Article

Discourse & Communication
2015, Vol. 9(2) 221–236

DOI: 10.1177/1750481314568542
dcm.sagepub.com
SAGE

Picking the right cherries? A comparison of corpus-based and qualitative analyses of news articles about masculinity

Paul Baker
Lancaster University, UK

Erez Levon
Queen Mary University of London, UK

Abstract

As a way of comparing qualitative and quantitative approaches to critical discourse analysis (CDA), two analysts independently examined similar datasets of newspaper articles in order to address the research question 'How are different types of men represented in the British press?'. One analyst used a 41.5 million word corpus of articles, while the other focused on a down-sampled set of 51 articles from the same corpus. The two ensuing research reports were then critically compared in order to elicit shared and unique findings and to highlight strengths and weaknesses between the two approaches. This article concludes that an effective form of CDA would be one where different forms of researcher expertise are carried out as separate components of a larger project, then combined as a way of triangulation.

Keywords

Corpus, discourse, masculinity, qualitative, triangulation

Introduction

Corpus approaches to critical discourse analysis (CDA) have grown in popularity since early pioneering work by Louw (1993), Caldas-Coulthard (1993, 1995), Hardt-Mautner (1995), Krishnamurthy (1996) and Flowerdew (1997). Proponents of this approach have

Corresponding author:
Paul Baker, Department of Linguistics and English Language, Lancaster University, Lancaster LA1 4YT, UK.
Email: p.baker@lancaster.ac.uk

put forward several arguments for the use of a large body of electronically coded text which is subjected to analysis of linguistic patterns via computer software. First, findings are more likely to be reliable and valid if shown to occur across a larger dataset. Second, a corpus approach guards against the accusation that critical discourse analysts could 'cherry-pick' or intentionally select (possibly atypical) data or linguistic features for analysis to prove a preconceived point (Widdowson, 2000, 2004). Third, corpus procedures offer findings based on frequency patterns so it becomes possible to indicate the commonly realised (and less popular or minority) discourses in societies. Unlike CDA, corpus linguists often take a 'bottom-up' or data-driven approach to their research, beginning with few strong (or conscious) hypotheses or expectations about what they will find. Instead, the corpus processes drive the analysis, and linguistic patterns based around what emerges as frequent or salient need to be accounted for. Furthermore, a corpus approach to discourse analysis is not always critical. For example, one school within such corpus approaches is Corpus Assisted Discourse Analysis (CADS), which is noted by Partington et al. (2013) as 'not tied to any particular school of discourse analysis… unlike CDA, it has no overarching political agenda' (p. 10). On the other hand, the approach developed by Baker et al. (2008) is inspired by the Discourse Historical Approach (DHA) (Reisgl and Wodak, 2001) and is more oriented towards providing a critical perspective. With corpus approaches being an 'emergent' technique for CDA, however, there are questions about the extent to which such an analysis can usefully supplement or replace more traditional and qualitative techniques. Additionally, does a larger corpus necessarily guard against cherry-picking? Ultimately, will different methods produce wildly different results?

The main aim of our research was to carry out an experiment where the two authors worked independently of each other, one using corpus-based techniques, the other focusing on a qualitative analysis, in order to address the same research question which involved the representation of different types of masculinity in the British press. The corpus analyst (Baker) used a large corpus of approximately 44.1 million words of newspaper articles stored in electronic form, while the qualitative analyst (Levon) studied a down-sampled set of 51 articles. We were interested in detecting whether there were broad similarities or differences between the research findings and how this related to the different methodologies that were undertaken.[1] A point of concern was that there would be no overlap between our findings, or worse, that the two analysts would produce contradictory conclusions. A secondary aim was to create and test a means of down-sampling the corpus data for the qualitative analysis as a way of selecting the most representative articles. While down-sampling is certainly a form of 'cherry-picking', we aimed to at least choose cherries from an informed position. Finally, we wanted to see whether a form of the 'triangulatory' approach that was undertaken here could be recommended to other researchers as a means of producing a better (more complete, rigorous, interesting and/or deeper) analysis.

After discussing relevant studies in corpus linguistics and/or CDA that have oriented towards triangulation, we then describe how we collected and down-sampled the corpus used in this study. This is followed by an account of the ways that the analyses were carried out, along with a brief summary of the two sets of results. The final sections of this article compare and critique our findings, discussing and attempting to find explanations

for the strengths and weaknesses of the two techniques as well as making suggestions for how other researchers could gainfully combine corpus approaches with CDA.

Triangulation

In social science, research triangulation (Cicourel, 1969) involves carrying out two or more approaches as a means of checking results. It is derived from land-surveying techniques which 'determine a single point in space with the convergence of measurements taken from two other distinct points' (Rothbauer, 2008: 892) and can involve the implementation of multiple datasets, investigators, theories or methods (Denzin, 1970). In qualitative research, triangulation can be used as an alternative to traditional measures of reliability and validity, enabling researchers to overcome limitations associated with a single method or their own biases. CDA often involves triangulation; for example, Van Dijk (2006) describes his approach to CDA as having a 'theoretical framework [that] is multi-disciplinary, articulated by the fundamental triangulation of discourse, cognition and society' (p. 115), while Wodak (2007) notes that 'One of the most salient features of the discourse-historical approach is its endeavour to work interdisciplinarily, multimethodically and on the basis of a variety of different empirical data as well as context theories' (p. 210).

