The utility of topic modelling for discourse studies: A critical evaluation

Gavin Brookes
Lancaster University, UK

Tony McEnery
Lancaster University, UK

Abstract
This article explores and critically evaluates the potential contribution to discourse studies of topic modelling, a group of machine learning methods which have been used with the aim of automatically discovering thematic information in large collections of texts. We critically evaluate the utility of the thematic grouping of texts into ‘topics’ emerging from a large collection of online patient comments about the National Health Service (NHS) in England. We take two approaches to this, one inspired by methods adopted in existing topic modelling research and the other using more established methods of discourse analysis. In the study, we compare the insights produced by each approach and consider the extent to which the automatically generated topics might be of use to discourse analysts attempting to organise and study sizeable datasets. We found that the topic modelling approach was able to group texts into ‘topics’ that were truly thematically coherent with a mixed degree of success, while the more traditional approach to discourse analysis consistently provided a more nuanced perspective on the data which was ultimately closer to the ‘reality’ of the texts it contains. This study thus highlights issues concerning the use of topic modelling and offers recommendations and caveats to researchers employing such approaches to studying discourse in the future.

Keywords
Corpus linguistics, corpus-assisted discourse studies, latent Dirichlet allocation, patient feedback, topic modelling

Corresponding author:
Gavin Brookes, CASS, FASS Building, Lancaster University, Lancaster LA1 4YW, Lancashire, UK.
Email: g.j.brookes@lancaster.ac.uk
Introduction

The emergence of any technique of data collection, storage or analysis poses important questions about the extent to which that technique might supplement or even replace existing techniques in a given field (Baker et al., 2008). This article sets out to answer such questions with regard to topic modelling by critically evaluating its utility for discourse studies. Focusing on one popular type of topic modelling, latent Dirichlet allocation (LDA; Blei et al., 2003), we analyse the ‘topics’ emerging from a large database of online patient feedback about healthcare services in England. We then analyse the same data using methods typical in discourse analysis and compare the findings obtained using each approach. Based on this, we assess the usefulness of topic modelling for providing insights that would be of value for studying discourse in large collections of texts, a use for topic modelling that has been proposed by some (e.g. Törnberg and Törnberg, 2016; Underwood, 2012). This evaluation is carried out with a focus upon the viability of developing approaches that combine topic modelling techniques with more established methods in linguistics, such as discourse analysis and corpus linguistics, to produce more complete, rigorous and insightful accounts of textual data. The following section provides a critical introduction to topic modelling. We then outline the data and methodological procedures used in this article. The findings of both analyses are then reported, compared and critically evaluated. We conclude the article by evaluating the overall utility of topic modelling for discourse studies and by offering recommendations and caveats to discourse analysts planning to adopt such methods in the future.

Topic modelling and LDA

Since our aim here is to evaluate topic modelling from the perspective of the discourse analyst, we will not concern ourselves overly with the computational and statistical processes that it involves, as these are well documented elsewhere (see Blei, 2012; Blei et al., 2003; Murakami et al., 2017). LDA is a form of topic modelling that seeks to automatically discover thematically coherent ‘topics’ within a large collection of texts. Topics are extracted on the basis of word co-occurrence. Once a collection of texts has been loaded into the computer, specialist topic modelling software searches for patterns of word co-occurrence within individual texts. Where a group of words tend to co-occur within the same text with a degree of regularity, the computer assumes there to be a relationship between those words. For example, if a statistical procedure determines that the words once and end often feature within the same text in a corpus, then a relationship between them is assumed by the computer. Unlike corpus linguistic measures of collocation, which tend to operate with relatively tight spans of co-occurrence, often of between three and five words (Baker, 2006: 95), topic modelling usually accounts for co-occurrence on a broader scale, typically across an entire text. Therefore, in topic modelling it is possible for the computer to assume a relationship between any words that frequently co-occur within the same text, even if they are consistently thousands of words apart. Once identified, a co-occurring group of words is then taken by the computer to represent a thematically coherent topic.

The principle output of the topic modelling process is a series of lists of words which co-occur within the collection of texts in the corpus being analysed. Each word-list, in
principle, represents a discrete topic and the words contained within it are the most characteristic words of that topic. The topics output from a topic model span the texts in the corpus – each text may be strongly associated with many, few or no topics. For each text, a topic model outputs a set of results which show, by means of a score, the strength of each topic in each text in the corpus. For the purposes of study, then, the words displayed in a topic word-list may be viewed as ‘key words’ of that topic, while texts with a high probability of exhibiting a topic might be viewed as ‘key texts’ of the topic (Murakami et al., 2017: 246). Although the topics are generated algorithmically by computer software, it is up to human users/researchers to assign meanings to those topics and infer their thematic coherence once they have been generated by the computer.

Yet before the computer can ‘discover’ topics in the data, the user is required to set two main parameters: (1) the number of topics to be extracted and (2) the number of words that must be in each topic. The number of topics provided by the computer influences the nature of the topics generated; a large number of topics holds the promise of providing more granular themes, while a smaller number of topics is claimed to provide more general themes (Murakami et al., 2017: 245). Topic modelling is a non-deterministic method; the topics generated by the computer are subject to some degree of randomness and may be different each time the procedure is run, even if the data and user-specified parameters remain constant.

Topic modelling is claimed to be a flexible approach that can be applied, in theory, to a collection of any type of text. Topic modelling has been used to study the themes in an ever-widening range of text types, such as political press releases (Grimmer, 2010), online fora (Törnberg and Törnberg, 2016) and academic research articles (Murakami et al., 2017).

Given that the success of any corpus-assisted study of discourse rests on the interplay of computational and statistical measures with human-led, theory-sensitive interpretation (Brezina, 2018), topic modelling, blending as it does an automated analysis with human-supervised interpretation, could in principle have a useful role to play in corpus-assisted discourse analysis. If we accept the claim that the topics generated are thematically coherent, topic modelling may provide a means of showing what a given collection of texts is ‘about’ (Blei, 2012), offering a potential alternative to established corpus linguistic techniques such as keywords and semantic tagging (Jaworska and Nanda, 2016), both of which could be said to successfully identify the ‘aboutness’ of a corpus (Murakami et al., 2017: 244). These apparent overlaps, between topic modelling and corpus linguistics, might imply the suitability of the former to supplement or even replace, on occasion, the latter. However, other methodological particularities of topic modelling provide cause to question its usefulness for the study of discourse. Broadly speaking, these concerns pertain to issues regarding (a) linguistic sensitivity and (b) replicability. We will consider both of these issues now.

