Uses of corpus linguistics in education research: An adjustable lens

Alsop, S., King, V., Giaimo, G. & Xu, X.

Original citation:
'Uses of corpus linguistics in education research: An adjustable lens', in Theory and Method in Higher Education Research, ed. by Tight, Malcolm Huisman, Jeroen, pub 2020.
ISBN 978-1-80043-321-2

Publisher: Emerald

Draft Chapter: Dr Siân Alsop and **Dr Virginia C. King**, *Centre for Global Learning: Education and Attainment (GLEA), Coventry University, UK*; **Dr Genie Giaimo**, *Center for Teaching/Learning/Research and Writing and Rhetoric Program, Middlebury College, USA*; **Dr Xiaoyu Xu**, *Department of English, City University of Hong Kong, Hong Kong*

Uses of corpus linguistics in education research: An adjustable lens

Keywords: corpus linguistics, higher education, research, case studies, discourse

Abstract

In this chapter, we explore uses of corpus linguistics within higher education research. Corpus linguistic approaches enable examination of large bodies of language data based on computing power. These bodies of data, or *corpora*, facilitate investigation of the meaning of words in context. The semi-automated nature of such investigation helps researchers to identify and interpret language patterns that might otherwise be inaccessible through manual analysis. We illustrate potential uses of corpus linguistic approaches through four short case studies by higher education researchers, spanning educational contexts, disciplines and genres. These case studies are underpinned by discussion of the development of corpus linguistics as a field of investigation, including existing open corpora and corpus analysis tools. We give a flavour of how corpus linguistic techniques, in isolation or as part of a wider research approach, can be particularly helpful to higher education researchers who wish to investigate language data and its context.

1. INTRODUCTION

Corpus linguistics is an umbrella term for researching bodies of language data that occur naturally in the 'real world' (Bennett, 2010; Gries, 2009; McEnery & Hardie, 2012; Taylor, 2008). Corpus linguistic approaches to analysis employ computing power to examine these bodies of language data, combining quantitative techniques (such as comparing word frequency) and qualitative techniques (such as word-usage contexts) (McEnery and Hardie, 2012). These approaches are attracting growing attention within higher education research across disciplines, perhaps reflecting a global higher education landscape in which, running in parallel to the collection of big data, is an awareness of the importance of interpreting this data within a bigger picture; as MacNeil (2019) points out, when it comes to education, "while content is king, context is majesty". The content in this case is the wealth of authentic language data generated within higher education, including the papers and presentations that students, researchers and managers produce. The context comprises the various forms of related information that surround these written and spoken genres. Corpus linguistics enables researchers to harness relationships between this language content and educational context, which promotes trans-disciplinary and trans-context uptake of the approach, beyond its roots in linguistic disciplines and language teaching activities.

In this chapter, we examine different ways in which corpus linguistics can be used to make sense of higher education data, from the language of teaching to the language of research, administration and policy. Our aim is to demonstrate through examples how this approach may be applied by researchers across disciplines. The chapter begins with an introduction to the development of corpus linguistics and its fundamentals as an analytical approach. We then illustrate potential uses of the approach in higher education research through four short accounts by higher education researchers. These accounts give a sense of how corpus linguistics can be used in global higher education contexts, at various levels of focus. The first account by King discusses the use of corpus linguistic techniques for surveying and scoping corpora in order to better understand areas of higher education research interest. The next account by Giaimo explores how corpus linguistic techniques can be used to examine working practices within an Academic Writing Centre. The final two accounts by Alsop and Xu more narrowly explore how corpus linguistic techniques can be employed to interrogate custom-built corpora to understand higher education communication practices; Xu focuses on argumentation within published research articles and Alsop focuses on discourse functions within spoken lecture data and in written feedback. Following these accounts, we consider some of the challenges that using corpus linguistic approaches may pose. We conclude by arguing that the ever-increasing amounts of language data to which higher education researchers have access can be usefully interpreted if examined through a corpus linguistic lens.

2. Corpus Linguistics as a Field of Investigation

Corpus linguists study bodies of language data, or *corpora* (singular: *corpus*), that are made up of purposefully collected, naturally occurring (authentic) text (Sinclair 1991). The semi-automated nature of the analysis of corpora makes it possible to interrogate large amounts of data to reveal patterns of linguistic usage that would be difficult, or impossible, to identify manually in a reasonable timeframe. Corpus linguistic investigation enables researchers to answer questions about the presence of language patterns in data (Bennett 2010). The approach to corpus analysis has been characterised as either corpus-based (inductive) or corpus-driven (deductive), but the usefulness of this somewhat abstract distinction in practice can be questioned (e.g. McEnery & Hardie, 2012; Meyer, 2014). Quantitative corpus analysis can certainly provide results that complement those found through detailed, 'close' textual analysis of a corpus ; and close analysis can likewise inform more widescale, automated investigation.

2.1. Development of the field

The nature of corpus linguistic investigation has developed over time, but the underlying drive to understand patterns in a variety of authentic language data remains constant. Work to create pre-digital databases and manual concordancing—which involved highly labour intensive manual data collection, preparation and analysis—resulted in, for example, biblical concordances, to document grammars and to inform dictionary construction, as well as a pre-computational corpus of English language: Quirk's original (1959) *Survey of English Usage* (see Meyer 2008, p. 1-14). Shifts in the use of computer-based corpora, due to technological advances and increased access in the 1960s, enabled automation of much of the data preparation and analysis processes. Seminal to this period was the creation of the English-medium Brown Corpus of Standard American English and the Lancaster-Oslo/Bergen (LOB) corpus of British English which are both available online through a variety of interfaces. Composed of text from 15 categories ranging from newspapers to fiction, these 1-million word corpora provided easy access, for the first time, to a large amount of machine-readable text (McEnery & Hardie, 2012). In their wake, Johansson (2008, pp. 33-34) depicts the 1970-80s as a "breakthrough" period in which an "explosion" in variety and usage of corpora occurred internationally. Fuelled by the pairing of ongoing technological advances and human curiosity about the nature of language, the modern structure of corpus linguistics marks a continuation of such momentum. Today's corpora reflect such advances through variation in focus, form, and size.

2.2. Questions of Design

Options for corpus design include choices about size, representativeness, time span, and language plurality . Interested users from a range of backgrounds now have access to 'super-corpora' (mostly web-based and in English), such as the Hansard Corpus (n.d.) of British Parliament speech events 1803-2005 (1.6 billion words), alongside much smaller, tailored corpora. Some collections attempt to offer a balanced representation of particular languages or language varieties at particular times, as used in particular genres. The newest, 2014 version of the British National Corpus (BNC 2018), for example, promises to offer 100 million words of contemporary written and spoken British English. The various editions of this corpus (1994, 2001, 2007; BNC 2015) give snapshots of general language usage. The continuously expanding Corpus of Contemporary American English (COCA, n.d.), currently 600 million words, allows researchers to monitor change in language usage over time (1990-2019). Depending on design, corpora can be used as comparators for researchers' own datasets, and also enable synchronic and diachronic investigation. At the same time, corpus compilers continue to build smaller, specialised corpora to represent particular languages or language varieties. In UK higher education, for example, proficient student writing is collected in the British Academic Written English (BAWE, n.d.) corpus ; spoken data is likewise collected in the British Academic Spoken English (BASE, n.d.). Other specialist corpora include texts written in multiple languages, and 'parallel' corpora contain texts translated into one or more other languages. These corpora offer multiple opportunities and models for researchers to pursue.