A third form of triangulation that relates to CDA is in combining close qualitative readings with a corpus linguistics approach that uses computer software to identify frequent and salient linguistic patterns over large amounts of data. Such an approach is described by Baker et al. (2008) as a 'useful methodology synergy'. One strand of this research involved one team of researchers (Gabrielatos and Baker, 2008) working on an entire corpus of news articles about immigration (140 million words) taken from 19 British newspapers, while another team carried out a close reading (using the DHA) of a smaller set of 439 articles which had been down-sampled (KhosraviNik, 2010). The down-sampling was based on taking sets of articles from three newspapers during a small number of periods which had the most articles about immigration. The paper by Baker et al. (2008) noted a couple of differences between the two sets of findings. For example, the DHA 'at times facilitated a more detailed analysis, taking into account larger amounts of textual context as well as the structure and characteristics of the employed genres' (Baker et al., 2008: 296), while the

> corpus-based approach … uncovered a small number of articles where 'positive' topoi of RASIM were employed in the corpus. This was different to the CDA analysis, which, focusing on a smaller number of articles, concluded that positive topoi were almost non-existent. (Baker et al., 2008)

However, the paper was more strongly focussed around the ways that the two approaches could be combined into nine stages, advocating moving back and forth recursively between qualitative and quantitative forms of analysis in order to generate new hypotheses as well as to test existing ones. A pertinent question that arises out of this research is whether the technique for down-sampling produced a representative set of texts for a comparable analysis with the entire corpus. In focusing on just three newspapers (two broadsheets,

one middle-market) during a small number of short periods of intense discussion about immigration, the CDA was potentially restricted to the analysis around 'big news' events which generated a lot of discussion.

Two further studies have attempted to address whether corpus-based approaches to discourse analysis can help to provide a consistent set of findings independent of analyst or even method used. Research by Marchi and Taylor (2009) and Baker (forthcoming) involved experiments where more than one researcher independently worked with corpus data and then compared results. Marchi and Taylor (2009) used a corpus of newspaper articles to address the question 'How do journalists talk about themselves/each other and their profession in a corpus of British media texts?'. They uncovered a mixture of dissonant, converging and complementary findings. The former indicates findings that are incompatible with each other, while converging findings are those which confirm one another. Marchi and Taylor (2009: 6) warn that while such findings are often the aim of triangulation (e.g. to demonstrate greater validity), they do not necessarily indicate greater reliability as the researchers may be equally wrong. Complementary findings are seen as part of a jigsaw puzzle (via Erzberger and Prein, 1997) which 'may offer a more complete view of the construct which is being investigated, and as such is a highly productive aim' (Marchi and Taylor, 2009: 7). Baker's (2014) study compared five analysts who worked independently on a corpus of newspaper articles about foreign doctors, in order to uncover representations of this group. He attempted to summarise and quantify the findings from the reports produced by the analysts, noting that about a quarter of findings were shared by at least three out of five analysts (or the majority). This included mostly frequently mentioned representations in the corpus, such as foreign doctors constructed as having poor English language abilities, being incompetent and requiring tougher regulation. However, the majority of findings (about two-thirds) were only uncovered by single analysts. Such less-mentioned findings related to infrequent phenomena in the corpus, which may explain why they tended to be overlooked in a corpus analysis which usually foregrounds high frequency phenomena. Baker notes two strategies for analysis that appeared to be particularly productive. The first was to carry out one methodological technique in a very thorough way (e.g. by carrying out a concordance search of a relevant term and reading every concordance line carefully, rather than say, a sample of lines). The other productive technique involved using multiple methods as a form of triangulation.

With the two studies described above, the analysts were all corpus linguists, although they were free to choose from a range of different tools and techniques and as such the method sections of each analysis were unique and would have helped to lead researchers down particular analytical routes while closing others off. It is also important to bear in mind individual differences as being a factor that is difficult to control for. Even asking multiple researchers to try exactly the same technique could result in unique outcomes as certain aspects of the data may appear more interesting to one person than another. Studies like those described above, along with our own experimental comparison of methods, should not be taken to be definitively controlled in other words. Rather they are indicative of the types and amounts of difference and similarity that could be found between two methods, along with suggestions relating to strengths and weaknesses of each.

Thus, our main aim in this study is to carry out a more comprehensive comparison of critical corpus and qualitative methods, akin to the study by Marchi and Taylor (2009)

where two researchers (the authors) separately undertake to analyse either a whole corpus or a down-sampled set. Compared to the study by Baker et al. (2008), we have used a different method for down-sampling which we feel is more fully representative of the corpus in two ways (number of newspapers and periods of time). Our study focuses on the shared and unique findings elicited from the two approaches, as well as discussing possible reasons for any discrepancies. The following section describes the dataset we collected and the two ways in which it was analysed.

