can be used. The second part of the book contains eight case studies in which computer learner corpora are analysed for various lexical, discourse and grammatical features such as overstatement, phraseology, direct questions and adverbial connectors. In the third part, authors explore the application of how Computer Learner Corpus (CLC) studies can help improve pedagogical tools, such as learner grammars, dictionaries, writing textbooks and online writing tools. Collectively the studies offer a compelling argument for the role of corpora in in SLA research, which has tended to favour introspective and experimental data.

If you look at… lexical bundles in university teaching and textbooks. Doug Biber, Susan Conrad & Vivianna Cortes, 2004).

This is a pioneering corpus-based study of multi-word sequences in academic discourse. Taking a frequency-based approach, the research identifies the lexical bundles in classroom teaching and textbooks, and compares them to those in the authors' previous research on conversation and academic prose, showing that bundles are neither complete grammatical structures or idiomatic expressions, but function as the basic blocks for the creation of discourse. The study classifies the bundles by their structural patterns and by a functional taxonomy which includes stance expressions, discourse

organizers and referential expressions. The analyses show that classroom teaching uses an extremely wide variety of different bundles in comparison to the other genres and that the written genres contained relatively few stance and organizing bundles. The study not only adds to our understanding of academic writing, but offers a clear definition of lexical bundles and a structural and pragmatic description of bundle types. While the frequency cut-offs and spread across a given number of texts in a corpus have subsequently been increased to strengthen the criteria for identifying bundles in later studies, the structural and functional descriptions have offered both teachers and researchers a useful inventory of sequences and inspired a number of further studies in the area.

Disciplinary identities (Ken Hyland, 2012)

This study extends corpus research into a new area: the relationship between author identity and disciplinary practice. Drawing on corpora which include academic bios, acknowledgements, undergraduate essays, academic homepages, book reviews and prize applications, the analyses seek to show how we can understand identity as a performance of writers which is informed and re-inscribed over time through their use of language in disciplinary communities. What we say and write aligns us with or separates us from other people and other positions, so the command of a disciplinary idiom can therefore be an assertion of oneself as a particular kind of person: one who has a right to be taken seriously in the academic world. By studying how language is routinely used in particular genres it is possible to see how disciplinary identities are performed and recognised as legitimate.

In most cases the analyses start by focusing on potentially productive items from interviews with writers or prior studies, while in other cases the task of identifying features is delegated to the computer, generating lists of high lexical items and keywords for further study. Items identified from either of these starting points then provide the basis for investigation through collocation and comparisons to see how particular academics and disciplinary communities used these features to express social identities.

Such approaches reveal the regularity and repetition of what is socially ratified and independently variant and in so doing offer insights into the preferred practices of both individuals and collectivities.

c. Corpus methods in studies of academic writing

Perhaps most corpus studies of academic writing have followed what Tognini-Bonelli (2001) calls a *corpus-based* approach where the researcher begins with a pre-selected list of potentially productive items and uses the corpus to examine their frequencies and the ways they behave in different contexts. This is, for example, how researchers have used corpora to study features such as self-mention (Hyland, 2001) and passive voice (Xiao et al, 2006). Researchers have not ignored more inductive *corpus-driven* studies, however, where the corpus provides the basis for frequencies and patterns. One example is the research on lexical bundles, exemplified by Biber et al's (2004) study to identify the most common multi-word patterns in textbooks and classroom teaching discussed above. In both approaches, corpus studies of academic writing typically use of the tools of frequency, keyness, concordance and annotation.

Frequency provides evidence of non-randomness, revealing what regularities, and exceptions, exist in the language use of a group of people when engaged in a particular activity. High frequency items represent repeated, taken-for-granted choices in academic writing as, from all the different ways of saying roughly the same thing, members of individual disciplines select the same items again and again. In-group abbreviations, acronyms, shorthand names for methods and theories, preferred argument patterns, preferences for author visibility or anonymity, particular lexical bundles and so on, all help define and identify disciplines and genres.