In terms of linguistic sensitivity, the first feature worth considering is that topic modelling uses a very naïve model of a text. A text is viewed as being composed of a collection of words, a simple unordered set. The ordering of the words in those texts is irrelevant, except in so far as words co-occur within a text. However, the treatment of words as independent of their original order decouples those words from their syntactic and grammatical contexts of use. While this contextual stripping does not present much of an issue for extracting word frequency information, it disregards words’ contexts of use (including order of occurrence) which can hold clues about meaning (Van Valin and LaPolla, 1997).
Also problematic from the point of view of discourse analysis is the common use within topic modelling approaches of stop-word lists in order to exclude closed class words from the analysis. However, such grammatical or ‘functional’ items often hold clues as to words’ meanings and facilitate text coherence, and can thus have an important role to play in the exploration of discourse (Van Dijk, 1977). Thus, when we remove grammatical ‘noise’, we run the risk of removing with it important information that contextualises natural language, creates meaning and resolves lexical ambiguity.

Also common in topic modelling is the process of so-called stemming, or lemmatisation (e.g. Törnberg and Törnberg, 2016) – that is, reducing morphological variants of words to their common base form. Although it reduces problems of data sparsity for computational modelling, the approach is inimical to discourse analysis and ignores long-standing findings from corpus linguistics that show that morphologically distinct forms often have quite distinct meanings and discourse; for example, McEnery et al. (2015) found the words *muslim* and *muslims* to consistently index different discourses about Muslims. Collapsing the frequency counts of both words together clearly risks at the very least blurring an important distinction and, at worst, losing or mischaracterising it.

Given the absence of any theoretical account of what constitutes a ‘topic’ within topic modelling, it is hardly surprising that from a linguist’s perspective the approach appears problematic. While Blei et al. (2003) and others have asserted the ‘thematic coherence’ of the topics produced by topic modelling methods, concepts like ‘theme’ and even ‘coherence’ are vague and remain ill-defined within this body of research. What constitutes a theme and what it is claimed makes that theme coherent are unclear. The topic model studies we have referenced so far appear to regard the results produced by topic modelling methods as propositional topics. In natural language, propositional topics can be expressed using a variety of words that can belong to different semantic domains and can be combined in a variety of ways. Likewise, any single word can be used to express an innumerable array of propositions (Van Dijk, 1977). However, the lack of any (linguistic) theoretical grounding concerning the definition and conceptualisation of topics as understood within topic modelling means that topic discovery procedures risk the (as we have seen, false) assumption that a propositional topic always corresponds neatly to a particular word or semantic set.

Building on the notion that propositional topics cannot be reduced to any semantic set or other collection of words, Van Dijk (1980) argues that propositional topics, broadly conceived, should be interpreted with due consideration of the macro-structural level of discourse – that is, of how topics are established not by individual words alone, but across chains of sentences, paragraphs and texts. As Van Dijk (1977) puts it,

> a notion such as topic of discourse cannot simply be explained in terms of semantic relations between successive sentences. Rather, each of the sentences may contribute one ‘element’ such that a certain structure of these elements defines the topic of that sequence in much the same way as, at the syntactic level, words can be assigned a syntactic function only with respect to a structure ‘covering’ the whole clause or sentence. (p. 6)

Thus, Van Dijk’s approach favours inferring propositional topics in view of the broader (at least textual) contexts in and across which those topics are established linguistically.
However, in many existing topic model studies, topics are inferred by researchers without actually inspecting the texts assigned to that topic (at least in any systematic way) in order to confirm or revise original inferences (Blei and Lafferty, 2009; Grimmer, 2010; Zhao et al., 2011). Thus, as a set of methods that ‘discover’ topics on the basis of individual, unordered and ultimately de-contextualised words, we might have cause to question the extent to which topic modelling is able to facilitate such an analysis. Murakami et al. (2017) rightly caution against the inference of topics from this list of words alone. Having described the term ‘topic’ (as in ‘topic modelling’) as a ‘mismomer’, they argue that ‘[t]hese groups of co-occurring words characterise “topics”, and researchers may choose to refer to them using topic-like titles, but these are only convenient abstractions from lists of words’ (Murakami et al., 2017: 244; see also Jaworska and Nanda, 2016). These abstractions have a proposed – not axiomatic – utility, and hence they may prove to be far from convenient for discourse studies.

On a practical level, even if we set aside the issues discussed so far, we are still faced with problems pertaining to replicability. One of the major contributions of corpus linguistics is that it offers a more objective approach to analysing text, underpinned by the principle of total accountability (Leech, 1992: 112), and its findings may be falsifiable (Popper, [1935] 2006). Replicability is important here, since the findings of any study can only be tested and falsified if that study can be replicated.

Total accountability and falsifiability present a way for discourse analysts to avoid the charge of cherry-picking examples to prove a particular argument (Widdowson, 2004). Similarly, topic modelling appears objective and data-driven in as much as the topic discovery procedure is automatic and there seems to be limited scope for the human user to impose any extra-data or pre-analytical categories onto the topics generated. However, as we pointed out earlier in this section, the human researcher is still required to make an important decision about the number of topics generated. This decision is relatively arbitrary. Some mathematical rigour may be applied when selecting the number of topics (see e.g. Zhao et al., 2015), but the discovery procedure in topic modelling seems guided more often than not by a ‘try and see’ approach. As Zhao et al. (2015) note, ‘subjective evaluations are needed to compare models’ and there is ‘no easy way to choose the proper number of topics in a model beyond a major iterative approach’ (p. 1). Faced with the difficulty of finding the ‘right’ number of topics to study, Zhao et al. (2015) note that researchers ‘have no recourse beyond an informed guess or time-consuming trial and error evaluation’ (p. 2).

