How humans and machines identify discourse topics: a methodological triangulation

Contents lists available at ScienceDirect

Applied Corpus Linguistics

journal homepage: www.elsevier.com/locate/acorp

Articles

How humans and machines identify discourse topics: A methodological triangulation

Mathew Gillings[a,*], Sylvia Jaworska[b]

[a] Institute for English Business Communication, WU Vienna University of Economics and Business, Building D2, Welthandelsplatz 1, 1020, Vienna, Austria
[b] Department of English Language and Applied Linguistics, University of Reading, Whiteknights, Reading, RG6 6AW, United Kingdom

ARTICLE INFO

Keywords:
Triangulation
Topic modelling
CADS
Concordance analysis
Close reading
LLMs

ABSTRACT

Identifying and exploring discursive topics in texts is of interest to not only linguists, but to researchers working across the full breadth of the social sciences. This paper reports on an exploratory study assessing the influence that analytical method has on the identification and labelling of topics, which might lead to varying interpretations of texts. Using a corpus of corporate sustainability reports, totalling 98,277 words, we asked 6 different researchers to interrogate the corpus and decide on its main 'topics' via four different methods: LLM-assisted analyses; topic modelling; concordance analysis; and close reading. These methods differ according to the amount of data that can be analysed at once, the amount of textual context available to the researcher, and the focus of the analysis (i.e., micro to macro). The paper explores how the identified topics differed both between analysts using the same method, and between methods. We conclude with a series of tentative observations regarding the benefits and limitations of each method, and offer recommendations for researchers in choosing which analytical technique to select.

1. Introduction

Identifying and analysing discourse topics is crucial for discourse analysts and social scientists working with or on texts, as it helps to understand the significance, cohesion, thematic connectedness and focus within a given text or a collection of texts. By examining discourse topics, researchers can gain insights into the central themes, key ideas and beliefs (discourses), which can contribute to a deeper understanding of the social, cultural, and linguistic factors at play. Despite the currency and significance of the term *topic*, it seems to be surrounded by a kind of definitory fuzziness with various names and conceptualizations given by different scholars, ranging from a simple definition of "whatever it is that is being talked about" (Brown and Yule, 1983: 62) to more specific ones such as sentence topics, discourse topics, topics, global proposition, subject, and aboutness (Watson Todd, 2016). The notion of *aboutness*, as discussed by Scott (2006), is particularly useful for understanding discourse topics, as it suggests a cline or continuum ranging from no aboutness to minor aboutness to great aboutness, with the latter being akin to a concise text summary. For the purpose of the present study, we understand discourse topics as one- to four- or five-word short labels, mostly nouns or noun phrases, that distil and condense whatever is talked about in a text or texts into a distinctive, relevant and overarching topic. It needs to be said at the outset that for us, the process of identification and labelling of topics is inseparable in that once a topic has be given a label (e.g., *health, human rights, diversity and inclusion*, etc.), we assume that it has been identified in some way - either computationally (by machines) or cognitively (by human readers), or using a mixture of both. In this sense, a given label reifies a topic. Through examining the given labels, we can therefore understand something about the process – that is, what has been identified as having a thematic saliency in the data at hand.

Traditionally, for the analyst interested in conducting relatively small scale interpretive analyses of discourse, a form of (computer-aided) content analysis, thematic analysis, discourse analysis, or pragmatic analysis has been the answer. This work tends to be fine-grained, and whilst it may explore macro argumentative structures, word choice and grammatical patterning, "its analyses are based on close reading, thick (i.e., very detailed) description, and hermeneutic interpretation" (Gillings et al., 2023: 6). Carrying out fine-grained discourse analyses are possible with a small number of texts, yet there will always be a cut-off point (in terms of the number of texts under analysis), where the person-power, financial implications, or time commitment becomes too

* Corresponding author.
E-mail addresses: mathew.gillings@wu.ac.at (M. Gillings), s.jaworska@reading.ac.uk (S. Jaworska).

https://doi.org/10.1016/j.acorp.2025.100121
Received 15 February 2024; Received in revised form 9 January 2025; Accepted 15 January 2025
Available online 16 January 2025

large. The obvious answer is to therefore only focus on smaller datasets; but then questions around bias (e.g., primacy bias), cherry-picking and representativeness begin to emerge (Mautner, 2015).