Frequency can therefore lead us to what is worth discussing as it often indicates what is *salient* for groups of language users. We find, for example, that all disciplines shape words for their own uses, as

demonstrated by their clear preferences for particular meanings and collocations. Thus science and engineering students, for example, are very unlikely to come across the noun *volume* in the meaning of "a book or journal series" while the noun *strategy* has different associations across disciplines, often appearing in the multi-word unit *marketing strategy* in business, *learning strategy* in applied linguistics and *coping strategy* in sociology (Hyland & Tse, 2007). Everything we know about a word is a result of our encounters with it, so that when we formulate what we want to say, our wordings are shaped by the way we regularly encounter them in similar texts. This helps to explain why it is, for example, of all the different ways of expressing thanks, over a third of all gratitude in PhD acknowledgements is expressed as nominals (*My sincere thanks to; My gratitude to*) (Hyland, 2003).

Keyness is another frequency approach often used in studies of academic writing. The basic idea is that a word form or cluster of words which are common in a given text are *key* to it, it is what the text is 'about' or 'what it boils down to …once we have steamed off the verbiage, the adornment, the blah blah blah' (Scott & Tribble, 2006). The text analysis programme *Wordsmith Tools* (Scott, 2007), identifies keyness by comparing frequencies in one corpus against those in another to determine which ones occur *statistically* more frequently using a log-likelihood statistic. This gives a better characterization of the differences between two corpora than simple frequency comparison as it identifies items which occur with unusually high frequency and so which are most prominent and not just common.

Keywords are therefore useful for identifying which words best distinguish the texts of a particular author of group of authors from another. Comparing different disciplines, for example, Scott and Tribble (2006), found the most frequent keywords in Humanities to be *of, the, in, early, war, theory, as, century* and *between*; in Medicine to be *clinical, patients, treatment, disease, of, study,* and *diagnosis;* and in the Natural sciences to be *are, Fig, shown, observe, sequence, obtained, surface, and analysis*. Similarly, Granger and Paquot (2009) identified the lexical verbs that are prominent in business, linguistics and medicine compared with a one million word reference corpus of fiction writing. While recognizing the

different meanings these might have in different fields, their keyword analysis identified potential candidates for inclusion in a list of EAP verbs, finding 106 shared key verbs. These largely consisted of verbs that typically serve organisational or rhetorical functions in academic writing: reviewing the literature (*maintain, present*), describing research (*investigate, describe*), reporting (*show, identify*), expressing cause and effect (*suggest, result*), describing tables and figures (*illustrate, highlight*) and contrasting and summarizing (*summarise, compare*).

Keyness therefore reveals a kind of interdiscursive similarity and helps build a picture of particular disciplines and how they are distinctive from each other. It also offers a starting point for *corpus-driven* investigations of academic corpora by generating list of items which can be further explored in more detail using concordance analyses.

Concordances While frequency lists provide information about the *focus* of a collection of texts, they don't tell us how words are actually used. This is the function of concordance analyses, which provide information about users' preferred meanings by displaying repeated co-occurrence of words, allowing us to see the characteristic associations and connections they have and how they take on specific meanings for particular individuals and in given communities. One example of this is Hyland & Tse's (2012) study of bios and how collocation allows us to see differences in the ways that senior academics and graduate students refer to themselves in the bios accompanying research articles. So, by checking the frequency of definite, indefinite and 'zero' articles in a corpus of bios and then looking at concordance lines for each, we find that professors are far more likely to use naming terms that collocate with definiteness (*she is professor of, he is the author of*) which serve to uniquely identify them. In the bios of students and non-professorial faculty, on the other hand, such attributive choices signal class membership rather than a unique identity (*she is a PhD student, he is an editor of*).

Annotation refers to adding linguistic information to a corpus. While a raw corpus is a highly useful resource, annotation provides an extra layer of information, which can be counted, sorted and compared. Lemmatizers, for example, retrieve word lemmas, the "canonical root" of a word such as *cook* from *cooking, cooks, cooked*, but while potentially useful for lexical analyses and mapping semantic relationships they are rarely used in academic writing research. POS-tagged corpora, on the other hand, are very powerful resources and have contributed to our understanding of academic writing by allowing for detailed studies of the use of grammatical categories, such as prepositions, phrasal verbs, modals, passives, etc., although the search and retrieval possibilities depend on the sensitivity of the tagset which can range from 50 to 250 tags.

