Language, Sexuality and Corpus Linguistics: Concerns and Future Directions

Paul Baker

Abstract

In this paper I discuss the potential that corpus linguistics approaches have to make in terms of enabling research on language and sexuality. After giving some background relating to my involvement in the development of this approach and discussion of some of the benefits of using corpus linguistics, I then outline some potential areas for concern, including: misconceptions of the field as only quantitative, the danger of reading only concordance lines, over-reliance on the idea of removing bias, the tendency of corpus approaches to focus on difference or easily searchable features and issues with copyright and ethics. I then discuss potential future directions that the approach could take, focussing on work in non-western and non-English contexts, the development of new tools such as Lancsbox, and the integration of multimodal analyses, using examples from my own work and others.

Keywords

sexuality, corpus linguistics, methods, language

1. Introduction

This concluding paper to this special issue on Corpus Approaches to Language and Sexuality reflects on a number of present concerns and debates that the field would be advised to consider, and then discusses future directions that could be taken, giving illustrative examples from my own work and others. However, I begin by making a few personal notes about my involvement in this field and its development over the past decade.

Original contributions to knowledge are usually incremental rather than involving the creation of a completely new field or method of research. Combining aspects of two existing fields together can also result in novel forms of triangulatory research, although the story of how such combinations are arrived at can sometimes owe a lot to serendipity or being at the right place at the right time. From 1995 I was lucky to be working at the Department of Linguistics and Modern English Language at Lancaster University, a large department with many research groups and opportunities to hear talks by a wide range of linguistic scholars. Crucially, the department contained the critical discourse analyst Norman Fairclough and the corpus linguist Geoffrey Leech. I had been employed as a researcher on a corpus building project and was simultaneously conducting doctoral research on an unrelated topic, uses of Polari – a form of language spoken by gay men, which broadly fell under 'sociolinguistics'. When I completed my PhD, it felt like a natural progression to combine what I had learnt as a corpus researcher with my continued interest in language and sexuality.

Motschenbacher (this volume) gives an overview of the areas where corpus approaches have been applied to studies of language and sexuality, listing studies on a) representation of LGBT people b) public discourse of sexual relationships and c) practices in sexualised communication. It should be noted that the field is skewed somewhat towards studies that have focussed on LGBT identities (as opposed to say heterosexuality, sexual practices like BDSM or asexuality, monogamy and polygamy and the wider notion of sexual desire).

In Baker (2005) I used corpus methods to examine a range of texts (news articles, television scripts, erotica, personal adverts, parliament debates and safe sex leaflets) written by and/or about gay men. This book, called *Public Discourses of Gay Men*, was mostly written between 2002 and 2004, and was my first attempt to use corpus linguistics to address social questions. While the topic of research (language around gay men) was relatively narrow, I realised that the techniques could be applied to a much wider range of research foci, resulting in a broader book called *Using Corpora in Discourse Analysis* (Baker 2006). A couple of years later, I collaborated with several colleagues at Lancaster: corpus linguists (Tony McEnery and Costas Gabrielatos) and critical discourse analysts (Ruth Wodak, Majid Khosravinik and Michal Krzyzanowski) on a project looking at representation of refugees in the press. This was a fruitful collaboration in terms of identifying different ways that the two approaches could work together to form an iterative framework where different stages helped to both identify and test new research questions or hypotheses (Baker et al 2008).

I was not the first person to use corpora in social research, and my earlier work was inspired by Susan Hunston's analysis of concordance lines about deaf people (Hunston 2002) and Michael Stubbs' (1995, 2002) work on discourse prosody, while around the same time Alan Partington and colleagues were developing Corpus Assisted Discourse Studies, a related approach which eschewed the explicitly 'critical' stance taken by myself and colleagues at Lancaster (see Partington et al 2004). I also note a scant amount of even earlier research in the 1990s, including a seminal technical paper about corpus linguistics and CDA by Gerlinde Hardt-Mautner (1995) and a couple of studies on gender representation using corpora by Carmen Caldas-Coulthard (1993, 1995). I may have been one of the first people to use corpora to examine sexuality though, and in 2004 I gave a workshop at the Lavender Linguistics conference in Washington DC to demonstrate some of the techniques I had used in *Public Discourses of Gay Men*, resulting in a range of responses (including interest, disinterest and disapproval) from an audience of varied backgrounds.