Total accountability and replicability become yet more problematic at the interpretative phase of the analysis where, as mentioned earlier, the human researcher is required to interpret topics from the word-lists provided by the computer. This phase has tended to be quite poorly specified within existing research and seems to involve human users/researchers inferring topics merely by ‘eye-balling’ the topic word-lists, without recourse to any standard or framework. For example, consider the following description of the topic discovery procedure employed by Hall et al. (2008):

We first ran LDA with 100 topics, and took 36 that we found to be relevant. We then hand-selected seed words for 10 more topics to improve coverage of the field. These 46 topics were then used as priors to a new 100-topic run. The top ten most frequent words for 43 of the topics
along with hand-assigned labels are provided. Topics deriving from manual seeds are marked with an asterisk. (pp. 364–365)

As a discovery procedure, this leaves much to be desired. There is an apparent objective procedure, topic modelling, but that is shot through with subjective decisions – the choice of 36 relevant topics, the discarding of 64 topics deemed irrelevant, the selection of seed words, the decision to generate 10 additional topics, the decision to run the topic model procedure again, the narrowing down of the study to the top 10 most frequent words in 43 topics, and the abandonment of 57 other topics and everything but the top 10 words in the 43 topics selected. The article offers no justification for any of these decisions and no insights into the considerations that underpinned them. While the article relies on many equations and technical terms to establish its credibility, these points relating to the discovery procedure are passed over in favour of a discussion of what appear to be plausible topics. This clearly leaves scope for Widdowson’s cherry-picking objection to be levelled at the work.

Even if we are transparent about our procedural decisions, the replicability of any topic modelling study is undermined by virtue of the method being non-deterministic; the topics generated by the computer in topic modelling are subject to some degree of randomness and may be different every time the procedure is run, even if the data and user-specified parameters remain constant. This means that, in contrast to methods typically used in corpus-assisted discourse analysis, the replication – or more specifically, the repetition – of a topic model study is not necessarily possible. Consequently, a user may modify and re-run the topic modelling many times, evaluating outputs until finding an analysis that they deem to be credible and usable. This introduces the possibility of a high degree of subjectivity into the analysis that would run directly up against Widdowson’s cherry-picking objection.

Despite its growing popularity in other fields, to our knowledge, topic modelling has yet to be critically evaluated in the context of its application to discourse studies. Yet there is value in human analysts using any method (including topic modelling) to test its asserted utility and to be cognisant of its limitations as well as its assumed strengths. In this article, we aim to provide this kind of insight with respect to the utility of topic modelling for discourse studies. Specifically, we seek to answer two questions: (1) How reliable an indicator are the topic word-lists alone of the themes in the texts assigned to that topic?; (2) How thematically coherent were the topics generated by the computer? The next section introduces the data and analytical approaches used to carry out this evaluation.

**Methodology**

**Data**

This study is based on a corpus of online patient feedback about the National Health Service (NHS) in England: 228,113 comments (approximately 29 million words) posted to the NHS Choices online service between March 2013 and September 2015. The comments exhibit clear thematic topics and are amenable to relatively simple classification.
On average, the comments were short (mean length: 128 words). This may be an issue to consider in the analysis, as some authors have claimed that the topic modelling technique we are using is more prone to error when texts are less than 1000 words in length (Törnberg and Törnberg, 2016: 410). Yet most experimental findings that discuss the problems that short texts produce for LDA analysis focus on Twitter data, where the messages are much shorter (maximum 140 characters). Also, even with Twitter data, it may be claimed that over half of the topics produced by LDA topic modelling appear coherent (Qiang et al., 2016: 8). Assuming that the longer length of our patient feedback comments (relative to tweets) should therefore enhance the performance of LDA topic modelling (Mazarura et al., 2014), our data should be amenable to demonstrating any potential usefulness of this method for discourse studies.

Analytical procedure

We decided to use an off-the-shelf topic modelling program – Mallet (McCallum, 2002), as this is a relatively easy-to-use package which has been adopted by linguists (e.g. Jaworska and Nanda, 2016). The corpus was processed by Mallet, which was instructed to produce 20 topics each composed of 20 words. We chose 20 topics because it is a common choice in the literature (see e.g. Ghayad et al., 2016). Also we were mindful of Zhao et al.’s study, which seemed to show that models using around 20 topics were successful for short message data (in their case, social media data) that might be viewed as somewhat analogous to our own. To reflect standard practice in topic modelling, we imposed a stop-word list to exclude closed class words from the analysis. However, we did not engage in lemmatisation in this article, partly because we do not think it is a helpful process for discourse analysis, but also because our goal was to see what a discourse analysis, carried out by a discourse analyst using Mallet, which does not perform stemming or lemmatisation, may look like. We ran the procedure once, accepting that output as the one to be used, and proceeded with our analysis from that point.

Our analytical procedure broadly mirrors the approach taken by Gkotsis et al. (2017), who were also looking at online health-related texts and had the advantage of having a prior qualitative analysis to use when initially classifying their topics. After that initial classification, they went on to explore the topics in question in some depth. Gkotsis et al.’s approach allowed us to explore two approaches to using topic model data: (1) a naïve model, where topics are interpreted largely on the basis of an analyst who may be familiar with a topic area, interpreting what the topic means with limited recourse to close reading, and (2) a blended approach, in which the topic model is evaluated with reference to close reading of down-sampled material. We took the relatively modest results of the Gkotsis et al. study as an indicator of the potential usefulness of the topic model approach. They reported topic allocation to be 71.37% accurate (p. 1).

Accordingly, our analysis proceeded in two phases. Phase 1, the naïve approach, reflected a commonplace approach to topic modelling; indeed, it may even be used as the basis for the evaluation of competing topic modelling algorithms (see e.g. Qiang et al., 2016: 8). While this phase focused on the topic word-lists only, the analysts were guided, as was the case in the study by Gkotsis et al. (2017), by our prior analysis of the texts using other methods (e.g. Brookes and Baker, 2017). Phase 2, based on a more typical
discourse analysis approach, focused, for each topic, on the 20 texts that exhibited the highest proportion of that topic as judged by the topic modelling algorithm. This analysis was qualitative, focused upon close reading, and was structured around the identification of recurring social actors, processes and places (Fairclough, 2003). Throughout this phase of the analysis, we were also mindful of the potential for patient feedback – as a register that might be characterised as uninterrupted and evaluative story-telling – to exhibit the types of narrative components of discourse identified by Labov (1972). Thus, we were sensitive to the possibility for the texts in our data to contain one or more of the following narrative components: abstract (what the comment is about), orientation (who, when, what, where?), complicating action (then what happened?), evaluation (so what?), result or resolution (what finally happened?) and coda (how do the events relate to the present time and place?) (p. 370; see also Labov and Waletzky, 1997). Although we did not carry out a full narrative analysis on these texts, our awareness of these narrative mechanisms and their potential discourse functions aided our interpretation of the topics and the evaluations in the data. This was a particularly important consideration, given the afore mentioned tendency for propositional topics to be established not in a single word or sentence but across entire texts (Van Dijk, 1977).