One example of how a POS tagger has been used in academic writing is Granger and Rayson's (1998) identification of salient features of interlanguage essays. Using a reduced tagset of nine major word categories and 14 subcategories from Claws4, the analysts compared a corpus of argumentative essays by advanced French-speaking learners of English with a corpus of similar writing by Native English writers. While both groups were found to use articles, adjectives and verbs with similar frequencies:, the non-native speaker writers overused determiners, pronouns and adverbs significantly and significantly underused conjunctions, prepositions and nouns. Hinkel's (2003) study also discovered significant differences between the structures and lexical forms used by native and nonnative writers. Hinkel tagged her corpus of 1,083 essays corpus by hand because the texts were hand-written in class. She found that advanced nonnative-English-speaking students employed simple syntactic and lexical constructions, such as *be* -copula as the main verb; predicative adjectives; vague nouns; and public, private, and tentative verbs, at rates significantly higher than those found in texts by native English speakers. Both studies point to the fact that the academic writing of non-native English speaking learners displayed many of the stylistic features of spoken, rather than written, English.

Sinclair (1991), however, has cautioned against tagging as it disguises the interdependency between grammar and lexis by setting up artificially imposed categories on the language and so prevents researchers from seeing unnoticed patterns in the text. The 'probabilistic tendencies of language' is the basis of the pattern grammar approach advocated by Hunston and Francis (2000) which encourages researchers to allow the text to throw up new insights, a philosophy which underlies the 'corpus-driven' method (Tognini Bonelli, 2001), discussed above, where the corpus data are not pre-defined in terms of a particular theory of grammar. These different ways of approaching data remind us that it is always possible to talk about the same thing in different ways and that our descriptions of language use encodes particular assumptions and points of view. So, while grammatical analyses of academic writing using tagged corpora are likely to continue to offer insights into how language works in this domain, researchers should not ignore more exploratory methods which follow Sinclair's (2004) exhortation to 'trust the text'.

d. A summary of findings and gaps

Some key findings:

1. That academic texts are persuasive and structured to secure readers' agreement;
2. That there are variations in spoken and written academic genres;
3. That language groups have different ways of expressing ideas and structuring arguments;
4. That ways of producing agreement represent disciplinary specific preferences;
5. That academic persuasion depends on negotiating appropriate interpersonal relations.
6. That authors are everywhere in their texts, presenting a stance towards their topics and readers.
7. That academic texts are constructed through fixed phrases to a greater extent than we expected.
8. That academic conventions constrain both meanings and author identities, but also provide the resources for creativity and agency.

Some major research gaps which remain to be addressed:

1. We need more descriptions of the wide range of specific disciplinary genres students need to write and read.
2. We need a greater understanding of how particular genres are used within specific contexts, adding a focus on 'action' to balance the focus on 'language' by including research techniques such as interviews and observations in what Swales calls a 'textography' (Swales, 1998)
3. We need to expand corpus studies into multimodal academic genres where writing is frequently used with graphical and visual semiotic forms, such as academic websites and textbooks.
4. We need studies which focus on NNES students and how their academic writing in English is similar and distinct from each other and from NESs.
5. We need more studies to help us understand the nature of disciplinary identities and the meaning of expertise in particular fields.

Section 2: An Example study

As an illustration of corpus research, I want to consider a study which attempts to see what corpus research can contribute to the study of identity (Hyland, 2010). The study is important as it seeks to move away from the traditional ways of studying identity through narrative recounts or interviews to ground discussions of identity in what people actually *do* rather than what they *say* about themselves. There are two major innovations in this study: the use of interview methods to complement corpus research, adding a subjective, 'emic' perspective which is not normally considered in corpus research; and the use of corpus data to 'go beyond' claims made in interviews or decisions made on particular occasions of writing to explore the regularity and repetition of what is socially ratified and independently variant and therefore what represents the preferred practices of individuals and collectivities.