Designed to minimise some of the pitfalls, a suite of methods stemming from linguistics and computer science have emerged. In using these methods, not only can more data be analysed at once, but they also have a quantitative component counting linguistic patterns that can reduce the inference of some human biases. This patterning is instructive, because it allows the analyst to examine how topics and discourses are built up gradually through incremental usage (Stubbs, 2001; Baker, 2023). For the analyst interested in large scale analyses of textual data spanning millions or even billions of words, one might turn to forms of text mining such as sentiment analysis, network analysis, or topic modelling. Or, if working within linguistics, they might turn towards corpus-assisted discourse analysis (CADS), widely considered to span the quantitative-qualitative divide.

Since late 2022, generative AI seems to have taken the world by storm offering its users new possibilities of working with texts. Rather than only being accessible to those with the requisite technical knowledge to program them, as was the case up until recently, generative AI tools based on large language models (LLMs) such as ChatGPT and Claude have now been developed, aimed at providing assistance to the wider public in the form of user-friendly chatbots. Seemingly, researchers across the whole academy are in the process of determining what these developments mean for their own disciplines and the methods that they use. Linguistics is no exception; as such, we find ourselves in the middle of exploring what LLM-assisted analyses mean for our work (Lin, 2023; Crosthwaite and Baisa, 2023; Curry et al., 2024; Berber Sardinha, 2024; Gillings et al., 2024).

In academic research, the methods we opt to use should be appropriate to answer our research questions and hypotheses; and it thus also follows that the choice of method impacts on outcomes that we are able to gather. In light of this, our focus in the present paper is on how the choice of method might influence the results, specifically the identification and labelling of topics from a large corpus. It is designed as a quasi-experimental methodological investigation, exploring the extent to which different methods can assist analysts with retrieving topics from a corpus. The particular methods under the spotlight are those discussed thus far: LLM-assisted analysis, topic modelling, concordance analysis (as a specific component of CADS), and close reading. Each of these methods will be explained in more detail throughout Section 2.

The present paper endeavours to contribute to ongoing methodological explorations and discussions within corpus linguistics, specifically those in relation to triangulation of linguistic and non-linguistic methods (e.g., Egbert and Baker, 2020; Baker and Egbert, 2016) and the effects of methods on results (Curry et al., 2024). Our present investigation is similar to Egbert and Baker (2020), which asked contributing authors to combine a corpus approach with other linguistic methods (e.g., psycholinguistic analysis, pragmatic analysis) to explore how results differed depending on the method employed.

This study is a form of both investigator and methodological triangulation, but as Marchi and Taylor (2009) note, the latter has the potential to be conflated with mixed- or multi-method approaches.[1] Methodological triangulation refers to a scenario where two methodologies are applied separately, before comparing findings. On the other hand, a mixed- or multi-method approach means that two methods are intertwined and interdependent on each other. CADS is a good example of this latter category, where discourse analysis is combined with corpus-assisted methods. The benefits of triangulation are manifold. Broadly speaking, Egbert and Baker (2020) suggest that it can offer a more in-depth look at a particular research topic, provide evidence for the validity of findings and reliability of methods, and act as a way to identify (ir)regularities.

In the present paper, we explore how the results of a textual analysis might differ, depending on whether LLM-assistance, topic modelling, concordance analysis, or close reading is employed in the process of identifying and labelling topics. Our aim is not to advocate for one method or another, but it is instead to compare and contrast how using each method may influence results; here, the identification and labelling of discourse topics by comparing the outputs (i.e., the given labels). This gives us a consistent basis for comparison. This study is not intended as a discourse analysis proper, although identification and labelling of topics are in most cases the very first step of such an analysis (and indeed an important one which often guides subsequent analytical focus and interpretations). In Section 2, we introduce the four methods and clarify exactly how they are being used for our triangulation quasi-experiment. In Section 3 we present the corpus under analysis, and lay out the methodological design. In Section 4 we report on the results: first comparing the responses from analysts within the same methodological condition; then comparing the responses across analyses conducted in different methodological conditions. In Section 5 we take stock of our findings and conclude with a series of tentative observations regarding the benefits and limitations of each method, and recommendations for researchers when it comes to choosing an analytical technique for the identification of discourse topics.