2.3. Analysis Techniques

A common starting point of corpus-driven investigation is to create a list of all unique words in a text/s (a 'wordlist') and their frequency of occurrence. From here, researchers can view words or groups of words in which they are interested in multiple ways, with or without the use of statistical tests. Concordance lines, for example, present all examples of a chosen word as a central node with surrounding co-text, the window of words occurring either side of the node, giving a keyword in context (KWIC) view (see the example 'corpus' in Table 1). Automatic retrieval of 'n-grams', or multi-word units, allows continuous sequences of words to be investigated. The retrieval of such strings of words does not depend on tests of statistical significance. It does, however, enable the user to see contextualised patterns in large amounts of language data quickly, which is important because "[t]he language looks rather different when you look at a lot at once" (Sinclair 1991, p. 100). This idea of stepping back from texts to see the bigger picture enables a viewpoint beyond individual researcher intuition.

Within a corpus, or any chosen data subset (subcorpus), the relationship between a node word and those that occur in close proximity can also be understood through 'collocation', which gives a measure of the strength of association, or 'glue', between words (e.g. through a mutual information test) to show whether they co-occur more or less frequently than expected. This measure can be revealing of how certain words are used or what they mean, in context. For example, in our sample of corpus linguistics higher education research (see Table 1) the word 'corpus' collocated strongly with the words 'linguistics', 'analysis', and 'collaborative', which gives a sense of how researchers use the word.

Patterns occurring more frequently than expected in language can also be calculated by comparing a single corpus (or any subcorpus) against a reference corpus (generally a much larger corpus). Doing so allows the calculation of 'key words' and 'key terms' that are unexpectedly frequent in the target corpus when compared to the reference corpus. These fundamental techniques of corpus analysis can be deployed in various configurations, alone or in combination with other analysis techniques.

in a design class: A	corpus	analysis of student writing 2019
on Snapchat and Instagram. The	corpus	analysis of how stories as
characteristics. With reference to a	corpus	-based Critical Discourse Analysis (CDA
target forms between a learner	corpus	and a similar-sized native
then compared to a comparable	corpus	of essays by native speakers
of a spoken and written	corpus	with a qualitative analysis of

we introduced a concept from	corpus	linguistics-a "lexical bundle," which
is argued that Internet-based	corpus	tools and techniques are undervalued
engineers. The innovative approach uses "	corpus	linguistics" - computer-assisted techniques from
text) and the employment of	corpus	analysis software can help to

Table 1: An example of *keyword in context* (KWIC) from a selection of corpus-related articles published in higher education journals, created using the Voyant Tools, Context function (Sinclair & Rockwell, 2020)

At this point, it is worth differentiating techniques such as collocation and keyword analysis from 'topic modelling', which enables the identification and summarisation of topics in large amounts of text. Although all the techniques mentioned so far can be used to establish what a body of text is 'about', topic modelling arises from a different research tradition from corpus linguistic techniques. Corpus linguistics is a sub-field of linguistics which has capitalised on computer automation to scale up manual text analysis techniques used to understand patterns in language, often including syntax and structure (McEnery & Hardie, 2012). Topic modelling is a type of text mining which ignores the structure and syntax of a 'bag of words' to reveal underlying semantic themes (Perez-Encinas & Rodriguez-Pomeda, 2019). Where collocation measures in corpus linguistics typically calculate the co-occurrence of words within a restricted span of less than a handful of words, topic modelling looks at co-occurrence typically across whole texts, where the separation may be thousands of words (Brookes & McEnery, 2019, p. 4). This broader way of understanding relationships between words is an increasingly attractive addition to the toolkit of many corpus linguists (e.g. Murakami, Thompson, Hunston, & Vajn, 2017) , and some corpus tools incorporate functionality to perform such analysis, such as the open-source Voyant platform (Sinclair & Rockwell, 2020).

2.4. Data Preparation

The many ways of 'seeing' data through corpus linguistic and other related lenses are connected to decisions about the way in which the corpus data is prepared. In order to compare language patterns across multiple files within a corpus, it is necessary to include not only the text that is the object of analysis, but also information about that text, as *metadata*. For this reason each file in a corpus generally comprises two parts: body text and corresponding metadata about that body text. In the case of the BAWE corpus, for example, users can access c.2,700 files that include examples of student writing (body text) *and* metadata information about the student writer (e.g. discipline and level of study) and about the text (e.g. genre and length). The inclusion, or association, of such metadata makes texts filterable, allowing comparisons of language patterns to be made within the corpus and across corpora.

The body text can be prepared in a number of ways. Basic information about its structure is commonly added in machine-readable format, such as where a new sentence or paragraph occurs in writing, or a new utterance and its orginator in transcribed speech. Information about language structure is also commonly added, such as parts-of-speech (POS) tagging, through which nouns, verbs, adjectives etc. are automatically identified with a high degree of accuracy. Adding such levels of descriptive tagging allows language patterns at, for example, the level of grammar rather than words only, to be investigated. More subjective descriptions of the body text are also commonly added, for example where a researcher is interested in identifying instances of certain indicators of attitude, or certain types of humour. In such cases, the term 'annotation' can be used to distinguish between interpreting linguistic information in a subjective way and identifying objectively the structural elements of a text (or, 'markup'), as defined by Gries (2009). While both approaches to preparing linguistic data are valid, they yield different findings that the researcher may be more or less interested in pursuing.

2.5. Software and training

The success of corpus linguistics as a research field means that higher education researchers can choose to use pre-existing corpora or to compile their own, with relative ease. There is a vast array of tools for compiling and/or analysing corpora, which continues to grow. These range from paid and unpaid options that are web-based or downloadable, some of which allow users to input their own data and/or access pre-loaded corpora. At the time of writing, the Linguist (n.d.) site presented 61 different items of "Software related to Text/Corpus Linguistics". Questions over where to start – what data to use and how to use it – may seem overwhelming. There is, however, comprehensive open-access training provision available to support users, from software-specific 'how-to' guides (e.g. for AntLab's (2019) suite of software or for the UAM CorpusTool (n.d.)) to MOOCs such as that run by the University of Lancaster (available via Future Learn (2020)) to accredited courses at undergraduate and postgraduate levels . Much of the available software is geared towards enabling new users of corpus linguistics to get started through user-friendly interfaces and extensive glosses, whilst also providing facility for more complex work.

3. Researcher Accounts

The following accounts illustrate a range of available resources and entry points, and touch on this plurality by examining examples of specific uses of corpus linguistics in higher education research. These case studies capture some of the diversity of researcher positioning yet are characterised by a shared interest in making sense of higher education data to improve staff and student experience. Each of the authors in turn presents their personal use of corpus analysis techniques, which we subsequently jointly critique.