Background and rationale

Identity has come to be seen as something that we actively and publicly accomplish in our interactions with each other (e.g. Benwell & Stokoe, 2006) so it doesn't exist *within* individuals but *between* them. But identity research is largely characterized by autobiographical and interview methods which underplay the fact that our identities must accord with the responses and behaviours of others (Hyland, 2012). Language allows us to create and present a coherent self to others because it ties us into webs of commonsense, interests and shared meanings. *Who we are* is built up through participation in social communities and linked to the rhetorical strategies and positions we adopt in engaging with others on a routine basis. We construct an identity from our consistent patterns of rhetorical choices over time.

Academic contexts obviously privilege certain ways of making meanings, but we can also see these writing conventions as options which allow writers to actively accomplish an identity through discourse choices. This is because it is through our use of community discourses that we claim or resist membership of social groups, defining who we are in relation to others. This suggests it might be productive to compare the writing of individuals with the general practices of their discipline to find evidence for identity construction. How do their choices help them achieve credibility as insiders and reputations as individuals?

The main investigative technique in this study was therefore comparison. Comparing the features of target writers' texts with a much large reference corpus of work in the same discipline can help to determine what is general in the norms of a community and what represents more personal choices. We can, in other words, see that if a particular word, phrase or usage is common in a corpus of a particular writer's work, then it might be said to be a consistent preference which reveals something of that individual's routine expression of self: of a relatively unreflective performance of identity. To capture this I compiled a corpus from each of the published single-authored work of two experienced and well-known applied linguists, Deborah Cameron and John Swales. I selected these two academics partly because of their celebrity in the field of

applied linguistics and their contrasting personalities and careers, but largely because their highly distinctive rhetorical styles seemed to offer a good starting point for this kind of analysis.

Corpora and methods

My corpus of Cameron's published writing consists of 21 single authored papers made available by the author. It represents some 20 years of publishing and comprises 125, 000 words. The Swales corpus was compiled at the Michigan ELI and consists of 14 single-authored papers together with the bulk of his three monographs, representing 18 years of output and comprising 342, 000 words. These corpora were individually compared with a larger reference corpus representing a spectrum of current published work in applied linguistics and in the same genres as the target texts. It comprises 75 research articles from 20 leading international journals and 25 chapters from 12 books totalling 750,000 words.

Wordsmith Tools 5 (Scott, 2007) was used to generate word lists of the most frequent single words, and three- and four-word *lexical bundles* used by each of these two authors. I then compared each author corpus with the reference corpus using the KeyWords tool. As noted above, this program identifies words and phrases that occur statistically significantly more frequently in the smaller corpus. This meant I could identify which words best distinguished the texts of these authors from those in applied linguistics more generally as represented in the reference corpus. After reviewing the keyword lists and identifying individual words and lexical bundles, I concordanced the more frequent items to group common devices into broad pragmatic categories to capture central aspects of their writing. In other words, my approach was to use the corpora as a starting point of an investigation into the preferred rhetorical practices of these two writers. The corpus methods I used were unexceptional, but I hoped the repeated patterns they revealed would provide a basis for understanding the routine ways these writers interacted with members of their community and allow an interpretation of their discoursally constructed identities.

Main findings

The high frequency content words and keywords indicate the niche of specialization which these academics have carved out from the mass of disciplinary subject matter. They reflect the main themes of an individual's work and serve as motifs for their contribution to the field. Items such as *women, language, gender, men, social, talk, discourse,* and *work* indicate Cameron's concern with the ways language functions to structure social relations, particularly in work contexts, and in the ways gender linked patterns of language-use are made significant in social relations. The top content items from the Swales corpus are *research, genre, English, academic, writing, non-native speakers of English and the concept of discourse community* which similarly encompass his key areas of contribution.

More interesting, however, are the non-content words and phrases in the keywords lists which emerge as consistent individual choices. One example from Cameron's writing is the significantly above average use of *is,* which was the fifth most Keyword in her corpus. While one of the most common words in English, in Cameron's texts it often occurs in the pattern *it is+Adj.+to infinitive* (161 times). This structure not only shifts new or complex information towards the end of a sentence, to the rheme, where it is easier for readers to process, but asserts the writer's opinion and recruit the reader into it (e.g. It is important to; It is difficult to think of). This assertiveness in Cameron's authorial positioning is also realized through the frequency with which *is* occurs in the company of *that* (230 times) in '*evaluative that*' constructions (Hyland & Tse, 2005). Here a complement clause is embedded in a super-ordinate clause (It is my view that; it is problematic that), making the attitudinal meaning the starting point of the message and the perspective from which the content of the *that*-clause is interpreted.