2. Methodological background

There are an increasing number of methods being used within the social sciences to analyse textual data. Sometimes, selecting which method to use is a careful and considered choice, depending on what the analyst wishes to examine and how the research questions can be best answered. At other times, however, analysts may find themselves to be more proficient in one method over another (potentially as a result of their own academic socialisation), and lose sight of the fact that other options are available. Regardless, it is useful for analysts to be aware of how their choice of method can impact the results that they acquire through doing such an analysis. This is particularly relevant for those who wish to engage in interdisciplinary and collaborative projects with researchers from across the social sciences. Even if we are 'united' by the interest in text and discourse, different methods are applied and it is vital to understand what these methods can or cannot bring to the table. Given that topics are often the first, and for some the most important aspect to investigate in collections of texts, we focus squarely on exploring how method impacts the results of an analysis, focusing specifically on the identification and labelling of topics.

2.1. *Key differences between approaches*

The four methods under discussion in this paper differ in three key respects: (1) the amount of data that can be analysed; (2) the amount of textual context available to the researcher; and (3) the focus of the analysis (i.e., micro to macro). Firstly, concerning (1): in a close reading, especially that which aims to conduct some form of (critical) discourse analysis, the focus is on linking large and small linguistic elements within texts to large-scale discourses within society at large (Fairclough, 1992; van Dijk, 1993; Wodak and Meyer, 2015). This is primarily based on analysts connecting what they read in texts with both their own and general world knowledge, and on making inferences (Kintsch, 1998). Such an analysis might bring about some novel or unexpected results, yet it could at the same time be 'accused' of subjectivity and possible cherry-picking. On the opposite end of the scale – that is, methods which allow for the analysis of large textual datasets with little or no human validation - we have forms of text mining such as topic modelling (popular across the social sciences and discussed from a corpus linguistic perspective in Jaworska, 2018; Brookes and McEnery, 2019; Gillings and Hardie, 2023; Bednarek, 2024). The focus here is on large-scale

[1] In fact, a 2017 editorial in the *Journal of Mixed Methods Research* highlighted this, and laid out a set of principles for developing new terminology (Fetters and Molina-Azorin, 2017).

linguistic patterning across often hundreds or thousands of texts. A typical analysis using topic modelling is less concerned with individual texts and communicative strategies within them, but more with identifying quantitative networks based on word frequency and co-occurrence. Corpus-assisted work lies somewhere in the middle; whilst large amounts of text can be processed, the researcher must still manually interact with that data through the examination of, for example, concordance lines.

Concerning (2), LDA topic modelling algorithms reduce texts to a simple bag-of-words and completely strip texts of their linguistic structure and context, presenting only a list of co-occurring words to the researcher for analysis. Researchers utilising topic modelling typically then need to eyeball these words in order to infer topics and label them (Gillings and Hardie, 2023). At the opposite end of the scale lies the completely contextualised close reading; and somewhere in between lies corpus-assisted methods. Concordance analysis specifically allows analysts to see words of interest within their co-text (typically a few words on either side), which allows them to make inferences based on close readings of shorter stretches of text.

It is a similar pattern regarding (3): small-scale interpretive discourse analytical work simply does a different thing to a method such as topic modelling; whereas the latter concerns itself with identifying themes or topics within a corpus, the former focuses more on how language is used to gradually develop and interpret themes that might tell us something about the world around us. Both ends of the scale ultimately make different epistemological assumptions, but both have been used by researchers working within relatively similar discourse-oriented fields (Pollach, 2012; McEnery and Brezina, 2022: 83–82). Again, from our perspective, CADS lies somewhere in between at the quantitative-qualitative intersection; 'the numbers tell us where to look closer', being a popular adage. CADS allows the researcher to identify areas of statistical importance, then verify those findings with human interpretation. Checks-and-balances are therefore built into the process.

2.2. Research design

In this paper, we conduct a four-way methodological comparison. We take a corpus of 10 texts and, utilising four different methods, attempt to determine the topics found within them. Whilst these methods can be paired with different theoretical paradigms and leanings, and whilst they can (and sometimes are) used in disciplines outside of the ones in which they were developed, the extent to which they do achieve this in practice is limited. As such, these methods tend to come along with a set of theoretical and epistemological assumptions of use. It is important to note at the outset that our aim is not to confirm which method is the 'best' or 'superior', but instead to contribute to our understanding of what kind of results we can obtain using a particular method and what each method might contribute to the identification of discourse topics in large text corpora.