Data

The online news database Nexis UK was used to build a corpus of British newspaper articles about masculinity. Articles were collected from nine daily national newspapers (and their Sunday equivalents)[2] between the years 2003 and 2011. It was stipulated that articles had to include at least one of the following terms:

> masculine OR masculinity OR macho OR manhood OR manly OR machismo OR manliness OR maleness OR black men OR black man OR asian man OR asian men OR white men OR white man OR working class man OR working class men OR middle class man OR middle class men OR upper class men OR upper class man.

Additionally, a set of excluding terms were used in order to limit repeated articles due to some newspapers archiving second editions or regional editions of the same issue. The resulting corpus was 44.1 million words in size. The whole dataset was used for the corpus analysis, although clearly it would not be feasible for the close qualitative analysis to be carried out on such a large amount of text. Instead, a down-sampling method was used in order to reduce the number of articles to a workable amount. As our research question involved the representation of and discourses around six male identity groups, we first identified the articles which contained the most references to each group, for example, by locating the article which had most mentions of the terms *black man* and *black men* combined. This gave us six articles (one for each group), although the set was somewhat skewed towards broadsheet newspapers which tended to have longer articles. Therefore, in order to provide a more representative coverage from the corpus, we carried out the same exercise again, separately for each of the nine newspapers. This would have resulted in a down-sampled set of 54 articles. However, as *The Mirror*, *Sun* and *Star* did not mention upper-class men, the actual set came to 51 articles. We feel that this method of down-sampling produced a small set of salient articles where the particular identities we were interested in were likely to be foregrounded as a topic in themselves rather than mentioned 'in passing'. A question this raises for the comparison, however, is whether references to men as the main focus of an article will access different discourses to those where they are more fleetingly eluded to.

The two analysts agreed on a research question for their analyses, which was 'How are different types of men represented in the British press?'. They also agreed to produce an analysis consisting of around 3000 words, summarising the main patterns that they found. They then worked without any form of collaboration on their respective datasets for a period of three weeks. The resulting analyses were then exchanged and compared.

Methods of analysis

The corpus analysis was approached relatively 'naively' so hypotheses were not formed in advance of the analysis, nor were attempts made to link findings to existing theories or research on masculinity. Instead it involved using the corpus analysis software WordSmith Tools 5 to find the strongest collocates of the six identity groups within the entire 44.1 million word corpus. Collocates were obtained using the Dice Coefficient, an effect size statistic which is a measure of strength of association between two words. Singular and plural identities were considered separately (so, for example, collocates of both the phrases *Asian man* and *Asian men* were elicited in two separate searches), and the 20 collocates with the highest Dice score were collected for each search term. This resulted in potentially 40 collocates for each identity, although in a few cases collocates for the singular and plural identities were the same (e.g. *accused* collocates with both *black man* and *black men*). The collocational relationships were explored using concordances (tables which show all of the occurrences of a word, phrase or related pair of words in the immediate context that they occur in) in order to identify the most typical contexts that they occurred in. In many cases, concordance lines needed to be expanded in order to access more context, which sometimes involved reading an entire article. Collocates which contributed towards similar representations of masculinity were grouped together in order to indicate discourse prosodies where a group is frequently associated with a set of words that reference the same discourse. An example of this would be the way that black men were positioned as either suspected (with collocates like *accused*, *defends*, *custody*, *lawyer*, *DNA* and *database*), or actual criminals (with collocates like *raping*, *rape*, *prison*, *confessed*, *deaths* and *gang*).

The analysis also focused on whether discourse prosodies were unique to particular identities or shared between more than one group. Cases where a collocate occurred due to the repetition of a single quote across multiple articles on the same story were highlighted as potentially interesting but offering only weak evidence for the generalisabiltiy of a particular discourse. An example would be the collocate *inarticulate* which appears with *working-class man*. However, this pairing only ever occurred in a repeated quote across numerous articles that is attributed to the then Labour chairman Ian McCartney, who says 'When you're basically described, and the best way of paraphrasing, as an inarticulate working class man from Glasgow who's very liked but ain't that much good, you know it's a caricature too far'. The quote does indeed suggest that *inarticulate* is seen as a characteristic of being working class, but more evidence would be needed to support this.

For some of the less frequent terms, 20 collocates were not found. In particular, there were only five collocates for *upper-class man/men* and all of them were grammatical words like *of* and *the*. Therefore, rather than examining these collocates, concordance lines of all the cases of *upper-class man/men* were scrutinised (there were only 31 lines in total).

The qualitative analysis of the down-sampled set of articles was situated within a framework of stance-taking, or the act of linguistically evaluating a contextually relevant object, and, in so doing, positioning one's self (and others) in social space (Du Bois, 2007; Jaffe, 2009). Most commonly, stance-taking is viewed as an interactional phenomenon, something that people do in naturally occurring speech. In this analysis, however,

the argument put forward in Thurlow and Jaworski (2009) was followed, which states that written discourse, and particularly media texts, can also function as a stance-taking vehicle by virtue of its ability to instantiate an ideological framework both for the evaluation of social practice and for the association of those practices with pre-defined social categories and roles (see also Agha, 2007; Kress, 1995). In other words, in much the same way that variation in spoken language provides individuals with a means to adopt different footings with respect to the categories and characteristics referenced through talk, it is argued that the textual properties of written discourse serve to encode different socio-interpretive schema within which relevant social categories and category-affiliated activities are positioned and evaluated. Given this point of departure, the primary aim of the analysis of the down-sampled set is therefore to determine what ideological framework for masculinity the texts serve to instantiate, and to discover how the different masculine identities in question are positioned within this space.