2. Benefits

Corpus studies around sexuality were very rare up until the mid-2000s, although there has been a moderate increase since then. More generally, research on language and sexuality embraces a broad range of techniques and approaches, as well as encompassing or intersecting with fields like linguistics, anthropology, sociology, literature and media studies. So it could be argued that most methods used by language and sexuality researchers will not constitute a 'typical' approach within the field. To make a case for using corpus linguistics, I argue that it is ideally positioned to examine questions around discourses and representations of sexuality, particularly so in cases where text producers are perhaps more careful in terms of openly expressing prejudice.

From a critical perspective, a corpus approach can show how certain innocuous-sounding words or phrases may be relatively frequent but contain particular negative associations due to their repetition in less positive contexts. In tabloid news, for example, I found that the words *gay* and *homosexual* tended to collocate with a set of words suggesting transiency like *fling*, *affair*, *liaison, frolics*, *romps*, *encounter*, *casual*, *occasional*, *experimenting* and *adventure* (Baker 2005). The message that these words gives is that gay desire is temporary – either non-monogamous or not even based upon a real identity. It is a relatively subtle message though, less clearly negative than a text which openly advocates making

homosexuality illegal. Corpus techniques can help to identify these repeated patterns and the evaluative prosodies that accompany them.

And from a queer perspective, the corpus approach, in handling large amounts of text at once, is also well-placed to demonstrate the *range* of discourse positions or representations around a particular sexual subject, which is useful in cases where sexual identities or practices are contested. With so much language data being produced and consumed via online sources, it is now relatively easy to construct and access large reference corpora or very specialised corpora, while a new generation of freeware and online tools enable starting analysts to engage with language data in ways that would have been unthinkable a few decades ago.

Increasingly, corpus data can be gathered from a range of sources or from different time periods. The techniques used by corpus linguists allow comparisons between datasets to be made, and such techniques can help to show how discourse positions around sexuality have changed over time or differ between cultures. Such an approach also works well from a queer perspective in terms of demonstrating how sexual categories and understandings of desire are not fixed but changeable. For example, Love and Baker (2015) compared political debates about homosexuality from 1998-2000 and 2013, using a keywords and collocates approach to see how people who argued against equalising legislation for gay people represented their positions and gay people generally. In the earlier corpus, politicians tended to focus on sexual behaviours, using negative words like *indulge* and *dangers*, arguing that a young man of 16 would be 'ruined for life' if seduced by man, becoming a 'promiscuous… lonely old homosexual'. The debate focussed on predatory older gay men and contained references to a 'homosexualist agenda'. Speakers were much more likely to use the word *homosexual* (which referenced medical and legal discourses as well as being associated with sexual behaviour) than *gay* (a term which had more associations with identity and community). The key argument given for not changing the law (to equalise the age of sexual consent) was that homosexuality was wrong. Only a few years later, in the 2013 debate, the equivalent set of politicians now used the word *gay* more than *homosexual* and acknowledged the existence of a gay community, attributing positive qualities such as *talented* and *caring* to it, with homosexuals being described (albeit somewhat patronisingly) as *delightful*, *artistic* and *loving*. While these politicians were still opposed to equality (this time marriage equality) for gay people, their arguments were focussed around matters of procedure, redefinition of the term *marriage*, the effects of the change to the law on the church and the view that many gay people themselves did not want the law to change. In other words, they used any argument *but* the one that homosexuality was wrong. The analysis shows critically how representations of sexual desire can shift from a (bad) behaviour to a (good) identity in a relatively short amount of time.

3. Misconceptions

In the 2000s I was one of a very small number of people combining corpus linguistics and some form of discourse analysis (critical or otherwise), a number which was even smaller when only studies around sexuality were considered. It remains a small field, although it is an approach which has gained more acceptance over time, yet as I will argue in this section, it still has a tendency to be misunderstood by both its supporters and its detractors.

For example, corpus linguistics can be incorrectly conceived of as a merely quantitative approach, based only on computational procedures and statistical tests, with results published

in the form of tables of numbers and p values. While 'number-crunching' plays a part in corpus linguistics, for those with interests in social phenomena, this ought to be a start to the analysis, constituting the means to an end. For a corpus analysis to be effective, a close linguistic study of the texts in the corpus needs to be undertaken, along with consideration of a range of different types of relevant context. This could include, for example, examination of the practices around the production and reception of the texts in the corpus, and consideration of the legal, medical, religious, historical and social status of social groups (such as LGBT people) in the society where the texts came from. Tables of numbers do not interpret or explain themselves, so there needs to be considerable human input. The corpus tools help us to process large amounts of data so that potentially interesting linguistic patterns can be identified effectively, but they do little more than that.