Given that each method achieves a different aim, we opted to simplify the process by taking one method as our starting point: topic modelling. Topic modelling is a machine learning algorithm with roots in computer science and text mining, but has since been applied to, and used within, the digital humanities for social scientific research. It allows the user to input a large corpus of texts, and it then algorithmically detects bundles of co-occurring words for the researcher to then interpret and label as topics. After running the algorithm in Mallet (Machine Learning for Language Toolkit; McCallum, 2002), our topic modelling software of choice and undoubtedly the most widely-used LDA program for digital humanities, two outputs are presented to the researcher: the first is lists of co-occurring words (represented by ten lists of ten words each that are identified as strongly linked to a topic); the second is a composition document, which shows the distribution of those topics across texts.

Upon receiving the lists of co-occurring words, it is the researcher's interpretive task to determine what an appropriate label for each list of words might be. It is this label, then, that reifies the topic. It gives the topic a specific ontological status and thus makes it 'real'. According to Blei (2012), LDA (Latent Dirichlet Allocation, claimed to be the simplest of topic modelling approaches [Blei, 2012: 78]) can discover the hidden thematic structure in a collection of texts, and suggests that topic models can be used to explore, visualize, and summarise a corpus. For an overview of how LDA works below the surface, see Blei (2012), Murakami et al. (2017), and Gillings and Hardie (2023).

Before running the algorithm, the researcher has control over various parameters: the number of topics that should be found; the inclusion or exclusion of stopwords; the number of sampling iterations; and whether or not to use hyperparameter optimisation. Decisions made on each of those parameters at the outset affects the result that the researcher receives at the other side. On its surface, it appears to take the subjectivity of the researcher out of the equation, and returns a set of mathematically-derived "topics" for probing. Yet as any critical linguist would acknowledge, whilst the topic model itself may be objective, the topic *labelling* is as subjective as ever (Brookes and McEnery, 2019).

For our methodological comparison, we take the output of a topic modelling algorithm (that is, ten lists of ten words, which each make up a topic) and then attempt to interpret that output in different ways. Method A takes the output from the topic model, then instructs both ChatGPT 4 and Claude 3.5 to both assign topic labels to the lists of words. Method B follows the typical method in topic modelling by assigning topic labels on the basis of the researcher eyeballing those word lists. Method C assigns topic labels with the aid of concordance analysis. And finally, Method D identifies and labels topics based on a close reading of the 10 texts, and then determines which words are most likely to represent each topic. This is summarised in Fig. 1, then explained in more detail below.

2.2.1. Method A: LLM-assisted analyses

Since late 2022, large language models (LLM) have been cast into the public consciousness, increasingly being used in workplaces to aid in the completion of everyday tasks. One such LLM, ChatGPT, is "an artificial intelligence (AI) chatbot that processes and generates natural language text, offering human-like responses to a wide range of questions and prompts" (Doshi et al., 2023: 6). On its most basic level, an LLM is a text prediction system; they generate predicted text, based on a user-provided prompt. (For more on LLMs and how they operate, including a discussion of their relationship with Critical Discourse Studies, see Gillings et al., 2024.)

Sensationalist headlines are many and varied, suggesting that tools such as ChatGPT or its competitor Claude can be used to solve a whole manner of world issues. On a more modest level, though, there is evidence that it can help in professional writing (Cardon et al., 2023), providing solutions to business problems, and perform routine but knowledge-intensive tasks normally conducted by highly educated professionals in knowledge-based sectors (Dell'Acqua et al., 2023). Researchers in corpus linguistics are also looking to its usefulness in aiding with textual analyses (e.g., Lin, 2023; Crosthwaite and Baisa, 2023; Curry et al., 2024; Yu et al., 2024). In one such paper, Curry et al. (2024), the authors replicated three previously-published CADS analyses, but this time using ChatGPT 4 to perform (part of) the analysis, in place of traditional corpus methods. Interestingly, they found that ChatGPT 4 was reasonably effective at semantically categorising keywords and assigning a category label; yet poor at performing concordance analysis, and poor at form-to-function analysis (both of which require on additional context to interpret).

In light of these findings, we thought it useful to include LLM-assisted analysis in our four-way methodological investigation here. Thus, in Method A, we take the topic model output, and ask the newest versions of both ChatGPT 4 and Claude 3.5 to assign topic labels to them.