This task was approached by asking three basic questions of the texts in the sample (cf. Du Bois, 2007; see also Bourdieu and Wacquant, 1992 on *field analysis*): (1) What is the stance object, or the particular category or trait being described and assessed, (2) how is that object evaluated (affectively, epistemically and deontically) and (3) how do these evaluations serve to position that object relationally in social space? Due to the design of the current study, the first question – while often highly relevant with interactional data – is relatively trivial in our sample. The down-sampled set is structured according to frequency of occurrence of the relevant search terms, and so each of these terms (e.g. black man/men, white man/men) were taken to be the respective stance objects. The qualitative investigation therefore focuses on resolving the two latter questions, which was accomplished by examining a range of both formal and semantic features, including clause modality, presupposition and implicature, and verbal argument structure. The analysis began with an investigation of discourses of 'masculinity' more generally, as opposed to texts that focus on any one specific masculine type. This allowed the identifications of the broader ideological topography (Bourdieu, 1979) of masculinity that exists in the sample as a whole. Then, the analysis turned to an examination of the specific masculine types, and a discussion of the ways in which they are discursively constructed as being more or less desirable, worthwhile and even morally sound than others.

Comparison of findings

Rather than providing the full analysis reports, in this section we carry out a meta-analysis which involves comparing and contrasting the two reports together. We have first divided our meta-analysis into a discussion of shared and unique findings, and then we move on to attempting to provide an explanation for the differences and similarities of the two approaches, as well as more critically evaluating the strengths and weaknesses of each, and the extent to which a combined approach would eliminate weaknesses. We should note that unlike Marchi and Taylor (2009), the comparison did not uncover any discordant findings (where we directly or indirectly disagreed), so all of our findings were either converging (shared) or complementary (different but contributing towards a wider picture).

Shared findings

As noted earlier, the corpus analysis found a discourse prosody of black men as suspected or actual criminals due to the presence of collocates like *accused*, *custody*, *rape*, *prison*, *confessed* and *gang*. Black men were also constructed as physically over-powering or impressive due to the use of collocate *tall*. The qualitative analysis noted a related construction of black men as violent and also indicated an over-abundance of physicality. Both analyses noted some common stereotypes around representations of black men, as well as the acknowledgement that black men suffered from discrimination.

For Asian men, constructions around sexual grooming were identified by both analysts, where such men were described as engaging in sexual and other forms of violent behaviour with under-age white women. The qualitative analysis noted 'In these texts, themes of violence are very clearly present, and include discussion of Asian Men's "raping," "beating" and "torturing" young women', while the corpus analysis pointed out that there were '42 cases where *groom/grooming* collocates with Asian men who are described as grooming (usually white) teenage girls for sex'.

The qualitative analyst found constructions of white men as 'beleaguered', noting that the majority of the texts described different situations in which white men were unfairly excluded from participation in an activity. Two of the texts, for example, discussed how the Avon Fire Service 'restricted' four of its five open days for potential new staff to women and ethnic minorities. In these texts, white men are described as being 'banned', 'shunned' and 'rejected'. Similarly, the corpus analysis identified how the collocate *applications* was used to suggest criticism of 'positive discrimination' against white men by describing cases where job applications by white men had been (sometimes illegally) rejected because of their race and gender. Both analysts therefore found evidence that white men were viewed as victims.

Additionally, both analysts noted working-class men being constructed in relation to disadvantage, with the corpus analyst writing 'The collocates *jobs* and *gone* suggest that working-class men are disadvantaged in comparison to other groups' and the qualitative analyst noting how 'Working-Class Men in the sample are also presented as somewhat "beleaguered"'.

Similar to white and working-class men, middle-class men were also described as beleaguered, with the qualitative analysis noting 'There are numerous lexical and semantic resonances between descriptions of white Men and descriptions of middle-Class Men, including such terms as (unfairly) "excluded" and "overlooked"'. The corpus analysis noted that such constructions, while present in the corpus, were relatively rare, focused around the collocate *discriminated*, and constituted a 'minority discourse'. But both analysts noted that white and middle-class men were represented in similar ways.

With upper-class men, the qualitative analysis noted they were represented as effete or unmanly, while a related construction was found by the corpus analysis (of concordance lines) which spotted 'an insinuation that they may be gay with one case referencing their "romantic friendships" while another refers to an actor who has played "rotund (and often homosexual) upper-class men"'. The qualitative analysis indicated a representation of upper-class men as lacking morality, while the corpus analysis focused on them as 'having bad manners or few manners'.