The texts were analysed independently in both phase 1 and phase 2 by each author of this article. After phase 1, we compared our analyses, which were similar, and agreed upon a set of interpretations of the topics for the data that we thought were plausible. We then independently analysed the texts of phase 2 and came together once more to compare and agree findings. We concluded by comparing our findings from phase 1 and phase 2 in order to ascertain the accuracy of our initial impressions based on the topics’ word-lists. The results of the analysis are presented in the next section.

Results

This section is divided into two parts. The first part reports the results of phase 1 of the analysis, outlining the 20 most characteristic words of each topic before briefly describing our interpretation of each topic based on our inspection of these lists of words. The latter half of this section then reports the results of phase 2, in which we qualitatively examined the texts assigned to each topic to either confirm or revise our initial interpretations from phase 1.

Phase 1

For ease of presentation, the results of phase 1 are displayed in Table 1. This table focuses on how interpretable we found the topics – hence the brevity of it. We were able, based on the topic models presented, to agree on a topic that the topic model represented. As users, our input in this process was to dictate the number of topics (20) and number of words in each topic word-list (20) generated by the program, and then to infer what the topics are ‘about’ based solely on eyeballing the words provided in the central column.

The process of looking at the topics seemed relatively straightforward – the words composing each topic, when considered in the cultural context of the United Kingdom
# Table 1. Results of phase 1 of the analysis.

| Topic | Word-list | Interpretation of the comments |
|-------|-----------|---------------------------------|
| 1     | treatment, clinic, department, dept, scan, referred, x-ray, outpatient, breast, attended, process, urology, excellent, appointment, royal, dermatology, bladder, ultrasound, today, november | People are happy with a range of NHS services |
| 2     | royal, leave, hospital, insulin, site, hillingdon, infection, eventually, managed, gran, control, entrance, security, knew, diabetic, walk, found, sunderland, department, vernon | Diabetic patients who have experienced difficulty finding their way somewhere within a number of hospitals |
| 3     | blood, test, heart, tests, results, pressure, chest, absolutely, feel, carried, cardiac, give, pains, private, fact, check, arm, offered, monday, wouldn’t | Chest pain and heart attacks |
| 4     | care, staff, hospital, ward, surgery, team, day, nurses, excellent, received, stay, treated, unit, admitted, experience, treatment, doctors, nursing, operation, great | Providing positive feedback |
| 5     | staff, service, time, made, good, feel, friendly, department, helpful, caring, hospital, kind, treatment, visit, member, procedure, felt, consultant, extremely, professional | Satisfied patients providing positive feedback |
| 6     | wife, complaint, complaints, pals, find, attitude, man, patient, absolute, system, reason, catheter, written, confirm, year, husband, show, no-one, supervisor, knew | Partners of patients complain about staff |
| 7     | eye, foot, leg, eyes, infection, vision, cataract, clinic, mile, referred, problem, painful, problems, discomfort, cast, casualty, drops, thing, moorfields, plaster | Complaints about a number of services |
| 8     | cancer, weeks, results, procedure, kidney, treatment, surgeon, told, radiotherapy, due, oncology, time, diagnosis, biopsy, piles, oncologist, diagnosed, chance, chemo, ward | Cancer treatment |
| 9     | appointment, weeks, hospital, told, clinic, wait, letter, back, appointments, phone, months, consultant, time, waiting, received, times, date, number, cancelled, department | Complaints regarding cancellations and waiting times |
| 10    | baby, birth, midwife, maternity, midwives, daughter, pregnant, labour, delivery, student, pregnancy, wife, weeks, born, child, early, miscarriage, baby’s, epsom, awful | Maternity services |
| 11    | doctor, nurse, time, hospital, hours, told, & waiting, home, back, asked, people, room, left, morning, didn’t, pain, arrived, long, called | Communication and the management of pain |
| 12    | bed, patients, ward, mother, night, lack, nurse, floor, infection, toilet, eventually, change, beds, assessment, bay, stop, legs, fall, opposite, management | Vague, though likely about people who have limited mobility or are unsteady |
| 13    | amp, parking, car, park, hospital, pay, access, i’d, disabled, find, time, area, local, ticket, minutes, money, travel, hours, massive, paid | Car parking, including the related issues of access to disabled spaces and payment |

(Continued)
Table 1. (Continued)

| Topic | Word-list | Interpretation of the comments |
|-------|-----------|-------------------------------|
| 14    | trust, nhs, services, health, positive, centre, hard, feedback, work, condition, person, people, patients, management, working, completely, group, wards, acute, individual | Feeding back in the NHS |
| 15    | eliot, george, steve, block, nurses, week, nerve, visitors, steve’s, locum, restaurant, university, surgeons, trainee, britten, bottles, Leeds, endoscopies, curtains, controlled | Broadly related to endoscopies |
| 16    | ward, operation, knee, surgeon, replacement, hip, food, weeks, recovery, total, deserve, physio, orthopaedic, admitted, walking, wife, bed, feb, queen, happy | Knee and hip replacements |
| 17    | son, daughter, children, child, childrens, year, children’s, mental, dental, can’t, don’t, boy, disgusting, vomiting, details, play, parents, teeth, queens, dentist b | Children’s dentistry |
| 18    | care, ward, patients, patient, mother, mum, poor, father, nurses, dad, family, elderly, admitted, time, life, it’s, year, communication, days, levels | Poor communication in relation to the elderly |
| 19    | patients, patient, scan, asked, sat, communication, lack, clear, common, special, sonographer, sense, southend, offered, syndrome, sick, understand, results, front, complete | Poor communication in relation to scans |
| 20    | pain, life, relief, medication, sleep, symptoms, side, experience, worse, caused, left, prescribed, years, offered, agony, lot, severe, skin, injection, advice | The experience of pain |

NHS: National Health Service.