3.1. Scoping the Landscape (King)

Although commonly used as a tool to carry out new research, corpus linguistics can provide valuable insight into the nature of existing research. As in every discipline, researchers in the field of higher education strive to manage the volume of academic publications pertinent to their spheres of interest. Over recent years, I have developed uses of corpus linguistics that help to analyse particular higher education research foci. This approach could be classified as a type of 'distant reading' (Moretti, 2013) – that is, a way of gleaning the essence of a large body of text through automated means, rather than having recourse to the potentially impossible task of reading it in the traditional way.

Most successfully, I use the search tools provided by databases such as Scopus (n.d.) and Academic Search Complete (EBSCO 2020) to extract the abstracts and other key data of potentially interesting articles, download these, and then submit the downloaded data to a corpus linguistics platform. In this way, I can undertake more subtle analysis, no matter how many abstracts are selected. For example, if the abstracts are extracted in date order, the corpus I create can reveal trends in the use of particular words, phrases, or part-words, over time. Similarly, I can search for the collocates of the terms that interest me, creating new combinations of words to use subsequently in academic database searches. I can also pinpoint particular combinations of key terms within the extracted abstracts, titles and key words, and examine their contexts for disciplinary, methodological or theoretical themes. This simple use of corpus linguistics has proved useful in gaining an overview of new research areas, or in looking for developments in familiar ones. A similar approach is demonstrated in Taylor (2008) which uses corpus linguistics methods to explore dispersion of the term 'corpus linguistics' (and its synonyms) in research literature, in order to arrive at a definition of the term.

Clouder and King (2015) provides an example of surveying and scoping existing corpora in order to better understand the higher education research landscape. This article explores the use of Appreciative Inquiry (AI) in researching higher education. Scopus and Google Scholar were used to identify academic publications which concerned AI. Of over 2,000 items found, 13% (289) had been published in higher education -related outlets. Further investigation was undertaken in Voyant, creating one corpus for AI items published in all disciplines, and another restricted to higher education -AI items. The 'all-disciplines' corpus featured variants of 'appreciative inquiry' as the most frequent words. However, since the corpus was organised in date order, Voyant's Trends tool revealed a downward trend in usage of 'appreciative inquiry' over the period 1999-2014 alongside an upward trend in the use of the term 'research'. A suggested interpretation of these features was that the AI terms were repeatedly used at first while appreciative inquiry was being established, but that its increasing acceptance as a research approach reduced the need for such repetition. However, no increase in use of 'research' was mirrored in the higher education -AI corpus. Examination of the higher education- AI concordance confirmed suspicions that AI was used by researchers in this field to inform their work rather than as a focus for research. Few examples were found in the higher education corpus of full adoption of the AI approach.

I have also investigated comparative word-frequency in a body of text that comprised a month's-worth of my incoming emails as part of an autoethnographic study of my own academic practice (King, 2013). I used Voyant to interrogate my small (18,000 word), anonymised, synchronic corpus. This software platform provided independent insight into the shared discourse underlying the emails I had received from colleagues, students, managers, administrators, professional bodies and chance correspondents. This highly personal use of corpus linguistics enabled a systematic analysis of the themes and patterns occurring within my chosen body of text. I could have consciously influenced the content of a corpus constructed from my outgoing emails, presenting a particular persona, focusing on specific themes, reusing telling vocabulary. However, my incoming emails were the product of longstanding behaviours (my own, and those of my network of correspondents). The patterns of language use in this corpus might have been apparent to an outsider, but were difficult for me to discern. Using software to identify and contextualise the most frequent words that characterised these accumulated emails provided a means of reflecting on the elements of my academic identity which were bound up with my professional networks. I used the results of my analysis to create a table which visualised the most frequently used vocabulary under thematic clusters, almost as an artwork. This novel research strategy was helpful in examining a familiar higher education context where it would otherwise be difficult to distance oneself. I would recommend it as a useful way of exposing the semi-hidden data collected in this kind of 'close-up' higher education research (Trowler, 2012).

A final example of my use of corpus linguistics as a non-specialist is given in Billot and King (2017) which sets out to explore the effectiveness of academic staff induction. Creating a corpus of relevant abstracts proved problematic since the term 'induction' has multiple meanings, and there is a range of alternative terms for introducing a new member of staff to an organisation. As Perez-Encinas and Rodriguez-Pomeda (2019) found in their study of probabilistic topic modelling, it was necessary to develop the Scopus search algorithm heuristically until the initial 700,000 items was refined to 1,535 that concerned higher education. However, once downloaded into Voyant, this corpus proved largely uninteresting. Progress was made once a comparator corpus was created comprising abstracts on staff induction from the field of Human Resources (HR). It transpired that the HR corpus was characterised by vocabulary associated with the socialisation of new employees, and the setting of objectives and targets. Close examination showed that this corpus presented induction as a reciprocal process aimed at enhancing the effectiveness of new staff. None of this was evidenced in the higher education corpus, which focused on training. The concordance demonstrated that research in the higher education context presented induction as a series of hurdles for new academics to jump. The use of words related to 'community' in the higher education corpus appeared in abstracts calling for greater organisational socialisation of new academics.

These brief examples of ad-hoc corpus-building and interrogation show how researchers can use corpus linguistic techniques to varying extents and ends, even with fairly small datasets.

3.2. Corpus Linguistic Analysis as Writing Centre Research and Assessment Tool (Giaimo)

In a period of ever-expanding online data collection, corpus linguistic techniques are a powerful method of analysis. As a writing centre Director in the United States, I collect a lot of linguistic data from both peer tutors as well as student writers. These data include full essays, transcripts of online synchronous tutoring sessions, and session notes. Depending on the area of focus, I can assess the kinds of writing being brought to the writing centre (artifact collection), the kinds of interactions that occur between tutors and writers within their sessions (transcripts), or how tutors conceive of and reflect upon their tutoring practice (session notes).

Up to this point, I have applied corpus linguistic analysis to the examination of session notes. Session notes are the summative notes that tutors fill out at the conclusion of their sessions with writers. These notes can focus on several different elements of the session from those that evaluate the writer's engagement and learning to those that ask the tutor to reflect upon their pedagogical approach, in-session. The notes' questions vary from centre to centre, and are largely idiosyncratic, though most writing centres engage in some version of summative note taking practice.

I have published on corpus linguistic analysis of session notes (Giaimo et al., 2018) in cross-institutional contexts. In that article, I noted that, at the time, Ohio State University's Writing Center conducted over 10,000 appointments annually. Tethered to each appointment was a 75 – 150-word session note. The analysis was limited to analysing 7,000 notes collected over a year-long period, which amounted to 550,000 words. I reviewed responses to 2 of the 6 questions on the session note form. My co-authors and colleagues also limited their datasets in similar ways (by word count, and time period). Together, we produced a 2-million-word corpus comprising 44,000 session notes. From it, we found that our writing centres had particular tutor cultures, training cultures, and pedagogical cultures. And, using the contexts tool in Voyant, which provides an analysis of collocates connected to keywords, I found out that tutors used session notes to conceptualise the labour of tutoring, including grappling with emotional fallout from stressful tutoring sessions and attendant feelings of guilt related to being unable to "help" writers sufficiently (p. 245). In the article, I provided and analysed a number of collocates surrounding the keyword "help" to demonstrate how tutors evoke emotional labour through several frequently-utilised keywords (7 in a list of 20 most frequent keywords) in their session notes (p. 250).