However, when combining fields of research together, we inevitably require more from our analysts. A brief list of requirements of a corpus based language and sexuality researcher would include familiarity with statistical tests (how to use software to carry them out, how to know which ones to use and how to interpret the output), how to use corpus analysis software like WordSmith or SketchEngine, how to collect, clean and annotate corpora or how to find an appropriate existing corpus, how to develop categorisation schemes through analysis of keywords, collocates and concordance lines, how to carry out linguistic analysis via close reading of texts, and how to carry out analysis of social and historical context, as well as considering production and reception of texts. Unless we have unlimited time and money, along with a team of people with complementary skills, it is unsurprising that some aspects of such a research project will be backgrounded. I would stress here the importance of collaborative research of the kind carried out by Baker et al 2008. In other words, if you are not a corpus linguist, make friends with a corpus linguist.

A related misconception about corpus-based sexuality research (or indeed any kind of corpus-based research) is that it is the same as a set of quantitative approaches that have a computational linguistics slant. These include techniques or approaches like sentiment analysis, opinion mining, topic modelling and culturnomics and may be associated with terms like 'the digital humanities' or 'big data'. Such approaches often involve the use of 'black box' tools (meaning that their workings are not understood or accessible to users), and can thus be carried out by academics or professionals working in non-academic fields (e.g. healthcare) who have access to some text but little or no background in linguistics or discourse analysis. As a result, the analysis will often be at best a kind of content analysis which summarises (not always accurately) what a set of texts is about or makes broad claims about the amount of 'positive' or 'negative' language within it, sometimes relying heavily on automatic word tagging systems which assign meaning or grammatical function to words in ways that are not always the most nuanced or accurate. Such studies tend to rely heavily on reporting the output from the tools, at times engaging in guesswork about what the output means, and there can be limited close reading of texts or consideration of types of context beyond the texts themselves. Hardie (2017), for example, demonstrates how a topic modelling study on an academic corpus resulted in different topics being identified when the process was carried out a second time, and that some topics were impossible to label, being based on a what looked like an unrelated set of words. Another aspect of computationally-led studies is the use of visuals (e.g. word clouds) which look pretty but tend to be rather reductive in terms of the information they give and can be presented as the outcome of an

analysis rather than the starting point. When I see a word cloud I want to know why such words appear in a particular text, what their typical and atypical functions are, how they occur in relationship to one another and in their wider contexts, and whether they are evenly distributed across an entire text or set of texts. This all requires more detailed analysis, which is not usually done. In corpus linguistics we dig deeper and are informed by understandings about language, taking into consideration phenomena like grammatical agency, metaphor, synonymy, hyperbole, euphemism and over-lexicalisation as well as multiple levels of context.

For anyone who is not familiar with corpus linguistics, it is easy to assume that it is the same as other 'big data' approaches which also use computer software to do something with texts. It is not.

4. Concerns

In this section I would like to discuss five issues within the field which require consideration and debate. They do not preclude the use of corpus techniques in sexuality research but they should make us pause for thought, and encourage us shift or widen our focus.

First, there is an issue relating to the way that corpus linguists traditionally carry out analyses of co-text (verbal context) within the corpus itself. Typically, we first use software and attendant statistical techniques like keywords or collocates to identify salient words, clusters, part of speech tags or other linguistic features which enable us to narrow our focus. As intimated above, these linguistic features are subjected to more detailed qualitative analysis, which usually amounts to a thorough examination of concordance lines (although sometimes even this is not done). The choice of which concordance lines to examine does not always result in a representative analysis though. A common error when faced with hundreds of lines, is to simply look at the first 20 or so, meaning that only the first few texts in any corpus actually end up being interrogated. If sampling is to be used, it is best used on a random set from the whole corpus, and if the patterns found from the sample are complex or inconclusive (for example, in cases where a large number of patterns are found), it is sensible to consider a second or third random sample.