*aNote that the word ‘amp’ is short for ampersand – the system from which the corpus was gathered uses this ‘amp’ instead of the ampersand character. The ampersand character has been present, it may have been deleted by the stop-word list.

*Note that the enclitic ‘n’t is not filtered out by the stop-word procedure.

and taking into account what we knew about the texts, seem to be semantically transparent. It was easy to read into each topic an entirely plausible analysis that could be helpful for a discourse analysis of the dataset, though some topics did appear to be mixed, since not all of the words in the topic seemed to contribute fully to the apparent meaning of the topic. There was also some degree of overlap between the topics; for example, topic 1 and topic 2 both focused on – admittedly apparently different – hospital departments. The general coherence in our topics is a marked improvement on the findings of Qiang et al. (2016) who, also using LDA topic modelling but with Twitter data, found that only just over half of the topics they produced seemed coherent, that is, could be said to have an apparent meaning when looked at by a human analyst. In our study there was no topic that seemed semantically opaque, nonsensical or random. However, while the topics may appear coherent, the question we wanted to answer was whether they were in fact a guide to discourse in the texts. So in the next section we explore the textual reality of these topics to see whether they are of use to the discourse analyst wanting to characterise large corpora and down-sample for close reading.
Phase 2

In this qualitative phase of our analysis, we studied the 20 most characteristic texts for each topic in order to ascertain the accuracy and reliability of phase 1 of our analysis. Note that by looking at the top 20 texts in each case, we were maximising the possibility, at least according to the scoring procedure used by the program, to identify the thematic coherence of the texts. By way of reporting the results of this part of the analysis, we group the topics into three categories representing the degree of correspondence between the results for phase 1 and phase 2 of our analysis. These categories are as follows:

- close correspondence;
- mixed correspondence;
- limited or no correspondence.

Of the 20 topics we studied, 3 exhibited close correspondence between phases 1 and 2, 9 topics exhibited mixed correspondence and the remaining 8 topics exhibited limited to no correspondence. We will begin by presenting those topics with the closest correspondence between the analytical phases, followed by those topics for which there was mixed correspondence, and then conclude the section by considering those topics for which there was limited or no overlap between phases 1 and 2. As space is constrained for the presentation of the analysis, rather than provide a full account of our interpretation of each topic, we will provide only an illustrative example for each category. When discussing the initial analysis of the topic, we will cite the specific items from the topic word-lists which guided our interpretations.

Close correspondence

Of the 20 topics, 3 exhibited what we would describe as a ‘close’ correspondence between phase 1 and phase 2 of the analysis – in other words, our initial analysis of the word-list for these topics accurately reflected, for the majority of the texts in the topic, the reality of those texts. This was the case for topics (5), (13) and (14). Take as an example topic (13). In phase 1, we interpreted this topic as being about car parking (car, park, parking, ticket), including the related issues surrounding access to disabled parking spaces (disabled), time (time, minutes, hours) and money (pay, paid, money). Our more qualitative examination of the texts assigned to this topic in phase 2 revealed that, although there was no overall theme running through all of the texts associated with this topic, 15 of the 20 texts were indeed about car parking, which was discussed in relation to a variety of issues including, as per our analysis in phase 1, issues concerning the availability of parking spaces and parking-related payments (extract 1):

**Extract 1**

I am a pensioner who has to travel over 45 miles there and back to visit this hospital. I am on pension credits which proves how broke I am! So i take my car park ticket to the cash office to claim travel allowance, and while the ticket is off my car the Highview parking group decide i have not paid although I was there under 3 hours. they now request £75 fine? When getting my
ticket my number plate is registered so why get a fine. They know I paid. I arrived 13:15 and left after extensive tests at 16:05. This shows the hospital up as it is still their car park.

Likewise, as per phase 1 of our analysis, 5 of the 20 texts in topic (13) discussed disabled parking spaces specifically. Although there was close correspondence between the results in phases 1 and 2, there were nonetheless other themes to do with car parking that could not have been apprehended merely by inspecting the topic word-list alone. For example, as extract 1 demonstrates, although issues to do with parking spaces and payment were prominent in this topic, these issues often provided context where the main complaint was about fines (incurred as a result of the aforementioned problems with payment and spaces). However, this overall trend was not clear from the word-list we analysed for this topic during phase 1.

It is also worth reiterating that our initial analysis of this topic was not able to account for the themes emerging from all of the top 20 texts in this or any other topic produced by Mallet. Although car parking featured prominently in 15 of the texts assigned to topic (13), the remaining 5 texts exhibited themes that were not identified during the initial phase of the analysis, for example a delay in seeing a doctor. Comments such as these, in which the words featured in the word-list appear to be used to provide context rather than reflecting the overall theme of the text, raise concerns not only about the thematic coherence of the texts grouped into the topic, but also with regard to the usefulness of the topic word-lists alone for identifying the ‘aboutness’ of those texts, at least as evidenced by a study of the 20 texts thought to be most closely related to each topic. This is an issue that we will return to in the discussion following the analysis.

**Mixed correspondence**

Of the 20 topics we analysed, nine could be described as exhibiting mixed correspondence between phases 1 and 2 of this analysis. By this we mean that the theme we inferred from the word-lists in phase 1 could account for some, but not the majority, of the top 20 texts assigned to the topic. This was the case for topics (1), (4), (7), (8), (9), (10), (17), (18) and (20). Taking as an example topic (9), this was a topic that we identified in phase 1 of our analysis as being about complaints relating to cancellations and waiting times. Phase 2 of our analysis showed that 18 of the 20 texts assigned to this topic were indeed about appointments, 16 of which mentioned the word ‘appointment’, while a further two were also about appointments but used alternative phrases, specifically, ‘operation date’ and ‘surgery date’. Combined, the specific themes of appointment cancellations and waits – inferred in phase 1 of the analysis – accounted for half of the texts in the topic (nine texts and one text, respectively). As per our initial interpretation, this topic also tended to contain negative evaluations. All the texts took a negative evaluative stance, except for one that mixed positive and negative evaluative elements (complaining about having to wait, but ultimately happy about the standard of treatment received).