Corpus linguistic analysis gave me a sense of what was going on in my 60-person writing centre. For resource- and time-strapped administrators, this is a powerful assessment method that helps us to understand, broadly-speaking, what is going on in the minds of our tutors, as well as what is going on in their sessions with writers. From this project, we designed a study that used metadata (tutor academic rank, semesters of experience, semester of work, etc.) and tracked tutor development and growth, over time, through analysing the content (and potential changes) in tutors' session notes (Giaimo & Turner, 2019). We used corpus analysis to generate key terms and, from there, developed a 12-point rubric that we used to code 1,261 notes. Some of the variables that we coded the notes for included affective, semantic, and evaluative stances. A mixed linear model (three-way ANOVA) was used to track variation and a logistic principle component analysis (Logistic PCA) was used to model dimensional clustering and change within and between tutor cohorts. From this study, we found that undergraduate and graduate inexperienced tutors move through the writing centre in different ways, likely due to differences in hiring and training, but, eventually, converge on a similar set of practices as they become more experienced and enculturate to the writing centre through training and mentorship (p. 154). Though we used a number of qualitative and quantitative models, the development of the rubric was informed by corpus linguistic analysis, specifically findings from keyword frequency identified through the terms and summary tools in Voyant.

A third project (Giaimo & Turner, under review) was inspired by findings from the session note study (2018) that tutors report differentiated tutoring approaches that break down along talk-based and nontalk-based tutoring strategies (p. 247). We focused on tracing the activities that tutors reported using in their writing centre sessions, again, through discourse analysis. Coding questions focused on activity-use in session note reportage; we identified four different "types" of tutors who utilise specific kinds of activities such as grammar-focused ones, or discussion-based ones. This project attempts to answer a long-standing question about how tutors can engage with their work in "flexible" ways. In our study, we interrogated how tutors behave flexibly , if at all, and, if they do, what that flexibility looks likes. From our coding of session notes, we found that tutors use a wide range of activities in their sessions but that these activities tend to be thematically similar to one another (i.e. grammar-based, or discussion-based). Therefore, flexibility might be at the group level rather than the individual tutor level, which has implications for how we hire and train writing centre staff.

The two discourse analysis projects I mention here originated from the corpus analysis project (Giaimo et al., 2018). Without its findings that tutors struggle with emotional labour in their sessions and seem to rely upon a repertoire of similar strategies, these later studies on tutor growth and tutor practice would not be possible. They inform each other.

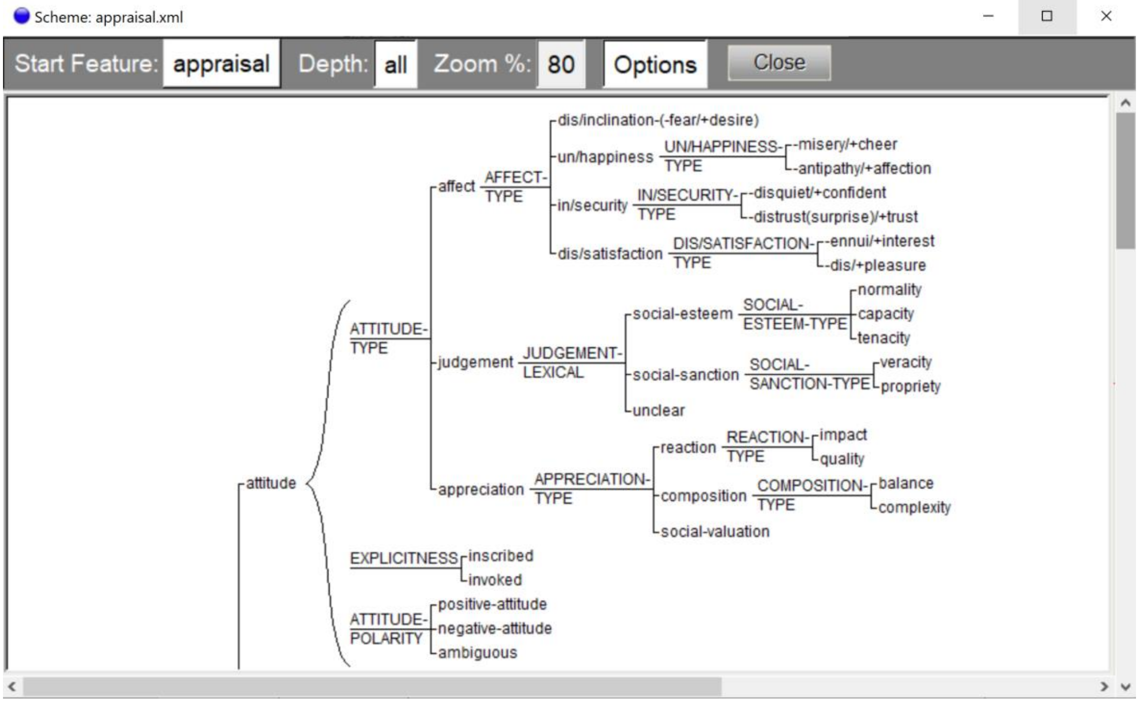
Higher education research, particularly at the programmatic or institutional levels, is made more comprehensive by incorporating corpus linguistic analysis. For writing centres, it offers a powerful lens into the inner workings of what are often high performing, if also under-supported, spaces within higher education . And, as we move towards the collection of ever-larger sets of data, this method offers an effective and speedy way into parsing some millions of units of information. For me, co-situating corpus analysis with discourse analysis, much like a mixed methods survey, gives depth to breadth of findings. It is an excellent starting point for writing centre research and analysis.

3.3. Written Research Communication (Xu)

Statistically speaking, findings made from big samples can be extended to a wider population with a high degree of certainty. In the field of corpus linguistics, large corpora are compiled particularly to pursue the transparency, reliability, and replicability of language analysis. However, applied linguists and EAP (English for Academic Purposes) practitioners often come across situations not ideal for compiling big corpora. In the context of higher education , for example, we may a have a limited number of participants (e.g. an analysis by an EAP tutor aimed at her/his own class of students), limited existing data (e.g. English publications written by academics from a particular, small developing country), difficulties in accessing the data (e.g. PhD viva transcripts), or a need for extensive manual annotation (e.g. analysis of evaluative meaning). It is therefore important to explore the ways in which small corpora can benefit from quantitative analysis using corpus tools.

I worked on a project that encountered a few of the situations mentioned above. The project (Xu & Nesi, 2019a; 2019b) investigated stance markers (e.g., words that express attitudes, certainties or evaluations) in applied linguistic research articles across Chinese and British cultures, using the Appraisal framework (Martin & White, 2005). The framework includes, for example, Inclination markers such as 'hope' and 'wonder', and Endorsing markers such as 'shows' and 'demonstrated'. In order to focus on cultural differences, we avoided controlled variables such as overseas educational background and low English language proficiency of the authors for the Chinese sub-corpus; we only collected publications in credible international journals, and we only collected publications written solely by so-called 'home-grown' Chinese academics (who received their doctorates in China, and currently work in China). Stance markers and patterns cannot be classified into hard-and-fast, mutually exclusive categories without close examination of the contexts in which they occur, and hence extensive manual annotation was involved. For these reasons, a small corpus of only 30 research articles was examined.