Unique to the qualitative analysis

The qualitative analyst noted a number of representations that were not identified by the corpus analyst. First, the qualitative analysis identified two major clines or factors which were viewed as relevant to representations in all six types of men examined: physicality and ambition. In terms of unique findings relating to individual identities, these included representations of black men as unambitious, destructive and lacking in self-control, as well as a representation of an 'anodyne black man' who was professionally successful and financially secure but was also depicted as a an inauthentic 'sell-out'. In terms of ambition, Asian men were seen as 'inherently entrepreneurial'. For example, one Asian man in the sample was viewed as a 'respectable entrepreneur'. However, some Asian men were viewed as over-ambitious, to the point that at an extreme end they were represented as achieving their ambitions through violent or immoral means. But the key trait uniting the different representations of Asian men was 'gritty pragmatism'.

The qualitative analysis thus noted two images of both Asian and black men (one ambitious, successful and acceptable, the other criminal, violent and deviant), while the corpus analysis only found evidence for the deviant identity. The qualitative analysis also noticed how journalists attempted to explain the deviant identities by focusing on their possible sources. For example, a reason for Asian men's criminal behaviour was linked to them being 'torn between two cultures', with the implication being that integration into British culture leads to respectability. Similarly, the reason why working-class men were described as 'beleaguered' was due to them being depicted as having been 'forgotten by society'.

Working-class men were also constructed as violent (indiscriminately so), due to a lack of education and 'moral anchoring', as well as lacking in ambition (similar to black men). However, upper-class men were also described in terms of a lack of ambition and ultimately having a deficient masculinity. The analyst points to middle-class and white men as existing at the moral centre of masculinity as an ideological space, with 'numerous semantic resonances' between these two identity groups.

Unique to the corpus analysis

The corpus analyst also noticed some representations that were not present in the other analysis, although these did not contradict anything found by the qualitative research. These included representations of black men as victims of two types of violent crime – historical lynchings in the United States and more recent stabbings in UK cities. Such men were also constructed as sexualised, particularly in relation to white women, with the collocate *women* being used to refer to them raping white women, dating them or being paid for sex by them. Asian men were also represented as victims of violence or racist abuse, but criminal Asian men were also constructed through the collocate *hunting*, which is normally associated with wild animals. Additionally, Asian men were sometimes identified by the collocate *beards*, particularly as an indicator of possible terrorism or other people's suspicion of terrorism. It was noted that white men are also described as hunted or wanted by the police, although there were no associations with particular *types* of crime with white men, unlike the constructions of black and Asian men. White men were described as dominating powerful institutions, engaging in historical cases of racism and potentially lacking in physicality compared to other ethnic groups.

A collocate related to working-class men was *self*, which occurred in descriptions of such men who were *self-made*, *self-educated*, *self-motivated* and so on. The implication being that success for such men is due to their own efforts. The collocate *I'm* was also used in autobiographical constructions, where there was a sense of (rueful) pride in disclosing a working-class status, and this was linked to the idea of working-class men having their own *values* (a collocate not found for any of the other groups). However, working-class men were also represented as having drab or tedious lives. Middle-class men collocated with *well-* in constructions like *well-known*, *well-educated* and *well-dressed*, suggesting a particular way that a discourse of privilege is realised. Unlike *self*, which makes agency very clear, a term such as *well-educated* does not directly attribute agency so it is often not clear who or what has caused these men to be well-educated. Finally, upper-class men were often found in historical contexts rather than being associated with a modern-day masculine identity. Table 1 summarises the shared and unique findings between the two types of analyses.

Discussion of the techniques

An obvious advantage of the corpus approach was in providing the analyst with a much larger data-set so that the analyst could claim fuller coverage of representations that were found. This helps to explain some of the findings that were unique to the corpus analysis, such as (mainly white) British journalists' seeming fascination with a range of different types of sexual relationships between black men and white women or the way that Asian men's beards are sometimes used as a signifier for possible terrorist intent.

The corpus analysis was particularly useful at identifying repetitive lexical combinations that indicated more subtle ideological representations and would have perhaps otherwise have been missed, even if researchers had access to the full dataset. Three examples of these include the association of *I'm* with *working-class man*, which indicated that of all the six identities, this was one where individuals who held it were most likely to 'claim' it, along with the association with working-class men and adjectives that were prefaced by *self-*. Ideologically, such marked expressions as *self-educated working-class man* imply perhaps that most working-class men are *not* educated, but also that those who are need to rely on their own resources and ingenuity to do so. Society then is implied to allow working-class men to succeed, although it must be on their own merits. Conversely, the collocate *well-*, appearing at the start of phrases like *well-educated* and *well-fed* with *middle-class man*, does not indicate who has bestowed good education or good food onto them. The collocate *well-* thus acknowledges the different sorts of privilege that middle-class men enjoy and also implies that such privilege is something that is awarded to them by an unmarked other.

A third advantage of the corpus method was in helping to give an indication of how frequently certain constructions were articulated. A representation which was just linked to a single collocate, and only appeared because that collocate was part of a single quote that occurred in multiple tellings of the same story (e.g. *inarticulate* with *working-class man*), means the analyst can describe the effect of such a collocate but also indicate its limitations towards generalisability. On the other hand, a set of collocates which have a similar meaning or discursive function and occur across numerous contexts, such as the collocates which link Black men to crime, enable the analyst to be more confident that

Table 1. Comparison of the two approaches.