John Swales, on the other hand, projects a very different identity. Here is an altogether more self-effacing and conciliatory writer, projecting a cautious and colleague using rhetorical choices which impart a clear personal attitude and a strong *interpersonal* connection to his readers, particularly through the use of self-mention and hedges. Both devices project the author as a participant in the text, indicating that the writer

is prepared to debate issues and contributing half of a dialogue with readers. Frequent use of the first person is perhaps the most striking feature of Swales' discourse, with both *I* and *my* occurring in the top ten keywords. Self-reference, in fact, occurs 9.1 times per 1000 words in the Swales corpus compared with 5.2 in the reference corpus, imparting a clear authorial presence of a thoughtful reflective colleague thinking through issues. An interesting aspect of Swales' identity is the extent self-mention is used in a self-deprecatory way, explicitly associated with modality, or at least a deliberative attitude. The most frequent main verbs related to *I* are *think* (86), *believe* (71), *suspect* (35), *hope* (33*), tried* (31) and *guess* (29), all of which point to some degree of tentativeness and care in handling claims and in dealing with the alternative interpretations and understandings of readers.

In Cameron's discourse then, the analyses reveal a range of features used to confidently and forcefully advocate a position, projecting a distinctive identity as a radical disciplinary expert. John Swales' choices, on the other hand, convey a clear personal attitude and a strong interpersonal, rather than intellectual, connection to readers, projecting the identity of a cautious colleague rather than a combative advocate of truth. Overall, the analyses suggest that the ways we write do not simply mimic community patterns but are a means of constructing who we are, or rather, how we would like others to see us.

Review of the study

The value of a corpus in this kind of research is that it can highlight what is common and what is individual. Stubbs (2005) puts it like this:

> individual texts can be explained only against a background of what is normal and expected in general language use, and this is precisely the comparative information that quantitative corpus data can provide. An understanding of the background of the usual and everyday - what happens millions of times - is necessary in order to understand the unique.

This methodology therefore points to new ways of understanding and exploring identity that takes us beyond what individuals say about themselves to what they do in interaction on repeated occasions, thus building a consistent persona through discourse.

In this view, identity can only be understood by close analysis of the ways writers routinely draw on the rhetorical repertoires of their communities to position themselves in recognisable ways as both individuals and as members of collectivities. It might be argued, however, that using corpora in this way fails to provide sufficient context to understand identity performance as it ignores the detailed biographies of interview techniques and perhaps draws instead on assumptions about the writers which are not in their texts at all. After all, we know something of these academics and their styles and I selected two of the most rhetorically aware individuals writing in applied linguistics today. Both writers are professional discourse analysts and so are highly attentive to the effects of their choices (see Hyland, 2010). My method, however, draws attention to an important aspect of corpus analysis: that although it is informed by numbers, largely frequency counts of keywords and collocations, it is ultimately constructed on interpretation. While repeated uses represent each writer's more or less conscious choices to project themselves and their work in particular ways, my take on them is necessarily subjective.

It is a methodology, however, which offers a way of exploring other unanswered questions about disciplinary constraints. Do all academic writers have a relatively consistent stylistic 'signature', for example, or is this something that only develops over time? Are novice writers more tightly constrained by conventions? What changes in their repertoire with greater experience and confidence? What variations exist across disciplines and between individuals in other fields? Not least it makes sense to address the wider political operation of discourse communities and to ask, with Bizzell (1989: 225), 'who gets to learn and use complex kinds of writing' and who has rights to manipulate or resist the conventions of a discipline rather than merely accommodate to them?

Conclusions

Corpus studies have made a considerable contribution to our understanding of academic discourse and revealed many of the ways that writers in different disciplines, genres and languages represent themselves, their work and their readers in different ways. In particular, they have shown that a range of features occur and behave in dissimilar ways in different disciplinary environments and underlined the importance of community, context and purpose in writing which has helped inform EAP course design and teaching.