Although the analysis requires manual annotation rather than using Corpus Query Language (CQL) or regular expressions (to automatically identify grammatical or lexical patterns in tagged corpora), it was impossible to reach conclusive findings without using corpus tools to quantify the annotated features and to conduct statistical analysis. UAM CorpusTool (O'Donnell, 2011) was the main tool we used for our project. The 30 research articles in plain text form were first imported into the tool. The manual annotation was recorded in two steps: 1) *Edit annotating schemes* (see Figure 1). 2) *Annotate each text* (see Figure 2). In this way, our 66 categories of annotated Appraisal features were sorted and stored for statistical analysis.



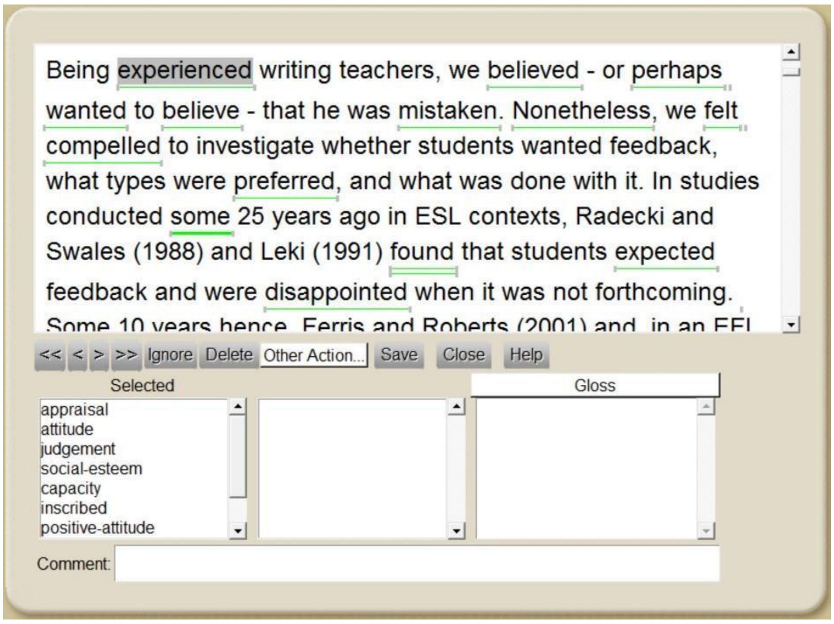
Scheme: appraisal.xml

Start Feature: appraisal

Depth: all

Zoom %: 80

Figure 1: Scheme editing



Being experienced writing teachers, we believed - or perhaps wanted to believe - that he was mistaken. Nonetheless, we felt compelled to investigate whether students wanted feedback, what types were preferred, and what was done with it. In studies conducted some 25 years ago in ESL contexts, Radecki and Swales (1988) and Leki (1991) found that students expected feedback and were disappointed when it was not forthcoming. Some 10 years hence, Ferris and Roberts (2001) and, in an EFL

Selected

appraisal
attitude
judgement
social-esteem
capacity
inscribed
positive-attitude

Comment:

Figure 2 Annotation window

The Statistics, Explore, and Search functions in UAM allowed for various statistical comparisons and keyword analyses for our project. For example, one of our research questions was 'how do Chinese academics construct their stance in applied linguistics research articles?' The following steps were taken to address this question:

1) Compare the use of the 66 categories of Appraisal markers across the Chinese and British sub-corpora .
2) Compare keywords in each of the 66 Appraisal categories across the two sub-corpora .
3) Retrieve instances of a particular Appraisal category.

.

It would have taken months just to manually count the frequencies and calculate the statistics. We were instead able to quickly discover the statistically significant features using UAM CorpusTool. For example, we found that the British authors used more Inclination markers (e.g., it is our *hope* that, researchers may *wonder*, we were *curious*). This indicates that the British authors tend to express their passion for research in their writing. On the other hand, the Chinese authors used less engaging language than the British authors; in particular, the Chinese authors used fewer Entertain markers (e.g., *may, can*) and used more Endorsing markers (e.g., his study *shows* that, she *found* that, previous studies *demonstrated* that), indicating that they regarded research writing as being less dialogic and more descriptive of facts that are regarded as correct, valid, and warrantable. Such findings can inform the teaching of English academic writing to Chinese academics in applied linguistics.

The UAM CorpusTool made the quantification possible for our analysis and allowed us to follow the common practice observed by Schmied (1993) in corpus linguistics whereby the qualitative research stage is often a precursor for quantitative analysis, and a quantitative result may need to be qualitatively interpreted. In other words, the corpus tool helped improve the transparency, reliability and replicability of our findings as well as the efficiency of our work with this small corpus.

3.4. Spoken and Written Classroom Communication (Alsop)

In my role as researcher in education at a UK university, I use the linguistic data that lecturers and students produce to study communication across higher education contexts, particularly spoken lectures and written feedback. I build custom corpora to investigate the functions of these texts beyond the level of individual words.

In terms of spoken data, I constructed with colleagues a corpus of lecture data, the Engineering Lecture Corpus (ELC), which to date includes 76 English-medium lectures from three countries (the UK, Malaysia and New Zealand), totalling just over half a million words of transcribed text (Alsop 2016). The aim of this ongoing project is to understand lecturer discourse outside discipline-specific content (or technical jargon), how such discourse can be characterised, and how it differs across educational contexts. We created a taxonomy of *pragmatic* discourse functions including: storytelling, summarising, and humour, most of which contain multiple sub-functions, or types. Summary, for example, includes four types: reviews of previous and current content, and previews of current and future content.

Refining this taxonomy involved repeatedly cycling through the relatively large spoken corpus, manually annotating in the transcriptions the start and end of each function. We started with a working list of functions based on pilot investigation and revised this significantly in terms of hierarchy and content during each annotation cycle. We also undertook multiple rounds of inter-annotator reliability testing, which fed back into how we shaped the taxonomy. I chose to work with a simple text editor to prepare my data and used scripts to extract information about language patterns, but I could have used a variety of platforms to do similar work (e.g. UAM CorpusTool or Voyant).

By pragmatically annotating our corpus, we compiled information that allowed us to answer questions about the linguistic character and usage of the lecture functions using corpus techniques. For example, in addition to exploring lexicogrammatical structures, we counted occurrences and words per occurrence: *summary*, for example, accounted for 11% of all lecture discourse. We used keyword analysis to identify which words are most salient to which functions, and probed the context of their usage through concordances, collocations and n-grams. The annotation also allowed us to compare similarities and differences based on metadata variables, particularly country of delivery, which showed, for example, that although all functions were common to all lectures, certain types of *humour* are used more heavily in some contexts, as are certain types of *story*. As King notes, this approach revealed patterns that we would have been unlikely to notice based on intuition. As in Xu's work, such findings can inform practice both in a disciplinary sense and across cultural/educational contexts.