2.2.2. Method B: eyeballing

Method B aims to simulate a traditional topic modelling analysis.

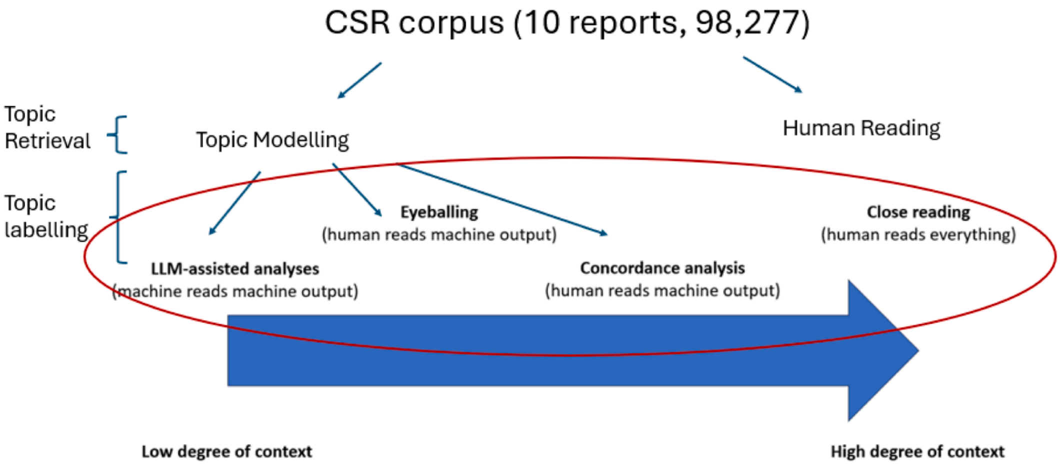

Fig. 1. The four methods compared in our investigation, organised according to the degree of context available to the researcher.

Here, topic labels are assigned based on a simple eyeballing: viewing the list of co-occurring words, and on that basis, assigning a label which adequately describes them. Undoubtedly the most common use of topic modelling in research is to identify the key messages found within a corpus, and thus identify overarching topics, rather than carry out an in-depth exploration of texts. As discussed in Gillings and Hardie (2023: 534), there is (as yet) no established procedure for generating meaningful labels, and researchers instead "characterise each topic based on their interpretation of commonalities across those words". It is of course conceivable that (some) scholars conducting such research do in fact assign topic labels in another manner, perhaps by returning to the original texts, interrogating specific co-occurring words in more detail, or by conducting a full close reading of those texts with a high percentage of a particular topic, but without an explicit statement, we are unable to replicate those procedures. We thus took eyeballing to be the central approach to topic labelling.

2.2.3. *Method C: concordance analysis*

Method C uses concordance analysis to help interpret a topic model output. Concordance analysis is one of the four key techniques used within CADS; on its most basic level, concordancing refers to a way of viewing a particular search term (a 'node' word) down the middle of the screen, with its wider linguistic co-text stretching off to the left and right. To conduct a concordance *analysis*, however the researcher must seek and identify patterns across the different lines through vertical reading. Gillings and Mautner (2024) identify four different ways in which concordance analysis can take place, with analyses differing along two main axes: how systematic it is (i.e., whether the analyst goes through every concordance line in turn and assigns categories, or informally reads through them), and whether it is top-down or bottom-up (i.e., whether categories come from the data or elsewhere).

For Method C, we uploaded the corpus to Sketch Engine (Kilgarriff et al., 2014) and instructed our analysts to conduct a concordance analysis of each word making up each topic. They were instructed to examine 100 randomised concordance lines for each word, through a form of vertical reading, and use the insights gleaned from that process to inform their topic labelling.

It is worth noting here that this procedure does not directly mimic how a topic model works. In a topic model, it is possible for the same word to co-occur with several words, and thus appear in several word lists. Or, put differently, it is possible that a word type belongs to several topics. This is in fact the case: in our ten topics (see Table 2), four words are repeated across topics (*food, energy*, and *climate* are in two topics, whereas *health* is in three). This means that, ideally, our analysts would be presented with different sets of concordance lines depending on the topic in which it is found (rather than simply a list of concordances from the general corpus). However, this information is not available via Mallet, the topic modelling software used for our analysis. The procedure we outline is thus designed to get as close to the ideal scenario as possible, with the additional caveat that it does not represent the topic modelling algorithm perfectly.