However, the issue here is not so much with which concordance lines to look at but the fact that we rely on concordance lines per se. A concordance line typically contains the search term (usually a word) with around 5-10 words of co-text either side of it. The limited nature of the concordance line can mean that analysts will only receive a keyhole glimpse of the co-text. At best, this forces researchers into a mindset of only identifying the linguistic patterns around an item in the immediate vicinity. At worst, it means that people can jump to the wrong conclusion or overlook an important analytical point because it appears just outside the concordance line. For example, we may see a concordance line which contains a very negative construction of gay people within it. But this may be part of a quote which an author is using in order to show that some people are homophobic, which is part of an overall argument framing homophobia as problematic. While the homophobic pattern exists then, its use is more complex than just being to relay a homophobic attitude. But a concordance analysis is often ineffective in telling us about how authors orient to what other people have said, and at times we will not be able to know that something is a quote, just from looking at a concordance line.

Additionally, concordance lines are not texts. They are snippets of texts and appear within a much wider context. The position of a word in a text may be important in determining speaker or writer stance. For example, newspaper texts often contain what the editors feel is the most important information at an early point in the article. If a person is quoted in an article at the start (as well as being given a lot of space), it is often the case that the newspaper aligns itself with the quoted view (unless the speaker is clearly been quoted to provoke outrage). But a concordance analysis alone will miss issues like amount of text quoted and position of the quote. Other techniques, such as considering dispersion or employing a tagging system so that all quoted text is clearly distinguished in concordance tables may help to resolve this matter, but ultimately I would advocate that concordance lines are expanded as much as possible. A concordance analysis is perfect for spotting prosodies – sets of negative or positive collocates around a search term which subconsciously imbue the search term with that evaluative force, but research that involves the representation of sexuality (or other complex and highly debated concepts) requires greater consideration of co-text than concordance lines.

Second, as I argued in Baker (2006) the corpus approach could be conceived of as a scientific, neutral, unbiased, objective method, set up in opposition to more qualitative critical forms of analysis which supposedly 'cherry pick' texts to prove a preconceived point (Widdowson 2004). And while it is harder to cherry pick a whole corpus, the corpus approach in itself does not remove bias. The analyst has to decide which techniques of analysis to implement, and then has to impose cut-off points to determine what counts as a frequent item, a keyword or a collocate. Even then, there is often still too much data to analyse so researchers may then select a few items from the keyword list to look at in more detail, and unless they are very explicit about their decision-making process, this can start to resemble a cherry picked process where we question whether the analyst's own interests, positions and background knowledge have influenced the research to the point where it is little more than a biased polemic.

As an illustration, Baker and Egbert (2016) describe a study where the same corpus and research questions were presented to ten independent analysts, and the subsequent findings were compared. Between them the authors made 91 findings in total, although 82% of those findings were only mentioned by a single author. Only six findings out of the 91 were made by two or more authors, while in three cases, authors made claims that were the opposite of one another. Thus, it is difficult to conclude that any given set of analysts can be given the same corpus and questions and yield the same results. Instead, they are much more likely to obtain different but complementary outcomes. We could take care both to temper our criticisms about the cherry picking of qualitative approaches as well as take steps to be more systematic and explicit in our decision-making processes when analysing corpora. We can never produce a fully objective analysis, but as much as possible we want to avoid analytical holes that mean what we claim is unconvincing to others.

Third, corpus tools are good at counting, sorting and carrying out statistical tests on language data, but they can sometimes lead us down certain analytical paths that may preclude a fuller analysis. The keywords approach, for example, requires the analyst to compare two corpora together in order to identify sets of words that are statistically frequent in one corpus when compared to another. This technique can be used gainfully to give focus to analysis, but it can also mean that the analysis only considers the differences between the two corpora, rather

than the similarities. A more subtle but interesting story about our data might occur within the similarities. And in the application of frequency-based cut-offs in order to reduce the number of linguistic items we need to analyse, we may also end up only considering the most typical patterns in the corpus, at the expense of the less frequent ones. As long as the limits of the analysis are made clear, then this is not too problematic. There is at least the potential for a corpus analysis to identify the large and the small in terms of linguistic tendencies, something which a qualitative analysis of a few texts may not be so easily achievable. A related issue is that some linguistic forms (such as words or fixed sequences of words) are easier to identify than others, meaning that our analysis might be skewed towards them and overlook something important but linguistically too complex or variable for a corpus tool to identify (such as pragmatic features or zero grammatical features). Dedicated corpus researchers can sometimes find workarounds by using tagging and complex search terms, although even then, they may not identify every case (see Baker 2014 for a discussion of searching for disagreements in a corpus). Qualitative analysts who carry out close readings might be better placed to spot such phenomena, and I would advocate an approach which shifts between readings of samples and corpus techniques in order to have the best of both worlds.