	Shared	Qualitative	Corpus
Black men	Criminal/violent Physical Negative stereotypes Discriminated against	Anodyne black men Unambitious, destructive, lack of self-control	Victims of crime Focus on sexualised relations with white women
Asian men	Sexual grooming	(Overly) entrepreneurial Torn between two cultures Gritty pragmatism	Victims of crime and racism 'Hunted' Suspected terrorists
White men	Unfair exclusion	Moral centre of masculinity	Dominating institutions Lacking in physicality Hunted as criminals
Working-class men	Disadvantaged	Indiscriminately violent Lack of ambition Forgotten	Success due to self Value system Drab lives
Middle-class men	Discriminated against	Moral centre of masculinity	Privileged
Upper-class men	Effete Lacking morality	Lack of ambition	Historically realised

what she or he is seeing is a common representation or a hegemonic discourse, particularly if such representations go unchallenged in most of the texts. Louw (1993) refers to the way that a word has 'a consistent aura of meaning with which a form is imbued by its collocates' (p. 157) as instantiating a semantic prosody.

A weakness of a singular corpus approach to discourse analysis (compared to the qualitative analysis) is that the focus on collocates or other patterns based around word frequencies may mean that in some cases a purely descriptive analysis emerges which does not attempt to provide interpretation, critique or explanation for the patterns found. Nor may such analysis engage with the wider social and historical context beyond the corpus. With many collocates, it may be difficult for the analyst to fully account for them all by carefully reading expanded concordance lines for every case. In a worst-case scenario, the analyst might reach an erroneous conclusion, incorrectly assuming a collocate has a particular function when in fact it might be used in an unexpected way some or even most of the time. Similarly, a corpus approach focused on reading concordance lines may not be effective at identifying non-patterned uses of language such as legitimation strategies, topoi or intertextuality.

And in providing the analyst with so much frequency-based information, it can sometimes be difficult to separate the wheat from the chaff. Therefore, a corpus approach may

yield numerous 'so what' findings, where the frequency patterns simply confirm the expectations of people who are reasonably au fait with the society that the texts come from. For example, we were not surprised to find representations of Black men as violent or discriminated against (although such representations were also found in the qualitative analysis). It could be argued that such findings are useful in that at least they indicate the validity of the approach, but generally less value tends to be placed on unsurprising or 'already known' findings, and therefore corpus analysts who study discourse may need to be more selective in terms of what they focus on in their reports. Inevitably this can result in subjectivity, with the analyst needing to make decisions based on either what they personally find surprising, troubling or encouraging in the corpus from a critical perspective, or what they think their audience is most likely to be interested in.

The qualitative analysis shares some of the same potential pitfalls as the corpus approach. In qualitative analysis, it is certainly possible to engage in pure description of the themes and structural patterns identified in a text, producing what Antaki et al. (2003) term 'under-analysis through summary'. Likewise, qualitative analysts may also have a tendency to focus on well-established patterns that reflect popularly known societal discourses. The result of this could be the production of the so-called 'so what' findings. While these analytical dangers are certainly not restricted to qualitative studies, they may be exacerbated by the holistic approach of most qualitative work. Given the poetic and discursive complexity inherent in any interaction or text, it can be difficult to identify systematic patterns, let alone bring the 'analytic extra' (Antaki et al., 2003) that qualitative analysis requires. Yet it is precisely the qualitative approach's willingness to explore this complexity that enables the analyst to discover subtle, and perhaps unexpected, patterns of socially meaningful language use, and to situate those patterns within a broader social, historical and ideological context (cf. Rampton et al., 2004).

In terms of the current discussion, for example, the qualitative analysis was able to pick up on the existence of two complementary representations of both black and Asian men in the sample. While the corpus analysis and qualitative analysis converged in identifying the more frequent image of black and Asian men as 'violent' and 'criminal', the holistic qualitative approach discovered that these dominant representations were occasionally juxtaposed with alternative images of 'anodyne' and 'respectable' black and Asian men, respectively. For both black and Asian men, the characteristic that distinguishes the 'violent' representations from their more 'anodyne' and 'respectable' counterparts is the extent to which the former are portrayed as being overly and indiscriminately 'physical'. Physicality is thus shown to be a primary structuring dimension of the ideological field of masculinity, along which representations of different kinds of men are positioned and morally evaluated. This discovery, therefore, not only provides a fuller ethnographic picture of how both dominant and alternative masculine types are represented in the dataset, but it also helps us to understand the discourses and ideologies that organise representations of masculinity more generally.

A second strength of the broader analytical gaze of qualitative analysis is its ability to uncover the implicit representations that emerge. Many characteristics of upper-class men, for example, were not explicitly stated. Instead, they were dually indexed (Kulick, 2005) in the structure of the texts. A good illustration of this pattern can be found in a first-person article in the sample about why the author prefers her current middle-class boyfriend to the upper-class men she used to date. In that piece, the author states that her

current boyfriend is 'refreshingly direct', has 'unshakeable self-belief' and is 'a man of action'. Without stating it outright, the author thus constructs an image of upper-class men as being indirect, lacking in confidence and generally ineffectual. These traits are not necessarily ones that a corpus analysis would be able to pick up since they emerge via strategic omission. The qualitative analysis, in contrast, was able to identify how representations of masculinity are constructed both by what is said and by what is left out.