2.2.4. *Method D: close reading*

The third method we use is a form of close reading, designed to simulate a qualitative 'unassisted' form of topic identification and labelling. Because topic modelling was our methodological starting point, and we were evaluating interpretation in comparison with that, this is more of an exercise in close reading than a *true* discourse analysis. In Method D, then, we are attempting to determine which topics are identified by human researchers within the same corpus, without the aid of any computer software. We asked the analysts to conduct a form of human-based topic modelling, with the aim to explore whether human analysts identify the same salient topics based on the labels that they come up with, and the words that are in their view responsible for making up those topics.

For tasks such as this, when attempting to identify thematic areas and trends within a text, humans tend to rely on skimming and scanning. Skimming refers to rapidly reading a text in order to get an overview of the material, whereas scanning is rapidly reading to identify specific pieces of information. We expected our analysts, in this study, to mainly rely on skimming to get a general idea of the topics found within the corpus; yet we also expected them to combine that with scanning. After all, we were asking them to do a form of human-based topic modelling, and to do that, they must identify specific words that they feel made up each topic. A two-pronged approach was therefore likely.

3. Data and method

3.1. *Corpus*

The dataset used for the present methodological investigation was a 98,277-word corpus of sustainability reports from 2021. The focus on sustainability reports was justified by two reasons: first, the topic of sustainability is a highly relevant societal matter and we thought that it might be of interest to participants (important due to the fact that researchers typically conduct analyses on texts that they themselves are interested in). The latest report of the UK's Office for National Statistics reported that in 2023, the environment was one of the most important matters to people with 64% of adults in the UK worried or very worried about climate change.[2] Second, sustainability reports are written for wider audiences and stakeholder groups, of which members of the public are (supposed) to be the most important. The language of such reports is therefore less technical and they are written in a way that should be accessible to average adult readers.

The reports included in our corpus were taken from some of the largest multinational companies in some of the largest industry sectors. Specifically, we selected two companies from each of the following five major industries: pharmaceutical, food, oil, banking, and manufacturing. Table 1 below details the number of words within each report, adding up to 98,277 words in total. Although the reports come from various industries and companies, when it comes to sustainability reporting, companies need to account for similar aspects as required by various initiatives such as the Integrated Reporting Framework (see Jaworksa et al., 2024, for more information). Broadly speaking, these include: environmental performance (e.g., greenhouse gas emissions, waste, water usage, pollution prevention), social performance (e.g., labour practices, equality, diversity and inclusion polices, health and safety, wellbeing, human rights, training) and governance (e.g., management practices, business ethics, risk management), and these too were included in the selected reports.

In preparing the corpus, we made a number of edits to the texts. We decided to remove mission statements from CEOs because they are a particular genre within a genre, and we wanted to keep it as focused on sustainability reporting as possible. We also removed data from tables, graphs, footnotes, appendices, running headings, table of contents, etc., because these are often difficult for corpus analysis software to display effectively. Data was essentially stripped to the main themes discussed in the report. We also decided to remove section headings, as this may unfairly prime analysts into selecting their topics. Likewise, raw text was presented, rather than original PDFs, because the original files were full with multimodal data (graphs, images, etc.) which might skew or assist the analyst. Whilst it is true that a traditional discourse analysis would indeed be helped by these cues, we wanted the comparison to be as consistent as possible and based on exactly the same textual material. Including these elements would have primed the analysts in a way that we could not replicate for the other methods.

The corpus had to be large enough to make a topic modelling analysis worthwhile, whilst also short enough to allow our analysts to read in detail. Reading through ten sustainability reports is of course no issue for computer-assisted methods, but we wanted to keep it manageable for our human analysts too.

Table 1
Details of the 10 sustainability reports within the corpus, including the number of words per text.

Industry	Company	Number of words
Pharmaceutical	AstraZeneca	11,624
	GlaxoSmithKline	8611
Food	Arla	13,292
	Nestle	17,563
Oil	Exxon	10,070
	CNNOC	2932
Banking	Lloyds	10,430
	Santander	4193
Manufacturing	Apple	12,877
	Ikea	6685
Total number of words:		98,277

3.2. Topic model

6 analysts were involved at this stage of the project: 2 were involved in eyeballing the topic model output, 2 were involved in conducting concordance analyses of the words, and 2 were involved in close reading the texts. All analysts were either enrolled on a PhD programme or had completed one in the social sciences (including linguistics), thus representing the research community that our results have implications for. We, as co-authors, were involved in prompting ChatGPT 4 and Claude 3.5 to provide topic labels. Discourse analytical work, such as that carried out here, is very often carried out by lone researchers, and very rarely are inter-analyst reliability checks introduced (see Baker, 2015, for an extended discussion). Here, we decided to have two analysts per method in order to explore the similarities and differences not only between methods, but between analysts too.