Fourth and fifth, corpus studies yield a host of issues relating to collecting and sharing texts, involving copyright clearance and ethics. The large amount of online data now available to us is attractive to corpus builders but copyright is a grey area and despite it being possible to download online data from large parts of the planet, actual laws differ from country to country. The UK changed a copyright law in 2014 to allow large amounts of online data to be downloaded for academic research: 'The new copyright exception allows researchers to make copies of any copyright material for the purpose of computational analysis if they already have the right to read the work (that is, work that they have "lawful access" to). They will be able to do this without having to obtain additional permission to make these copies from the rights holder.' (Intellectual Property Office 2014: 6). However, this law only applies to academic research carried out within UK institutions. I know of no corpus linguistics who have been sued for taking publicly accessible online data for academic purposes, although I have found interactions with copyright holders of large databases to be inconsistent, as well as having had similarly frustrating experiences with publishers when I have tried to publish corpus studies using online data. In one recent edited collection I was involved in, the publishers removed a chapter involving a corpus of online gay dating profiles, despite the abstract being OK'd at the outset and the chapter containing no examples of text that were longer than a couple of sentences.

Publishers can sometimes take time to understand what corpus linguistics involves and may try to apply rulings based on experiences working in other academic fields. I would advise researchers who are concerned about publishers rejecting their corpus analysis to check with them in advance, and ensure their questions are accompanied by examples of previously published research using similar corpora and links to the 2014 copyright exception if applicable.

In the interests of replicability, it is good practice to try to make corpus data available to others where possible, yet this can also result in copyright concerns. Some corpus builders have tried to reach a compromise by putting their corpora online and allowing concordance searches to show snippets of text, but they do not allow the corpus to be downloaded in full. Instead, analysts must work with the online tool that comes with the corpus. While this

enables access to a much wider range of data, it also means that limitations are placed on analysts in terms of what can be done with it.

A related concern relates to the ethical side of collecting corpus data, especially large amounts of online data. While corpus builders may not want to consider asking permission to build a corpus of newspaper articles, a lot of online data is created by individuals, not writing for profit and not backed by a large institution. Qualitative researchers may find the permissions process frustrating but often they are only dealing with a small number of authors so the task of contacting them is achievable. If our corpus contains hundreds or thousands of texts, all from different sources, the permissions-seeking task is exponentially more difficult. Not everyone is contactable and not everyone appreciates the value of academic research (especially if our values do not match those of the original author). Returning to the example above, if we build a corpus of online dating profiles, should we seek permission from all the people who posted the profiles, even if we do not intend to share the corpus with anyone? Should we only ask for permission if we want to quote an excerpt of someone's profile in the analysis? It could be argued that even if it was feasible to contact the creator of every advert or dating profile, if some did not give permission the aim of creating a representative corpus would be compromised.

Perhaps anonymization is more important than obtaining permission in such cases. But to give another example (which is now increasingly common in corpus based research), anonymization can be difficult in online contexts. Imagine that we want to research homophobic language and quote some homophobic tweets from a corpus we built. Even if we try to anonymise the tweet, an online search of it may be able to locate it fairly quickly – making true anonymization very difficult. It could be argued that people who use a social networking platform like Twitter are 'fair game' because they know their tweets are public. But what if the person who wrote the tweet is a child and we have no way of knowing this? Or what if they are older, and if by bringing attention to their homophobic tweet they lose their job and their children suffer as a result of that? There needs to be a balance between taking an ethically sensitive approach so that individuals are not compromised by our research, but we do not get so tied up in ethical ruminating that corpus research becomes impossible. In Baker and McEnery (2015) when we analysed a twitter corpus about people receiving government benefits we made the decision to not directly quote tweets which advocated violence. We need to work towards the creation of guidelines for ethically-responsible corpus building, but should also bear in mind that it might not be possible to produce a definitive set of guidelines which can apply across every study. Instead each study needs to be carefully assessed in its own right.