It is this observation about students' target needs which helps clarify future directions for research. Quite clearly we need more descriptions of the specific disciplinary genres students need to write and read. Reports, essays, articles, critiques, presentations, case notes, lectures, and so on all differ across disciplines and knowledge of their structures and salient features can demystify them for learners. As corpus research into academic genres continues to grow, therefore, we can anticipate an ever increasing broadening of studies beyond texts to the talk and contexts which surrounds their production and use, beyond the verbal to the visual, and beyond tertiary to school and professional contexts. Corpus studies will, in tandem with other methods, have a continuing and important role to play in this endeavour.

It is important to mention, however, that generalising from a corpus will always be an extrapolation – it provides the evidence for interpretations about how language works. Intuitions remain in the explanations analysts bring to the data that is collected, making a corpus approach a unique combination of empirical analysis, deduction and human sensitivity.

References

Benwell, B. & Stokoe, E. 2006. *Discourse and Identity*. Edinburgh: Edinburgh University Press.

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

Biber, D., Conrad, S. & Cortes, V. (2004). *If you look at...*: Lexical bundles in university teaching and textbooks. *Applied linguistics*. 25L 371-405.

Bizzell, P. 1989. Cultural criticism: a social approach to studying writing. *Rhetoric Review* 7. 224-230.

Bondi, M. (2012), Voice in textbooks: Between exposition and argument. In Hyland, K. & Sancho Guinda, C. (eds.) *Stance and voice in written academic genres*. London: Palgrave. pp 101-117.

Bondi, M. & Silver, M. (2004). Textual Voices: A Cross Disciplinary Study of Attribution in Academic Discourse. In Anderson, L. & Bamford, J. (eds.), *Evaluation in Oral and Written Discourse.* Rome: Officina Edizioni. Pp 117-136.

Bruce, I. (2009). Results sections in sociology and organic chemistry articles: A genre analysis *English for Specific Purposes, 28, 2*: *105-124*

Bunton, D. (1999). The use of higher level metatext in PhD theses. *English for Specific Purposes*. 18, S41-S56.

Connor, U. & Upton, T. (2004). The genre of grant proposals: a corpus linguistic study. In Connor, U. & Upton, T. (eds.) *Discourse in the Professions*. Amsterdam: Benjamins. pp 235-255.

Coxhead, A. (2000) A New Academic Word List. TESOL Quarterly, 34(2), 213-238.

Ding, H. (2007). Genre analysis of personal statements: Analysis of moves in application essays to medical and dental schools. *English for Specific Purposes*. 26, (3): 368-392

Dueñas, PM (2007). I/we focus on…': A cross-cultural analysis of self-mentions in business management research articles. *Journal of English for Academic Purposes.* 6 (2): *143-162*

Granger, S. (Ed.) (1998). *Learner English on computer*. London: Longman,

Granger, S. & Paquot, M. (2009) In search of General Academic English: a corpus driven study. In Katsampoxaki-Hodgetts, K. (ed). *Options and practices of LSP practitioners conference proceedings.* University of Crete. pp 94-108.

Granger, S., and Rayson, P. (1998). Automatic profiling of learner texts. In S. Granger (ed.) *Learner English on Computer*. Longman, London and New York, pp. 119-131.

Gray, B. & Cortes, V. (2011) Perception vs. evidence: An analysis of *this* and *these* in academic prose *English for Specific Purposes, 30, 1*: *31-43*

Hinkel, E. (2002). *Second Language Writers' Text.* Mahwah, NJ: Lawrence Erlbaum.

Hinkel, E. (2003). Simplicity Without Elegance: Features of Sentences in L1 and L2 Academic Texts. *TESOL Quarterly*. 37, 2: *275-302.*

Hunston, S. & Francis, G. (2000). *Pattern grammar*. Amsterdam: Benjamins.

Hyland, K. (2001). Humble servants of the discipline? Self-mention in research articles. *English for Specific Purposes.* 20 (3). 207-226.