To better understand the patterns we found, we developed a visualisation tool, *ElVis*, which renders each lecture as a stacked line on the Y axis, from start to end point on the X axis. ElVis shows any combination of functions and subtypes, in line with Sinclair's view on the value of looking at a lot of data at once. Figure 7 shows all 76 lectures and all occurrences of only the function *summary*.

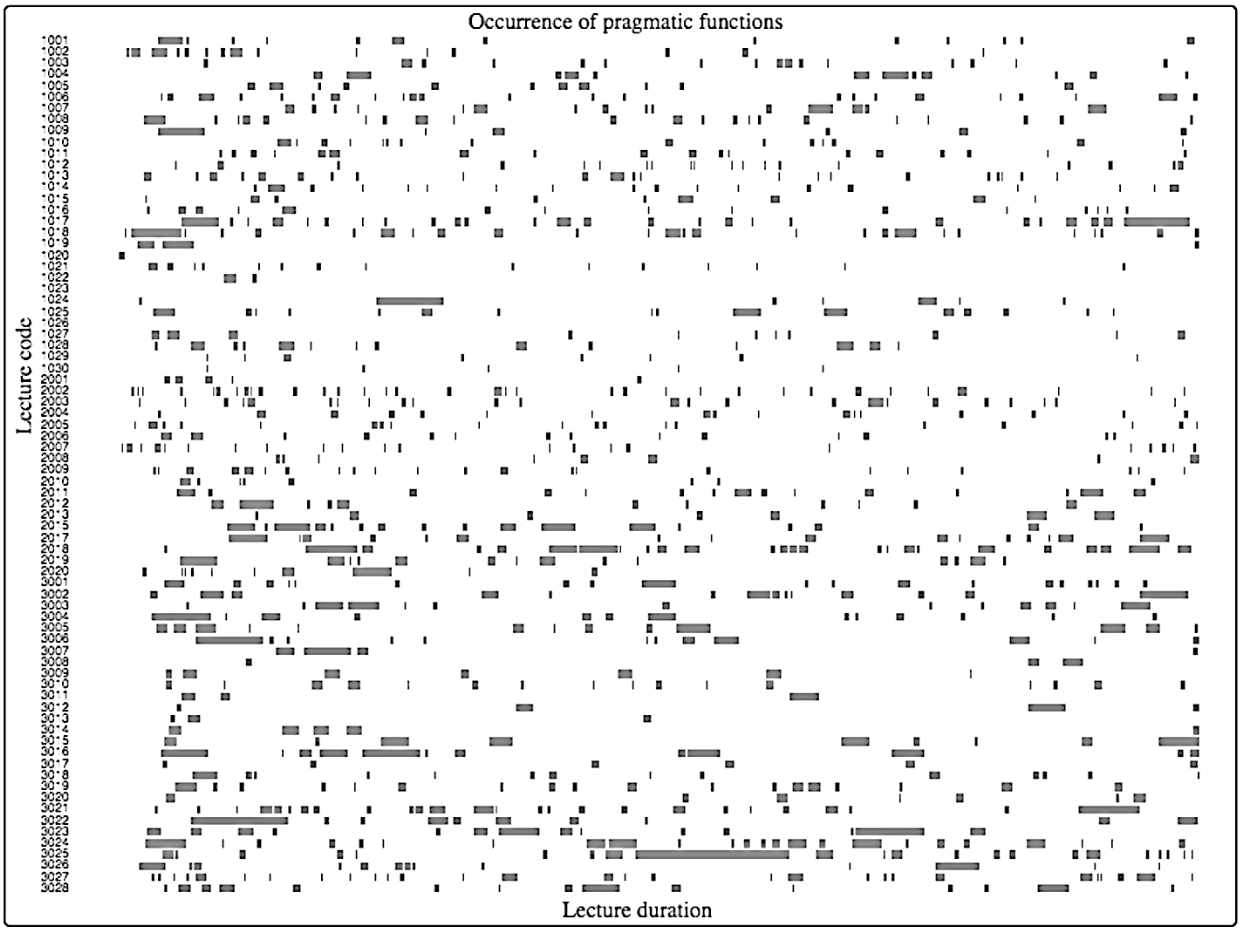


Figure 7: Visualisation of ELC lectures and the Summary function

Visualising the ELC annotation let us *see* at various levels of granularity where and for how long discourse functions occurred and how they *chained* in terms of function and sub-function types, augmenting and at some points directing finer language analysis. In the case of summary, for example, visualising where in the lectures different types of *reviews* and *previews* occur showed, that summary does not clearly follow the type of *beginning, middle* and *end* model promoted in some pedagogical advice, and that previews and reviews employ quite different language strategies. In the same way as Giaimo discusses the benefits of 'co-situating' corpus techniques with discourse analysis, we found the process of moving between the annotated data and its visualised form extremely useful for understanding, and also for disseminating, the communication strategies we identified in the data.

In terms of understanding communication strategies in written texts, I am currently developing with colleagues a corpus of the summative written feedback that lecturers give to assessed student work. Although this project is in the early stages, we have developed a working taxonomy to describe functions within written feedback. Our taxonomy builds on previous systems by offering a more systematic and symmetrical approach to categorisation, whereby five main categories (*advice*, *critique*, *observation*, *praise* and *query*) branch three times according to *aspect*, *orientation*, and *focus*. This fairly complex branching structure comprises 60 functional annotation categories and an unclassified category, which allowed us to annotate with high reliability all text in two small pilot corpora (Alsop and Gardner 2019; Alsop, Gardner and Priestley 2020).

In this project, the metadata variables encoded within each corpus file are central to our research questions. We are interested in whether relationships exist between the language of the written feedback that students receive and the sociodemographic and academic backgrounds of those students. We are particularly interested in the relationship between written feedback and ethnic attainment disparity in UK higher education. We have therefore included detailed information about, for example, students' prior educational experience, discipline, current grade, gender, and self-identified ethnicity. We have to date begun to identify correlations between language function and discipline (Alsop and Gardner 2019) and also language function, self-identified ethnicity and grade (Alsop, Gardner and Priestley 2020). Gauging the impact of this on higher education experience and outcomes will be done through triangulation with other qualitative methods, including focus groups with staff and with students. The process of mapping functional language patterns within the written feedback against recipient metadata will be undertaken on a wider scale as part of our ongoing project (involving a multi-million-word corpus). As with work on the ELC, movement between qualitative identification of language-level functions and quantification of language and usage patterns is important to this way of seeing texts, in context; my work aligns with the view that the two approaches are not only complementary but are intertwined (Schmied 1993, p. 85). Employing corpus linguistic approaches in this way allowed me to ask and answer questions about language delivery choices (*what*, *when*, *how*, and *to whom*) across large amounts of authentic data. The outcomes inform lecturer and student training, as well as further research understanding, in the context of facilitating equity of opportunity in local and global higher education settings.

4. Challenges to Corpus Analysis

Despite their diversity, our accounts of using corpus analysis are largely positive. However, there are issues which may make its use challenging to higher education researchers. These include access issues related to specific corpora and consideration of data triangulation, technical inconsistencies between software platforms, questions of data security and ethics, and, more philosophically, the role corpus analysis might play in higher education research.