5. Future Directions

In this part of the paper I outline three future directions that the application of corpus linguistics to research in language and sexuality could take. First, I note that most research in this area currently involves English language corpora and usually involves British, North American or Australian texts. There is a gap then, involving the method being applied to non-western and non-English contexts. In the late 1990s I was involved in a corpus building project for Indic languages like Bengali, Urdu and Punjabi (McEnery et al 2000). One challenge that we faced was the fact that not a great deal of relevant data was available online and much of what did exist tended to be incompatible (either in image (gif or jpg) form,

which corpus tools could not interpret, or used a variety of bespoke fonts, where inconsistent encoding schemes had been applied (so to copy the same text from one Punjabi font to another, a different combination of keyboard presses would be required). The situation is now much improved, so technically speaking, there is no reason for this gap – most corpus tools do not place limitations on the writing systems that they work with. They view letters as codes, and are usually compatible with Unicode or UTF-8, both of which have the capacity to represent most of the world's languages and are increasingly commonly used as a standard. Instead, the main irritant for corpus builders are pdf files, which require conversion to plain text and can often contain so much background formatting code that they can be incomprehensible once converted.

Two recent non-English studies using corpus research are Bogetic (2013), who built a corpus of personal adverts posted by gay Serbian teenagers. This study examined collocates, finding that they indexed dominant values, and Silva Paredes (2017) who examined websites containing religious discourses around homosexuality in Chile. Her analysis found that the Catholic Church constructed homosexuality as a tendency and an act, as opposed to an identity, while at the same time addressing accusations of homophobia by positioning gay people as the beneficiaries of pastoral care from a caring church.

While corpus tools can work well with most languages, the amount of existing corpus resources tends to favour English, which is much better provided for in terms of reference corpora (which are often required in order to elicit keywords or provide examples of typical patterns around linguistic features) and user expertise. Corpus research which works across more than one language can present additional problems. For example, in comparing similar corpora of English and French, it can be difficult to apply cut-offs in a consistent way. French has more verb tenses than English, resulting in many verbs having low individual frequencies and not appearing in lists of frequent words. As noted earlier, such issues do not mean the research cannot be carried out (one solution would be lemmatisation), but it does require additional steps and thinking through of issues that might not arise otherwise.

Corpus linguistics is impossible without software (or the ability to code), and the field is lucky to have a small number of dedicated, helpful and responsive software engineers, who give very large amounts of their time to assist software users, as well as providing updates to incorporate new features that have been requested. This is essential for the field – especially for people who cannot code we can only be as good as the tools allow us to be. A second future direction then, involves the development of tools which enable more sophisticated forms of analysis. As an example of software opening up new areas of research, Figure 1 is a screenshot from a free tool called LancsBox (Brezina et al 2015), which has a facility to visually show collocates between words.

Unlike earlier tools, LancsBox views words as existing within a network of multiple collocates, rather than simply considering collocation as merely involving pairs of words. These networks give a unique picture of collocation, enabling the identification of different and new sorts of linguistic patterns. Figure 1 shows collocates of the word *man* in a corpus of online gay erotic narratives which I had analysed previously (see Baker 2005). For the analysis shown here, I first found the collocates of *man* (by clicking the mouse on this word) which is shown at the centre of the network. Nine collocates were elicited: *handsome, meat, woman, hey, young, juice, older, younger* and *married*. Exploration (via concordancing) of

these collocational relationship shows some of the characteristics that are regularly eroticised in these narratives – *handsome*, *young*, *older*, *younger* and *married*, indicating that age differentials are often viewed as attractive, while being married is also seen as an indicator of heterosexuality, masculinity and unavailability, helping to represent sexual partners who have these qualities as highly prized. Other collocates of *man* offer slightly less expected uses: *man juice* is a euphemism for semen whereas *hey man* occurs in account of spoken dialogue, an informal greeting which helps to establish the masculinity of the characters. The lengths of the lines show the strength of collocation, with words that are closer together being more strongly attracted to one another.

So far, this kind of analysis could have been carried out using most existing corpus software. But by clicking on the collocates of *man* we can obtain their collocates, adding new collocates to the network (e.g. additional collocates of *young* are *beautiful*, *men* and *guys*), as well as showing additional links between existing collocates (such as the collocational relationship between *young* and *handsome*). As a result, we start to see more complex patterns of collocation. For example, *man*, *handsome* and *young* form a triangle, collocating with one another. However, another word, *beautiful*, which is semantically similar to *handsome*, also collocates with *young* but does not collocate directly with *man*.