Finally, as the example above also demonstrates, qualitative analyses enable a more detailed examination of the structural properties of texts and the ways in which these properties serve particular discursive ends. Both the corpus and the qualitative analysis were able to identify dominant representations of black and white men, for instance, via an examination of patterns of lexical collocation and semantic resonance. The qualitative analysis, however, was also able to show that these representations were systematically located within grammatical constructions that serve to enhance their epistemic force. One common strategy for epistemic 'boosting' (Holmes, 1984) was the embedding of particular representations as presuppositions of clausal matrix verbs, thus making them resistant to negation (e.g. *I fear that [Black men] will never be able to regain control of their lives*, where the verb *regain* presupposes the assertion that Black men have already lost control). The effect of modalising tactics of this sort is to strengthen the authority of the representation that emerges in the text, and, as a result, to both reinforce and reproduce the dominant ideologies upon which these representations are based.

Overall then, the detailed scrutiny of a smaller dataset in the qualitative analysis is able to identify more subtle social and linguistic patterns in the texts and to situate its interpretations of these patterns within a multi-level understanding of the broader ideological context. A potential shortcoming of this kind of approach, however, is that the findings identified in this way may not be generalisable beyond the smaller dataset considered. In addition, and more so perhaps than in the corpus approach, the qualitative analysis runs the risk of being overly subjective, with pattern identification entirely a property of the perspective of the analyst and not accountable to any more 'objective' method. For this reason, it is useful to complement qualitative analyses with corpus approaches as we have done here, since the sociolinguistic empiricism (Woolard, 1985) of corpus methods helps to 'tie down' (Rampton et al., 2004) the qualitative interpretations, thus enhancing their overall reliability and validity.

Conclusion

Reassuringly, both approaches to the analysis uncovered a set of shared findings. These tended to be based around frequently articulated representations of the various groups – they were frequent enough to emerge both from a corpus analysis which tends to prioritise frequency and/or saliency, as well as being very likely to appear in the down-sampled dataset due to their high probability of occurrence. We could be reasonably confident that as long as the down-sampling is carried out sensitively, either method is likely to elicit dominant discourses (at least in terms of the frequency of their mention) around a topic.

The fact that neither approach yielded any contradictory findings is also reassuring, although we cannot claim that this state of affairs would be replicated had we used a different down-sampled set or even used different analysts. However, the existence of

complementary findings perhaps best indicates the strengths and weaknesses of the two approaches, with both uncovering more subtle representations in different ways. The corpus approach was able to pick up on repeated lexical pairings across very large amounts of text which pointed to a set of representations that were also very frequent (but did not appear in the down-sampled set), while the qualitative analysis was able to uncover discourses that were linguistically realised in more complex ways, for example, more easily being able to uncover how authors attempted to explain or imply certain representations. The qualitative analysis also picked up more readily on a set of findings that were not necessarily frequent across the corpus, although an issue here is that based on a small set, without relying on other forms of information, it is difficult to make strong claims about how typical a finding is from a small sample.

While further experiments of this type would be welcome in order to shed more light on the benefits and drawbacks associated with different approaches to CDA, we would conclude that this study shows a clear advantage to a triangulatory approach. As an alternative to the model proposed by Baker et al. (2008), which attempted to synthesise both methods into a nine-stage model, we would also argue that the two approaches could function well as separate components carried out by different experts in their field, with a final stage of the analysis involving a comparison and consolidation of research findings. We believe that such an approach would encourage collaboration between researchers with different skill sets, yet also allowing expertise to be put to best use, rather than requiring researchers to become discourse analytical polymaths.

Funding

The research presented in this paper was supported by the Economic and Social Research Council (ESRC) Centre for Corpus Approaches to Social Science, ESRC grant reference ES/K002155/1

Notes

1. While this article summarises the two sets of results, we focus more on a critical discussion of the methodologies used, with the aim of publishing a separate paper based around representations of masculinity in the press.
2. The newspapers collected were *The Express*, *The Guardian*, *The Independent*, *The Mail*, *The Mirror*, *The Star*, *The Sun*, *The Telegraph* and *The Times*.

References

Agha A (2007) *Language and Social Relations*. Cambridge: Cambridge University Press.

Antaki C, Billig M, Edwards D, et al. (2003) Discourse analysis means doing analysis: A critique of six analytic shortcomings. *Discourse Analysis Online*. Available at: http://extra.shu.ac.uk/daol/articles/open/2002/002/antaki2002002-paper.html (accessed 19 January 2015).

Baker P (forthcoming) Does Britain need any more foreign doctors? Inter-analyst consistency and corpus-assisted (critical) discourse analysis. In: Charles M, Groom N and John S (eds) *Grammar, Text and Discourse: In Honour of Susan Hunston*. Amsterdam/Philadelphia, PA: John Benjamins.

Baker P, Gabrielatos C, Khosravinik M, et al. (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 & Society* 19(3): 273–306.

Bourdieu P (1979) *Distinction*. Cambridge, MA: Harvard University Press.

Bourdieu P and Wacquant L (1992) *An Invitation to Reflexive Sociology*. Cambridge: Polity Press.