Not all corpora are designed, or accessed, equally. As corpora are digitised, user access is determined by the levels of access granted by compilers. For example, users can consult some corpora through open-access interfaces or download them directly; BAWE and BASE are available through SketchEngine (2020) software and can be requested in full from the Oxford Text Archive (2019) for research purposes. Useful lists of open access corpora exist (e.g. English-Corpora.org (n.d.) and Clarin (2020)), but not all corpora are made publicly available. Corpus metadata are also not always consistent, easily accessed or shared, depending on country and context in which the researcher is doing their work. A further complication for researchers seeking to use different corpora and corpus linguistics tools is a lack of interoperability due to historically different approaches to the encoding and structure of the underlying databases. This is being addressed through ISO standards and the work of such organisations as CLARIN (2019) and TEI (2019) but remains a challenge (Wigham & Poudat, 2020).

Existing, accessible corpora should comprise data for which ethical approval has already been granted or is not necessary, but researchers must be mindful of ethical and security considerations relating to data collection and use when compiling corpora and when using existing corpora. Such care is particularly pertinent when researchers are working with corpora that contain sensitive or restricted data, which affects choices such as whether to use web-based or downloadable analysis software. For example, uploading full data and metadata to a web-based platform may breach project data security protocols. Due to its reliance on computing power, corpus linguistic investigation, as much if not more so than other research approaches, must be grounded in sound understanding of data management and ethical procedures.

Methodologically speaking, corpus analysis software systems do not all utilise the same statistical measurements or include the same linguistic nuance. For example, Voyant can be set to ignore 'stop words' (such as: an, on, and, me), while the AntConc system (2019) automatically includes them. Voyant provides frequencies for keywords, but does not adjust its analysis based on corpus size. Similarly, without a reference corpus, researchers might struggle with identifying what, in the analysis, is expected/unexpected. Given all of these considerations, it is important for researchers to have a general sense of what is going on "under the hood" of the corpus analysis programs, as they might not include enough or the appropriate statistical tests to produce accurate outcomes. For example, in comparing two different responses, with two different word lengths, from session notes, Giaimo had to calculate the relative frequency because the raw frequency gave misleading information (because of the differences in corpora size) on term keyness. Researchers, therefore, need to have working statistical knowledge to shore-up and confirm, or disconfirm, the findings from corpus analysis software.

Part of the philosophy behind engaging with corpus linguistic analysis is to go beyond researcher intuition, as shown in Giaimo's work on understanding the emotional labour of a large number of writing tutors and in King's investigation of her own patterns of correspondence. Yet subjectivity can unquestionably play a role in data preparation and analysis choices, as the accounts by Xu and Alsop demonstrate. For example, in Alsop's exploration of written feedback, corpus analysis can reveal the presence of patterns in functional language such as praising or instructing, but understanding *why* this language is used (lecturer intention) and *how* it is received (student reception) requires further research methods to be employed, such as focus groups. We suggest that corpus linguistic investigation is one that lends itself well to triangulation with other approaches, and that, to answer some research questions, triangulation is not only beneficial but necessary.

It is a continuing matter of debate as to whether corpus linguistics is a methodology, collection of methods, or theory. Bennett (2010, p. 7) observes that "most corpus linguists are not willing to answer that question in such terms". Some have. Leech (1992: 106), for example, described early work as constitutive of “a new philosophical approach” and Teubert (2005, p. 2) later referred to a “theoretical approach”. Other corpus linguists describe a methodology composed of a system or set of methods/principles/procedures (e.g. McEnery, Xiao, & Tono 2006; Meyer 2008). Citing this divide, and aligning with the latter position, Gries (2009, p. 1225) points out that few practical differences are likely to arise from the perspective adopted because much corpus linguistic work is descriptive or applied, rather than directly reliant on theory. Equally, whether or not differences arise in practice is unlikely to subdue ongoing discussions over *what corpus linguistics is*. If it is viewed as a set of methods, or approach, it can certainly be characterised by plurality. As demonstrated by Taylor’s (2008) observations on the construal of corpus linguistics within its own literature and Lüdeling and Kytö’s (2008) detailed overview of the field, it is variation in compilation, type, interrogation and usage that characterises corpus linguistic work.

5. Concluding Remarks

Through these diverse accounts we aimed to highlight some of the ways in which corpus linguistic approaches can be used in higher education research. We demonstrated how these approaches can provide an adjustable lens through which to view large bodies of natural language data in use, which can augment the development of and response to research questions, inform higher education practice and procedures, and, potentially, policy, and also complement other forms of data analysis. We showed how the questions that we ask through corpus linguistic techniques can by design be local or global. Xu and Alsop explored how the process of making sense of language data often requires extensive qualitative analysis of context prior to or in conjunction with quantitative work. Their accounts demonstrated how qualitative and quantitative approaches can be not only mututally beneficial but also co-dependent in terms of both analysis and interpretation. Giaimo demonstrated how creating corpora to capture the rich linguistic information that we as institutions and as a sector already produce enables us to learn more about what is going on inside our local centres, disciplines, universities, and across the sector more widely. King illustrated what such analysis can tell us about our individual practice and about existing research. Despite their diversity, our accounts show how capturing and interrogating information via corpus linguistic methods can enable researchers to address problems by means that do not rely solely on intuition and experience. In so doing, we make the case that corpus linguistics is accessible to – and useful for – all researchers, and may be particularly helpful to those interested in higher education research.

REFERENCES

Alsop, S. (2016). The 'humour' element in engineering lectures across cultures: An approach to pragmatic annotation. In *Corpus Linguistics on the Move: Exploring and Understanding English through Corpora* 79: 337–361

Alsop, S. and Gardner, S. (2019). Understanding attainment disparity: A corpus-driven analysis of the language used in written feedback information to students of different backgrounds. *Journal of Writing Analytics, 3*, 38–68

Alsop, S., Gardner, S. and Priestley, J. (2020). A corpus-driven investigation of connections between the language of written feedback information and ethnic attainment disparity in UK higher education. Paper presented at the *9th International Conference on Writing* Analytics. University of Southern Florida, USA. 5-8th February 2020.

AntLab (2019) *Links*. https://www.laurenceanthony.net/links.html (7 January 2020)

BASE (n.d.) *British Academic Spoken English (BASE) corpus*. https://www.coventry.ac.uk/base (7 January 2020)

BAWE (n.d.) *British Academic Written English (BAWE) corpus*. https://www.coventry.ac.uk/bawe (7 January 2020)

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

Billot, J., & King, V. (2017). The missing measure? Academic identity and the induction process. *Higher Education Research & Development*, *36* (3), 612–624. DOI: 10.1080/07294360.2017.1288705

Brookes, G. And McEnery, T. (2019) The utility of topic modelling for discourse studies: A critical evaluation. Discourse Studies 21(1), 3-21

BNC. (2018). *British National Corpus 2014*. http://corpora.lancs.ac.uk/bnc2014/ . (7 January 2020)

BNC. (2015). *British National Corpus*. http://www.natcorp.ox.ac.uk/corpus/index.xml?ID=products. (7 January 2020)

CLARIN (2019). *Common Language Resources and Technology Infrastructure*. [website] https://www.clarin.eu/. (7 January 2020)

CLARIN (2020). Resource families. www.clarin.eu/resource-families (7 January 2020)

Clouder, L., & King, V. (2015). What works? A critique of appreciative inquiry as a research method/ology. In Tight, M., & Huisman, J. (Eds.). *Theory and Method in Higher Education Research, 1*, 169–190. doi:10.1108/S2056-375220150000001008.