Attractiveness is clearly associated with youth in this corpus, and the phrase *handsome young man* is reasonably common, but *beautiful young man* is not. Instead, the adjectives *beautiful* and *young* occur together with other nouns like *woman*, *boy*, *dudes*, *body* and *buns* (although the network would need to be expanded further to show these words). The analysis helps to show how attractiveness and masculinity are constructed in the narratives in different ways for different types of social actors or body parts.

[FIGURE 1 HERE]

Figure 1. Collocational network output from LancsBox

This kind of network therefore gives a multidimensional view of collocation which takes into account the fact that collocates do not occur as isolated pairs, but in endlessly connected relationships to other words. It also implies that we should be considering groups of collocates, and the different combinations of links between 3, 4, and more words can suggest different semantic and grammatical relationships (Baker 2016). LancsBox was developed at Lancaster University as a result of corpus linguists specifying the kinds of analysis they wanted to do and working with software engineers to create them.

A third possible future direction for corpus linguists working in sexuality to consider is to employ a wider range of text types in their analyses. Most corpus research is carried out on electronic, representations of written or spoken words which are rendered as letters on a screen. A spoken corpus normally contains a written transcription intended to represent the features of the speech, and may include paralinguistic information like laughter or pauses. Sound files of the original utterances can be aligned to the transcription, as is the case with part of the British National Corpus, although this practice tends to be the exception to the rule. Many spoken corpora do not allow access to sound files, let alone video data. Additionally, written corpora are often in a form where much of the original formatting has been stripped away, so textual information like font type, size and colour, as well as relative positions of text on a page, and use of boxes, lines or other graphics are absent. In my own

work with large newspaper corpora collected from online databases, the articles do not contain the original images that accompanied them. Yet words (spoken and written) do not occur in isolation. Language is multimodal, and corpus linguists need to be challenged to acknowledge and incorporate this fact into their work in a meaningful way.

One approach would be to carry out a corpus analysis occur alongside a multimodal analysis. For example, corpus techniques like keywords or collocates could be employed on the text only files, but as a supplementary approach we could examine the corpus texts in their original form, implementing some sort of visual analysis alongside the texts, perhaps on a sample of the texts if the corpus was very large. The two forms of analysis would therefore be separate but linked components, a form of triangulation in other words. This is an approach taken by Ismail (2017), who examined a corpus of news articles about male and female athletes.

A second approach would be one which tries to integrate the two forms of analysis together, viewing the written text and the visual analysis as operating in a relationship with one another. For example, McGlashan (2016) analysed a corpus of children's books with titles like *Jenny Lives with Eric and Martin* which contain same-sex care-givers. McGlashan identified frequent words, clusters and keywords in the corpus and then carried out concordance analyses in order to demonstrate how these linguistic features functioned in the texts, particularly in terms of challenging heteronormative assumptions. However, he also paired concordance lines to the images that co-occurred on the page where the text from each concordance line appeared, and analysed these sets of images in order to determine the ways that they contributed to overall understandings of the texts. In some cases the images helped to provide additional meanings which would not have been identified if only the words on the pages had been considered. McGlashan refers to a concordance which contains images linked to each line as a *collustration*, a blend of the words *concordance* and *illustration*. Meaning is thus made by a consideration of the juxtaposition of the image and the text together. His research shows the potential that a multimodal corpus analysis could have, allowing analysts to make connections and notice patterns that go beyond the written word. However, in the absence of tools to automate this kind of visual analysis, a great deal of manual work was required. It is hoped that newer generations of corpus analysis software will be able to more easily enable different forms of multimodal analyses.

6. Conclusion

In this concluding paper I have highlighted some of the misconceptions, challenges and future directions that I see for researchers who wish to use corpora to analyse questions of language and sexuality. As noted, this is still a small field, but the presence of a special issue in this journal suggests that there is growing interest. It certainly feels rather less lonely than it did at the start of the century. While a disadvantage of being around at the conception of a field is the resistance (passive or otherwise) which we may occasionally encounter, and the sense that we must continuously justify or 'sell' our approach to others (when we would often prefer to be just getting on with it), there are also advantages. The lack of established procedures or an existing research canon means that truly innovative steps can be taken, both in terms of developing the approach and in terms of what can be found from it. This special issue marks the end of the beginning. There is still much work to be done.