Despite these similarities, there were also important differences between our interpretation from phase 1 of the analysis and the nature of the texts in the topic. What is being complained about in these texts was often not simply appointment delays or cancellations, but more specifically, misinformation relating to appointments, and so communication about
appointment delays and cancellations actually becomes the central issue for the majority of the texts in this topic, as exemplified by the comment in extract 2:

**Extract 2**

Cancelled appointments. I booked an appointment online to this hospital, they confirmed it then after a few weeks they just sent me a letter that they cancelled my appointment, they offered me a new one but that was not convenient for me so I had to call them to book a new appointment, the member of staff on the phone booked me the appointment what I wanted and told me that I will receive the confirmation mail. I received the confirmation mail but again with wrong appointment time, they gave me a new appointment for more than a month later … I called them back and they said that they cannot do anything. Thanks for them now I can start the whole process again, go to my GP again … Thank you very much … Not recommended!

The focus on the issue of communication is also evident in recurring references to modes of communication throughout many of the texts in this topic. Most notably, the word *phone* occurs 14 times in nine texts and its patterns of use are diffuse, including patients phoning providers, patients’ calls being ignored, patients’ calls being answered, the wrong number being used for patients, patients being greeted by an answerphone and patients not being called when they should have been. Similarly, the word *letter* occurs 14 times in 9 texts, consistently in reference to a letter that was sent (and not received) by the comment poster, while the word *answer* occurs 7 times in 6 texts and is always related to an answering machine.

Within this topic, there is also a tendency for problems with communication to be related to broader and longer-term issues with the NHS, rather than to individuals working within the immediate contexts of care. Sometimes this link is implicit, and the NHS is the referent of general terms and is the *they* and *someone* responsible for poor communication, as in extract 2. Many of the comments assigned to this topic are therefore indeed about appointments and waiting times, as per our initial analysis in phase 1. However, they are also more broadly concerned with communication problems and how these relate to system and organisational issues within the NHS more broadly – a theme that was not apparent from the topic word-list alone.

**Limited or no correspondence**

The remaining eight topics exhibited limited or no correspondence between phases 1 and 2 of our analysis. This was the case for topics (2), (3), (6), (11), (12), (15), (16) and (19). Taking as an example topic (12), in phase 1 of the analysis we commented on the vagueness of this topic but inferred that the topic was about the experiences of people on wards (*bed, ward, beds, bay*) when they have limited mobility or are unsteady (*fall, legs*). Qualitative examination of the texts assigned to this topic in phase 2 of the analysis revealed that there was, in fact, a clear overall theme running through all the comments – hospital visits. However, this theme is different from, and much more general than, any of the themes we inferred during the initial phase of our analysis. Furthermore, contrary to our initial interpretation of this topic, the majority of the texts are about hospital cleanliness/hygiene and infection control. The word *bed*, which occurs 23 times across 9
comments, was used in 3 texts to complain about hospital bedding being soiled and not changed, in 3 texts to complain about staff not using the hand gel provided at the end of hospital beds, and in the remaining 3 texts to report a soiled protective pad being left by a bed for days (extract 3):

**Extract 3**

My friend was unable to reach her drink but her daughter and I were there to be able to help! I asked for her nasal oxygen tubing to taped to her face as she could not keep it in! again just a little thing that should have been taken care of! I don’t want to read excused of low moral or lack of nursing staff, too many managers, lack of money etc. etc.! these are basic nursing needs that a student nurse learns! This lady is not a complaining type in fact quite the opposite! over the years has had operations, but never have I seen her like this. She has become negative and not drinking! today a soiled protective pad had been left by the bed since Saturday! My key concerns are:- Infection control is not being adhered to!

The appearance of the word *infection* in this extract is important – overall, the word *infection* occurs 14 times in 8 texts, half of which relate to concerns about the spread of infection resulting from poor hygiene standards. Another theme that was not evident from the word-list for this topic was complaints relating to staff attentiveness. In some cases, these complaints were related to concerns about hygiene standards. Finally, it is also worth noting that most of the comments assigned to this topic (n = 16) were negative, while the remaining four contained a mixture of positive and negative elements. Thus, none of the comments were entirely positive. However, like concerns relating to hygiene standards and staff attentiveness, this evaluative trend was not apparent during our initial analysis of the topic word-list in phase 1.

**Discussion and conclusion**

Our first research question asked ‘How reliable an indicator are the topic word-lists alone of the themes in the texts assigned to that topic?’. Our response is that topic word-lists alone are of limited use for interpreting the topics produced by the computer. This approach, explored in phase 1 in our analysis, initially appeared to allow us to generally infer the themes running through the texts grouped into our topics. However, close reading showed that it enabled us to accurately infer the themes running through the majority of the texts assigned to just three of the 20 topics we considered. A more likely outcome of the analysis was to produce mixed results, illuminating themes that occurred in approximately half of the top 20 texts assigned to the topic in question, but not the others. This was the case for nine of the 20 topics analysed. In almost as many cases (eight topics), the results from phase 1 of the analysis bore limited to no correspondence to the reality of the texts assigned to those topics, accounting, at best, for only a minority of the texts assigned to the topic in question. It should also be noted that when we returned to the topic word-lists, having qualitatively analysed the assigned texts in phase 2 of our study, we would still not have been able to ascertain the themes based on the word-list alone. In other words, we, as human analysts, did not ignore plausible interpretations that would, in hindsight, have corresponded with what we found in phase 2 of our study.
Even in those cases in which we were able to accurately infer the topic themes from word-lists alone, the more contextualised analysis of the texts assigned to those topics in phase 2 – which adopted more of a macro-structural (Van Dijk, 1977, 1980) perspective on those texts – consistently provided a more refined and complex view of the topics, and one which usually downplayed or mitigated the overall significance of the topical themes we inferred from the word-lists in the initial phase of our study. This qualitative phase of our analysis also foregrounded the importance of function words in discourse analysis. While routinely ignored in topic modelling, often in our close reading of the texts in phase 2 function words had an important role to play in the discourse of feedback. For example, the coordinating conjunctions however and but tended to be used in comments that provided a mixture of positive and negative feedback; in one comment from our topics, a patient negatively evaluates the ‘sisters on the ward’ but positively evaluates the nurses: ‘My gran recently fell ill, she was taken onto ward 5a at [name of hospital], at first the care seemed fantastic, the sisters of the ward were very rude on several occasions, the nurses however were fantastic.’ We found many other examples – if tends to be used in negative feedback; then is used to link a series of problematic events; negation words are a marker of dissatisfaction. Beyond function words, even punctuation can have an important discourse role to play – in the data both the exclamation and question marks were associated with negative evaluation. The deletion of such features – as is standard practice in topic modelling research – fundamentally limits the utility, such as it is, of this approach for discourse analysis.