We began by running the corpus through Mallet. We opted to exclude stopwords (using the software's default list), and asked it to identify 10 topics consisting of 10 words each. We were then presented with both the list of topics and the associated composition document (detailing which documents have the highest proportion of each topic). Methods A, B and C were all given the same set of 10 topics, each made up of 10 words. Those topics can be found in Table 2.

The decision to extract 10 topics, made up of 10 words, from a corpus of 10 annual reports is in some ways an unconventional approach. A topic model is typically run on a much larger number of texts than the number of topics identified, but there is, as yet, no agreed method for deciding on how many. Green et al. (2014) suggest that extracting too many topics can result in over-clustering, and extracting too few can produce overly broad results, and the lack of any clear guidance to aid that decision-making process is a methodological issue. For us, given that we were employing human analysts, the corpus had to be readable in its entirety, and we had to request a large enough number of topics to examine potential variation. If we had extracted 5 topics, for example, it may have led to more distinct topics, and thus more distinct labelling. Whilst this approach is thus not an issue for this paper, it does mean that generalising our findings to other projects must be done with care. Here we are dealing with a relatively homogenous corpus and a large number of topics; yet not all projects will fit that description.

To summarise, topic model labels were assigned in 4 different ways, differing in the amount of context available to the researcher:

- Method A: LLM-assistance. Here, we uploaded the topic model output table to both ChatGPT 4 and Claude 3.5. We used the

Table 2
10 topics and 10 words making up each topic, computed via MALLET.

Column A Topic number	*Column B* Words making up that topic
1	*food nestlé water business systems approach forest supply regenerative agriculture*
2	*ikea energy products materials renewable product recycled emissions climate chain*
3	*climate business including working training development improve sustainability focus supporting*
4	*health healthcare water patients programme medicines clinical patient data systems*
5	*work impact local communities access reduce solutions make part environmental*
6	*support key provide million risks management future year natural employee*
7	*global people rights human health sustainable products emissions operations employees*
8	*board exxonmobil company gas energy management waste plastic employees development*
9	*arla dairy farming food milk consumers waste carbon owners sweden*
10	*colleagues customers group support financial santander health digital programme skills*

[2] https://www.ons.gov.uk/economy/environmentalaccounts/articles/climatechangeinsightsuk/august2023

following prompt to ask for a topic label: “In Column B, there are words that were identified by LDA topic modelling as constituting a topic (one per row) in a corpus of 10 sustainability reports. Look at the words in Column B and identify an overarching summary topic for each row. Please add a Column C with your topic for each row and create a new file for download.” Importantly, all 10 lists were provided in this single prompt so that the LLM had the opportunity to distinguish between 10 topics in the same way that humans (presumably) would for the other methods.

- Method B: Eyeballing. Analysts examined the 10 topics (of 10 words each) and, based purely on their knowledge of the dataset and their wider world knowledge, assigned what they felt were relevant labels.
- Method C: Concordance analysis. Analysts ran a concordance analysis for each of the words that appeared in the topic model (100 words in total). They used a random sample of 100 concordance lines and, based on their reading through those lines, they assigned what they considered relevant labels.
- Method D: Close reading of texts. Analysts read through all 10 texts and reconstructed the topic model backwards. Firstly, analysts identified 5 key topics for each individual text, and then secondly, based on that, they identified 10 key topics for the entire corpus. Thirdly, they decided on the 10 most salient words for each topic.[3]

4. Results

In Section 4.1, we begin by looking at the similarities and differences between analyst responses within the same condition. We were interested in whether analysts utilising the same method would arrive at the same outcomes when given the same texts and same instructions. Then, in Section 4.2, we compare those outcomes across methods.