Caldas-Coulthard CR (1993) From discourse analysis to critical discourse analysis: The differential re-presentation of women and men speaking in written news. In: Sinclair JM, Hoey M and Fox G (eds) *Techniques of Description*. London: Routledge, pp. 196–208.

Caldas-Coulthard CR (1995) Man in the news: The misrepresentation of women speaking in news-as-narrative-discourse. In: Mills S (ed.) *Language and Gender: Interdisciplinary Perspectives*. Harlow: Longman, pp. 226–239.

Cicourel A (1969) *Method and Measurement in Sociology*. New York: The Free Press.

Denzin NK (1970) *The Research Act: A Theoretical Introduction to Sociological Methods*. Chicago, IL: Aldine.

Du Bois J (2007) The stance triangle. In: Englebretson R (ed.) *Stancetaking in Discourse*. Amsterdam: John Benjamins, pp. 139–182.

Erzberger C and Prein G (1997) Triangulation: Validity and empirically-based hypothesis construction. *Quality & Quantity* 31: 141–154.

Flowerdew J (1997) The discourse of colonial withdrawal: A case study in the creation of mythic discourse. *Discourse & Society* 8: 453–477.

Gabrielatos C and Baker P (2008) Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press, 1996–2005. *Journal of English Linguistics* 36(1): 5–38.

Hardt-Mautner G (1995) *Only connect. Critical discourse analysis and corpus linguistics*. UCREL technical paper 6. Lancaster: Lancaster University. Available at: http://ucrel.lancs.ac.uk/papers/techpaper/vol6.pdf (accessed 19 January 2015).

Holmes J (1984) Modifying illocutionary force. *Journal of Pragmatics* 8: 345–365.

Jaffe A (ed.) (2009) *Stance: Sociolinguistic Perspectives*. Oxford: Oxford University Press.

KhosraviNik M (2010) The representation of refugees, asylum seekers and immigrants in British newspapers: A critical discourse analysis. *Journal of Language and Politics* 9(1): 1–28.

Kress G (1995) The social production of language: History and structures of domination. In: Fries P and Gregory M (eds) *Discourse in Society: Systemic Functional Perspectives. Meaning and Choice in Language: Studies for Michael Halliday*. Westport: Ablex Publishing, pp. 115–140.

Krishnamurthy R (1996) Ethnic, racial and tribal: The language of racism? In: Caldas-Coulthard CR and Coulthard M (eds) *Texts and Practices: Readings in Critical Discourse Analysis*. London: Routledge, pp. 129–149.

Kulick D (2005) The importance of what gets left out. *Discourse Studies* 7: 615–624.

Louw B (1993) Irony in the text or insincerity in the writer? In: Baker M, Francis G and Tognini-Bonelli E (eds) *Text and Technology: In Honour of John Sinclair*. Philadelphia, PA/Amsterdam: John Benjamins, pp. 157–176.

Marchi A and Taylor C (2009) 'If on a winter's night two researchers … A challenge to assumptions of soundness of interpretation. *CADAAD Journal* 3(1): 1–20.

Partington A, Duguid A and Taylor C (2013) *Patterns and Meanings in Discourse: Theory and Practice in Corpus-assisted Discourse Studies (CADS)*. Amsterdam: John Benjamins.

Rampton B, Tusting K, Maybin J, et al. (2004) UK linguistic ethnography: A discussion paper. Available at: http://www.lancaster.ac.uk/fss/organisations/lingethn/documents/discussion_paper_jan_05.pdf (accessed 19 January 2015).

Reisgl M and Wodak R (2001) *Discourse and Discrimination: Rhetorics of Racism and Antisemitism*. London: Routledge.

Rothbauer P (2008) Triangulation. In: Given L (ed.) *The SAGE Encyclopedia of Qualitative Research Methods*. Thousand Oaks, CA: SAGE, pp. 892–894.

Thurlow C and Jaworski A (2009) Taking an elitist stance: Ideology and the social production of distinction. In: Jaffe A (ed.) *Stance: Sociolinguistic Perspectives*. Oxford: Oxford University Press, pp. 195–226.

Van Dijk T (2006) Ideology and discourse analysis. *Journal of Political Ideologies* 11(2): 115–140.

Widdowson HG (2000) On the limitations of linguistics applied. *Applied Linguistics* 21(1): 3–25.

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

Wodak R (2007) Pragmatics and critical discourse analysis. *Pragmatics and Cognition* 15(1): 203–225.

Woolard K (1985) Language variation and cultural hegemony: Toward an integration of sociolinguistic and social theory. *American Ethnologist* 12: 738–748.

Author biographies

Paul Baker is Professor of English Language at Lancaster University. His research involves applications of corpus linguistics, and his recent books include *Using Corpora to Analyze Gender* (2014), *Discourse Analysis and Media Attitudes* (2013) and *Sociolinguistics and Corpus Linguistics* (2010). He is the commissioning editor of the journal *Corpora*.

Erez Levon is Senior Lecturer in Lingustics at Queen Mary University of London. His research interests include sociolinguistics, language, and gender/sexuality and intersectionality. His recent publications include the book *Language and the Politics of Sexuality* (2010) and the edited collection *Language, Sexuality and Stance* (forthcoming with Ronald Mendes).