Our second research question asked ‘How thematically coherent were the topics generated by the computer?’. When we qualitatively analysed the texts assigned to each topic in phase 2 of our study, we found that, of the 20 topics we considered, just 6 exhibited some theme or other commonality running through all 20 assigned texts (topics (4), (5), (10), (12), (16) and (18)). This meant that the majority (n=14) of the topics therefore lacked thematic coherence throughout and were mixed in terms of their reliability. Furthermore, of the six topics that did evince thematic coherence, only half provided what were arguably potentially useful thematic starting points for discourse analysis; topics (4) and (16) involved the positive evaluation of hospital ward/unit staff for their interpersonal skills following a surgical procedure, while topic (10) was broadly about childbirth and maternity. The themes running through the texts in the other topics were much vaguer in the context of a corpus of patient feedback; all the texts in topic (5) provided positive feedback (but about different aspects of health care), while all the texts in both topics (12) and (18) were about hospital stays but were otherwise thematically diverse. The themes found in this set of topics, though consistent, are therefore so vague in this context that they would likely be of limited use for down-sampling.

Of the genuinely thematically coherent topics in our data, our analysis of the topic word-lists in phase 1 allowed us to accurately identify the theme in question in just one case: topic (5), which grouped together texts that all contained positive feedback. In all other cases, our lack of access to the words’ original contexts of use meant that we over-interpreted the significance of certain words and inferred false relationships between others. For example, we interpreted topic (10) to be about people commenting on the experiences of a pregnant partner or relative. However, this was the case in only a minority of the texts assigned to this topic, for the majority were comments in which people described
their own experiences, written from a first-person perspective. This demonstrates the danger that arises from relying on the grouping of words in a topic word-list – it forces us to interpret meaningful relationships between words, and in turn to ‘see topics’. While this appears to be possible, in many cases the topics simply did not exist. An associated risk is that our efforts are spent trying to make sense of relationships between words and to collapse words into single topics which might not exist, rather than actually studying the content of the comments themselves, to apprehend the ways in which words might be related but also disparate in terms of their discourse functions. The qualitative discourse analysis undertaken in phase 2 not only proved to be much more effective for identifying coherent topics running through the texts, but was also useful for revising our hypotheses about the relationships between items in the topic word-lists. We found a very low level of correspondence between our topics and textual reality, a level much lower than that reported in studies of texts which, being shorter than the ones we studied, might reasonably be expected to have produced worse results (notably Qiang et al., 2016). Given that the results reported in such studies are commonly based on the methodology used in phase 1 of our research, there do seem to be grounds on which to doubt the validity not only of the results reported in such work, but also of the evaluations provided by them.

This discussion, and the analysis that preceded it, has cast doubt over the thematic coherence of the ‘topics’ generated using topic modelling methods. This in turn raises questions regarding the usefulness of topic modelling as a means of down-sampling within a large collection of texts, as topics tended to exhibit limited thematic consistency. It has also raised doubts about the usefulness of topic word-lists alone as a basis on which to infer the ‘aboutness’ of a group of texts in a given topic, and has thus highlighted issues concerning the topic discovery procedure, which lacks standardisation and is based on a rather reductive output in terms of the topic word-list – one that can not only obscure the themes in the texts in a given dataset, but also potentially imply (or at least facilitate the inference of) non-existent themes and relationships between words. Although these concerns cast doubt on the utility of topic modelling for discourse studies, we conclude this article with some recommendations for those who decide to use this form of analysis in spite of these issues.

As discussed at the beginning of this article, a major concern with topic modelling methods is the present lack of an adequate theoretical underpinning of what a topic actually is. This absence has given rise to ill-defined (and likely inconsistent) procedures of topic discovery that lack linguistic sensitivity and are, it seems, liable to make the false assumption that unordered and de-contextualised words can be mapped neatly onto propositional topics. Future research employing topic modelling methods should therefore endeavour to engage more deeply with linguistic theory when inferring the presence of topics in their textual data. For our purposes, we have drawn upon fairly established concepts advanced within semantics (Van Dijk, 1997), discourse analysis (Fairclough, 2003) and, to a lesser extent, narrative analysis (Labov, 1972). However, the theories and concepts used when defining and discovering topics are likely to depend on the aims of the research as well as the register and type of texts being analysed. Whatever the case, deeper engagement with linguistic theory in this process could give rise to more standardised procedures for topic discovery, which is particularly attractive given that the non-predictability of topic modelling approaches means that analyses cannot be replicated or tested easily by others. This is of particular concern to quantitative discourse studies of large amounts of textual data since, unlike qualitative
studies of relatively small datasets which can be accessed and inspected with relative ease, the sheer size of most corpora means that they cannot simply be read in this same way.

Our study has demonstrated that the analysis of a topic word-list alone is unlikely to be sufficient for accurately inferring the themes present in the texts grouped into a given ‘topic’. Contrary to some topic modelling research to date this phase of analysis should not replace close, qualitative analysis of the texts in the data. Instead, the examination of topic word-lists should form the initial part of the investigation and be followed up with qualitative, more contextualised analysis of the texts making up the topics (akin to phase 2 of our analysis). Only through that should a topic be determined to be credible, and the grounds for deeming a topic to be credible or not should be clear and data-driven. In addition, this more qualitative stage would enable analysts to inspect the macro-structural properties of discourse and disambiguate between word senses, distinguish between literal and idiomatic/figurative language use, begin to apprehend any themes that were obscured by the topic word-list and confirm or revise their initial interpretations. Such a qualitative analytical phase would also provide analysts with the option to filter and discard thematically incoherent topics and focus their analytical efforts on more interesting and thematically coherent topics. Although taking such measures would undoubtedly improve the accuracy of topic interpretation, the requirement to filter and select topics in this way does provide cause to question the usefulness of the topic modelling approach in the first place. Moreover, we would also raise concerns as to the practicality of carrying out such granular analyses of topics as part of a topic modelling study. This study was based on a relatively small number of topics and texts, yet phase 2 of our analysis was still laborious and time-consuming. The more ambitious the study in terms of the size and diversity of the data analysed, the less practical this most essential step becomes.