COCA. (n.d.). *Corpus of Contemporary American English*. https://www.english-corpora.org/coca/. (7 January 2020)

EBSCO (2020) *Academic Search Complete*. https://www.ebsco.com/products/research-databases/academic-search-complete (5 February 2020)

English-Corpora.org (n.d.). English-Corpora.org. https://www.english-corpora.org (7 January 2020)

Future Learn (2020). Corpus Linguistics: Method, Analysis, Interpretation. https://www.futurelearn.com/courses/corpus-linguistics (7 January 2020)

Giaimo, G., Cheatle, J., Hastings, C., & Modey, C. (2018). It's all in the notes: What session notes can tell us about the work of writing centers. *Journal of Writing Analytics, 2,* 225–256.

Giaimo, G. N., & Turner, S. J. (2019). Session notes as a professionalization tool for writing center staff: Conducting discourse analysis to determine training efficacy and tutor growth. *Journal of Writing Research*, *11,* (1), 131-162. doi: 10.17239/jowr-2019.11.01.05

Gries, S. (2009). What is corpus linguistics? *Language and Linguistics Compass, 3* (5), 1225–1241

Hansard Corpus (n.d.).*Hansard Corpus (British Parliament*). [website] https://www.english-corpora.org/hansard/ (7 January 2020)

Heron, M. (2019). Pedagogic practices to support international students in seminar discussions. *Higher Education Research & Development*, *38*, 2, 266–279. DOI: 10.1080/07294360.2018.1512954

Johansson, S. (2008). Some aspects of the development of corpus linguistics in the 1970s and 1980s. In A. Lüdeling, & M. Kytö (Eds.). *Corpus linguistics: An international handbook* (Vol. 1). Berlin, New York: Mouton de Gruyter. pp. 1-13.

Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on the Web as corpus. *Computational Linguistics*, *29*, *3*, 333–347. DOI: 10.1162/089120103322711569

King, V. (2013). Self-portrait with mortar board: A study of academic identity using the map, the novel and the grid. *Higher Education Research & Development*, *32*, *1*, 96–108. DOI: 10.1080/07294360.2012.751525

Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (ed.).Directions in corpus linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991. Berlin and New York: Mouton de Gruyter. pp. 105–122

Linguist (n.d.). *Software related to Text/Corpus Linguistics*. https://linguistlist.org/sp/SearchWRListing-action.cfm?subclassid=7223&SearchType=LF&WRTypeID=2 (7 January 2020)

Lüdeling, A., & Kytö, M. (Eds.) (2008). *Corpus linguistics: An international handbook* (Vol. 1). Berlin, New York: Mouton de Gruyter.

MacNeil, S. (2019). Facing the future - Higher education in the era of artificial intelligence. AdvanceHE webinar, 13 December 2019. https://www.advance-he.ac.uk/programmes-events/calendar/facing-future-higher-education-era-artificial-intelligence-members-only

Martin, J. R., & White, P. R. R. (2005). *The language of evaluation: Appraisal in English*. Basingstoke: Palgrave Macmillan.

Meyer, C. F. (2008). Pre-electronic corpora. In A. Lüdeling, & M. Kytö (Eds.). *Corpus Linguistics: An International Handbook* (Vol. 1). Berlin, New York: Mouton de Gruyter. pp. 1-13.

Meyer, C. (2014). Corpus-based and corpus-driven approaches to linguistic analysis: One and the same? In I. Taavitsainen, M. Kytö, C. Claridge, & J. Smith (Eds.). *Developments in English: Expanding Electronic Evidence. Studies in English Language*. Cambridge: Cambridge University Press. pp.14-28

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

MICASE. (n.d.). *Michigan Corpus of Academic Spoken English.* https://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;page=simple (7 January 2020).

MICUSP. (2009). *Michigan Corpus of Upper Level Student Papers*. [website]. http://micusp.elicorpora.info. (7 January 2020).

Moretti, F. (2013). *Distant reading*. London: Verso.

Murakami, A., Thompson, P., Hunston, S., & Vajn, D. (2017). 'What is this corpus about?': Using topic modelling to explore a specialised corpus. *Corpora, 12*, *2*, 243–277. DOI: 10.3366/cor.2017.0118

O'Donnell, M. (2011). *CorpusTool (Version 3.0)*. Retrieved 4 March 2014, from http://www.wagsoft.com/CorpusTool/index.html

Oxford Text Archive (2019) Oxford Text Archive: A repository of full-text literary and linguistic resources. https://ota.bodleian.ox.ac.uk/repository/xmlui/# (7 January 2020)

Perez-Encinas, A., & Rodriguez-Pomeda, J. (2018). A probabilistic approach to studies in higher education. *Theory and Method in Higher Education Research* (*Vol. 4*), Emerald Publishing Limited, pp. 19–30. https://doi.org/10.1108/S2056-375220180000004003

Schmied, J. (1993). Qualitative and quantitative research approaches to English relative constructions. In C. Souter & E. Atwell, *Corpus-based Computational Linguistics*. Amsterdam: Rodopi. pp. 85-96.

Scopus (n.d.) *Scopus preview*. www.scopus.com (5 February 2020)

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, S., & Rockwell, G. (2020). Contexts. *Voyant Tools.* http://voyant.tools.org (5 February 2020)

Sketch Engine (2020) Sketch Engine: Learn how language works. https://www.sketchengine.eu (7 January 2020)

Taylor, C. (2008). What is corpus linguistics? What the data says. *ICAME Journal 32,* 179–200

Text Encoding Initiative. (2019). *Text Encoding Initiative Consortium*. [website]. https://tei-c.org/. (7 January 2020)

Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics, 10*, *1*, 1–13

Trowler, P. (2012). Wicked issues in situating theory in close-up research. *Higher Education Research & Development*, *31*, *3*, 273–284.

UAM CorpusTool (n.d.) UAM CorpusTool: Documentation. http://www.corpustool.com/documentation.html (7 January 2020)

University of Wolverhampton (2020) Practical Corpus Linguistics for ELT, Lexicography, and Translation. https://www.wlv.ac.uk/courses/ma-practical-corpus-linguistics-for-elt-lexicography-and-translation/ (7 January 2020)

Wigham, C. R. & Céline Poudat, C. (2020). Corpus complexes et standards: Un retour sur le projet CoMeRe (Complex and standard corpora: A look back at the CoMeRe project). *Corpus* [online], 20, 4736. http://journals.openedition.org/corpus/4736. (4 February 2020)

Xu, X. & Nesi, H. (2019a). Differences in engagement: A comparison of the strategies used by British and Chinese research article writers. *Journal of English for Academic Purposes*, 38, 121–134.

Xu, X. & Nesi, H. (2019b). Evaluation in research article introductions: A comparison of the strategies used by Chinese and British authors. *Text and Talk*, 39, *6*, 797–818.