References

Baker, Paul. 2005. *Public Discourses of Gay Men*. London: Routledge.

Baker, Paul. 2006. *Using Corpora for Discourse Analysis*. London: Continuum.

Baker, Paul. 2014. *Using Corpora to Analyse Gender*. London: Bloomsbury.

Baker, Paul. 2016. The shapes of collocation. *International Journal of Corpus Linguistics* 21(2): 139-164.

Baker, Paul, Gabrielatos, Costas, Khosravinik, Michal, Krzyzanowski, Majid, McEnery, Tony & Wodak, Ruth. 2008. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse and Society* 19(3): 273-306.

Baker, Paul & McEnery, Tony. 2015. Who benefits when discourse gets democratised? Analysing a Twitter corpus around the British Benefits Street debate. In *Corpora and Discourse Studies: Integrating Discourse and Corpora*, Paul Baker & Tony McEnery (eds), 244-265. London: Palgrave.

Baker, Paul & Egbert, Jesse. (eds) 2016. *Triangulating Methodological Approaches in Corpus-Linguistic Research*. London: Routledge.

Bogetić, Ksenija. 2013. Normal straight gays: Lexical collocations and ideologies of masculinity in personal ads of Serbian gay teenagers. *Gender and Language* 7(3): 333-367.

Brezina, Vaclav, McEnery, Tony, & Wattam, Steve. 2015. Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2): 139-173

Caldas-Coulthard, Carmen Rosa. 1993. From Discourse Analysis to Critical Discourse Analysis: The differential re-presentation of women and men speaking in written news. In *Techniques of Description: Spoken and Written Discourse*, John M. Sinclair, Michael Hoey and Gwyneth Fox (eds), 196-208. London: Routledge.

Caldas-Coulthard, Carmen Rosa. 1995. Man in the News: The misrepresentation of women speaking in news-as-narrative-discourse. *Language and Gender: Interdisciplinary Perspectives*, Sara Mills (ed.), 226-39. Harlow: Longman.

Hardie, Andrew. 2017. Exploratory analysis of word frequencies across corpus texts: Towards a critical contrast of approaches. Plenary paper presented at Corpus Linguistics Conference, Birmingham, UK, 25th July 2017. <https://www.youtube.com/watch?v=ka4yDJLtSSc> (February 22, 2018)

Hardt-Mautner, Gerlinde. 1995. *Only Connect. Critical Discourse Analysis and Corpus Linguistics.* UCREL Technical Paper 6. Lancaster, UK: Lancaster University.

Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Intellectual Property Office (2014) *Exceptions to copyright: Research*. IPO: Newport. <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/375954/Research.pdf> (February 22, 2018)

Ismail, Habibah. 2017. *A Corpus-assisted Multimodal Discourse Analysis of Malaysian Sports News Discourse: Exploring the Representation of Female and Male Athletes*. Unpublished PhD thesis. University of Sydney, Australia.

Love, Robbie & Baker, Paul. 2015. The hate that dare not speak its name? *Journal of Language Aggression and Conflict* 3(1): 57-86.

McEnery, Anthony, Baker, Paul, Gaizauskas, Robert, & Cunningham, Hamish. 2000. EMILLE: towards a corpus of South Asian languages, *British Computing Society Machine Translation Specialist Group* 11: 1-9.

McGlashan, Mark. 2016. *The representation of same-sex parents in children's picturebooks: a corpus-assisted multimodal critical discourse analysis*. Unpublished PhD thesis. Lancaster University, UK.

Partington, Alan, Morley, John & Haarman, Louann. 2004. *Corpora and Discourse*, Bern: Peter Lang.

Silva Paredes, Daniela. 2017. Discourses of gay people and homosexuality in Chilean Church Discourse. (Paper presented at the 24th Lavender Languages and Linguistics Conference, University of Nottingham, UK 29th April 2017).

Stubbs, Michael. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative methods. *Functions of Language* 2(1): 1–33.

Stubbs, Michael. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. London: Blackwell.

Widdowson, Henry. 2004. *Text, Context, Pretext. Critical Issues in Discourse Analysis*. Oxford: Blackwell.

Author details

Paul Baker

Department of Linguistics and English Language

Lancaster University

Lancaster

Lancashire

LA1 4YL

UK

Email: p.baker@lancaster.ac.uk