Hyland, K. (2002). Authority and invisibility: authorial identity in academic writing. *Journal of Pragmatics*. 34 (8): 1091-1112

Hyland, K. (2003) Dissertation acknowledgments: The anatomy of a Cinderella genre. *Written Communication.* 20 (3): 242-268.

Hyland, K. (2004) *Disciplinary Discourses: social interactions in academic writing*. Ann Arbor: MI. University of Michigan Press

Hyland, K. (2006). Representing readers in writing: student and expert practices. *Linguistics and Education.* 16: 363-377

Hyland, K. (2007). Applying a gloss: exemplifying and reformulating in academic discourse. *Applied Linguistics.* 28: 266-285

Hyland, K. (2008a).Genre and academic writing in the disciplines *Language Teaching.* 41 (4): 543-562.

Hyland, K. (2008b).As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes.* 27 (1): 4-21

Hyland, K. (2012) *Disciplinary Identities.* Cambridge: Cambridge University Press

Hyland, K. & Diani, G. (Eds). (2009). *Academic evaluation: review genres in university settings*. London: Palgrave-MacMillan.

Hyland, K. &Tse, P. (2005). *Evaluative that* constructions: signalling stance in research abstracts. *Functions of Language*. 12 (1): 39-64

Hyland, K. &Tse, P. (2007). Is there an 'academic Vocabulary'? TESOL Quarterly. 41 (2): 235-254

Hyland, K. & Tse, P. (2012). 'She has received many honours': Identity Construction in Article Bio Statements. *Journal of English for Academic Purposes.*11: 155–165

Loi, CK (2010) Research article introductions in Chinese and English: A comparative genre-based study *Journal of English for Academic Purposes* 9 (4): *267-279*

Matsuda, P. & Jeffery, J. (2012), Voice in Student Essays. In Hyland, K. & Sancho Guinda, C. (eds.) *Stance and voice in written academic genres*. London: Palgrave. pp 151-165.

Molino, A. (2010). Personal and impersonal authorial references: A contrastive study of English and Italian Linguistics research articles *Journal of English for Academic Purposes*. 9 (2): *86-101*

Moreno, A. & Suárez, L. (2008). A study of critical attitude across English and Spanish academic book reviews *Journal of English for Academic Purposes. 7 (1): 15-26*

Nesi, H. & Gardner, S. (2012). *Genres across the disciplines.* Cambridge: CUP.

Ozturk, I. (2007). The textual organization of research article introductions in applied linguistics: Variability within a single discipline, *English for Specific Purposes* **26** (1): 25–38

Samraj, Betty. (2005). An exploration of a genre set: Research article abstracts and introductions in two disciplines. *English for Specific Purposes*, 24, 141–156.

Scott, M. 2007. *Wordsmith Tools 5.* Oxford University Press.

Scott, M. & Tribble, C. (2006). *Textual patterns*. Amsterdam: John Benjamins

Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford: OUP.

Sinclair, J. (2004). *Trust the text.* London: Routladge.

Stubbs, M. (2005). 'Conrad in the computer: examples of quantitative stylistic methods. *Language and Literature*, 14:1. 14, No. 1, 5-24

Swales, J. (1996) Occluded genres in the academy: The case of the submission letter. In E. Ventola & A. Mauranen (Eds.) *Academic Writing: Intercultural and Textual Issues.* Amsterdam: Benjamins pp. 45-58.

Swales, J. (1998). *Other Floors, Other Voices: A Textography of a small University Building.* Mahwah, NJ: Lawrence Erlbaum.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work.* Amsterdam: John Benjamins.

Tse, P. & Hyland, K. (2008). 'Robot Kung fu': gender and the performance of a professional identity. *Journal of Pragmatics*. Vol 40 (7): 1232-1248.

Wulff, S., Römer, U. & Swales, J. (2012) Attended/unattended *this* in academic student writing: Quantitative and qualitative perspectives. *Corpus Linguistics and Linguistic Theory* 8: 129 – 157

Xiao, R., McEnery, T. & Qian, Y. (2006) Passive constructions in English and Chinese: A corpus-based contrastive study. *Languages in Contrast*, 6, 1: 109-149