Acknowledgements
The authors would like to thank Paul Baker, Vaclav Brezina and Charlotte Taylor for commenting on earlier drafts of this article.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the UK Economic and Social Research Council’s Centre for Corpus Approaches to Social Science, grant number ES/K002155/1.

Notes
1. It should be noted that non-standard spelling is not endemic in the data, though we were sensitive to the possibility of it causing issues in our analysis.
2. Note that we used the optimise-interval option when running Mallet so that the output for each document would report the relative weight of each topic within it.
3. Or rather, we follow it in so far as Mallet’s settings allow – the stop-words function of Mallet does not remove all clitics.
4. Although the recent studies by Jaworska and Nanda (2016), Törnberg and Törnberg (2016) and Murakami et al. (2017) go beyond this by synthesising topic modelling with more contextualised analysis.
5. Note that here we are following the evaluation method used by Qiang et al. (2016: 8, section 4.2.2) and will compare our findings to theirs at the end of this section.

References
Baker P (2006) Using Corpora in Discourse Analysis. London: Continuum.

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.

Blei D (2012) Probabilistic topic models. Communications of the ACM 55(4): 77–84.

Blei D and Lafferty J (2009) Topic models. In: Srivastava A and Sahami M (eds) Text Mining: Classification, Clustering, and Applications (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series). London; New York: Chapman & Hall/CRC Press, pp. 71–94.

Blei D, Ng AY and Jordan MI (2003) Latent Dirichlet allocation. Journal of Machine Learning Research 3(4–5): 993–1022.

Brezina V (2018) Statistical choices in corpus-based discourse analysis. In: Taylor C and Marchi A (eds) Corpus Approaches to Discourse: A Critical Review. London and New York: Routledge, pp. 259–280.

Brookes G and Baker P (2017) What does patient feedback reveal about the NHS? A mixed methods study of comments posted to the NHS Choices online service. BMJ Open 7: e013821.

Fairclough N (2003) Analysing Discourse: Textual Analysis for Social Research. London and New York: Routledge.

Ghayad R, Cragg M and Pinter F (2016) Elections and the economy: What to do about recessions? The Economists’ Voice 13(1): 9–25.

Gkotsis G, Oerlich A, Velupillai S, et al. (2017) Characterisation of mental health conditions in social media using Informed Deep Learning. Science Reports 7: 45141.

Grimmer J (2010) A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. Political Analysis 18: 1–35.

Hall D, Jurafsky D and Manning C (2008) Studying the history of ideas using topic models. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, HI, 25–27 October, pp. 363–371. Honolulu, HI: Association for Computational Linguistics (SIGDAT).

Jaworska S and Nanda A (2016) Doing well by talking good: A topic modelling-assisted discourse study of corporate social responsibility. Applied Linguistics 39: 373–399.

Labov W (1972) Language in the Inner City: Studies in the Black English Vernacular. Philadelphia, PA: University of Pennsylvania Press.

Labov W and Waletzky J (1997) Narrative analysis: Oral versions of personal experience. Journal of Narrative and Life History 7(1–4): 3–38.

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

McCallum A (2002) Mallet: A Machine Learning for Language Toolkit. Available at: http://mallet.cs.umass.edu/

McEnery T, McGlashan M and Love R (2015) Press and social media reaction to ideologically inspired murder: The case of Lee Rigby. Discourse & Communication 9(2): 237–259.
Mazarura J, de Waal A, Kanfer F and Millard S (2014) Topic modelling for short text. In: Proceedings of the 2014 PRASA, RobMech and AfLaT International Joint Symposium, Cape Town, South Africa, 27–28 November 2014.

Murakami A, Thompson P, Hunston S and Vajn D (2017) ‘What is this corpus about?’ Using topic modelling to explore a specialised corpus. Corpora 12(2): 243–277.

Popper KR ([1935] 2006) The Logic of Scientific Discovery. London: Routledge.

Qiang J, Chen P, Wang T and Wu X (2016) Topic modelling over short texts by incorporating word embeddings. In: Kim J, Shim K, Cao L, et al. (eds) Advances in Knowledge Discovery and Data Mining. Cham: Springer, pp. 363–374.

Törnberg A and Törnberg P (2016) Combining CDA and topic modeling: Analyzing discursive connections between Islamophobia and anti-feminism on an online forum. Discourse & Society 27(4): 401–422.

Underwood T (2012) Topic modelling made just simple enough. Available at: https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/

Van Dijk TA (1977) Text and Context: Explorations in the Semantics and Pragmatics of Discourse. London: Longman.

Van Dijk TA (1980) Macrostructures: An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition. Hillsdale, NJ: Lawrence Erlbaum.

Van Valin RD and LaPolla RJ (1997) Syntax: Structure, Meaning, and Function. Cambridge: Cambridge University Press.

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

Zhao WX, Chen J, Perkins R, et al. (2015) A heuristic approach to determine an appropriate number of topics in topic modelling. In: Proceedings of the 12th Annual MCBIOS Conference, Little Rock, AR. Available at: https://bmcbioinformatics.biomedcentral.com/track/pdf/10.1186/1471-2105-16-S13-S8?site=bmcbioinformatics.biomedcentral.com

Zhao WX, Jiang J, Weng J, et al. (2011) Comparing Twitter and traditional media using topic models. In: Clough P, Foley C, Gurrin C, et al. (eds) Advances in Information Retrieval. New York: Springer, pp. 338–349.

**Author biographies**

Gavin Brookes is Senior Research Associate in the ESRC Centre for Corpus Approaches to Social Science at Lancaster University.

Tony McEnery is Distinguished Professor of English Language and Linguistics in the Department of Linguistics and English Language at Lancaster University.