4.1. Comparing analysts’ responses within the same condition

Methods A, B, and C all used the same set of 10 topics (see Table 2), and labelled them using either LLM-assistance, eyeballing or concordance analysis respectively. Analysts utilising Method D were given the texts and asked to produce 10 topics. Their responses can be found in the following tables. The right-most column of each table contains a similarity score; this was a metric that we used to quantify the similarity of the topic label across analysts (1 = The labels are exactly or almost the same; 2 = The labels have some degree of similarity; 3 = The labels are quite or completely different).

Table 3 compares the topic labels assigned by ChatGPT (Analysis 1) and Claude (Analysis 2) when given the same prompt. Topics 2, 5 and 10 were judged as having ‘some degree of similarity’ between the two sets of labels, whilst the remaining 7 topic label comparisons were considered ‘exactly or almost the same’. There does not appear to be any pattern as to why one output differed from the other (e.g., why some labels were more or less specific or differed just by one word; why certain labels were longer than others etc.). Yet nor should there be. Because LLMs work on text prediction, it is simply a statistical product, and as such the LLM is not “thinking” about producing one response over another. The slight differences observed may stem from the distinct textual datasets on which the models were trained. However, verifying this is challenging due to the limited information about the texts in LLMs’ training data. The developers only confirmed that the models were trained on massive amounts of texts available on the internet. While variations in word choice might appear minimal, the chosen lexis can suggest different associations, perspectives, and interpretations. For instance, ChatGPT’s emphasis on "digital transformation" in Topic 10 conveys a sense of more substantial impact compared to Claude’s use of "digital skills". Furthermore, there is repetition in the terms used to label topics; for example, "sustainable" and "sustainability" appear three times in ChatGPT’s responses and five times in Claude’s, resulting in similarly themed labels. In contrast, and as will be seen below, human analysts typically attempt to create more distinct topic labels. Despite the slight differences and nuances, the overall similarity score is, however, high.

Table 3
Comparison of topic labels assigned with LLM-assistance (Method A).

Topic number	Analysis 1: ChatGPT 4	Analysis 2: Claude 3.5	Similarity score
1	Sustainable Food Systems	Sustainable Agriculture & Food Systems	1
2	Renewable Energy and Climate Solutions	Sustainable Manufacturing & Energy	2
3	Business Sustainability and Development	Corporate Sustainability Development	1
4	Healthcare and Patient Services	Healthcare Systems & Patient Care	1
5	Community Impact and Environmental Solutions	Community & Environmental Impact	2
6	Risk Management and Employee Support	Risk Management & Employee Support	1
7	Global Sustainability and Human Rights	Global Sustainability & Human Rights	1
8	Corporate Energy and Waste Management	Energy & Waste Management	1
9	Dairy Industry and Carbon Impact	Dairy Industry Sustainability	1
10	Customer Support and Digital Transformation	Financial Services & Digital Skills	2
Mean similarity score:			1.3

Table 4 shows the topic labels produced in both Analysis 3 and 4 (using Method B; that is, two researchers assigning labels via eyeballing). Looking at the similarity score, we can see a high degree of similarity and thus a relatively strong convergence of opinion between analysts. 5 of the topic labels were judged to be exactly or almost the same; 3 were judged to have some degree of similarity; and only 2 were judged to be quite or completely different.

With that said, even for those labels that were judged to be exactly or almost the same, we find that there is still some minor differences in the label. For example, Topic 6 was labelled as “Risk mitigation” in Analysis 3, but as “Risk management” in Analysis 4. Whilst on the surface these two labels appear relatively similar (mitigation is a form of management, after all), this slight difference in construal might indicate that the analysts were building up slightly different narratives of what the topic represented. Clearly, in Analysis 3, the focus was on future-proofing and

Table 4
Comparison of topic labels assigned via eyeballing (Method B).

Topic number	Analysis 3	Analysis 4	Similarity score
1	Sustainability pipelines and processes	Regenerative agriculture production	2
2	Renewable materials	Sustainable products	2
3	Educating employees about sustainability	Training for sustainability	1
4	Clinical healthcare systems	Healthcare	1
5	Environmental impact on local communities	Impact on local communities	1
6	Risk mitigation	Risk management	1
7	Sustainable workplaces	Human rights	3
8	Sustainability management	Corporate management around sustainability	1
9	Carbon reduction	Impact of Swedish dairy farming	3
10	Education programme	Corporate training	2
Mean similarity score:			1.7

[3] Our interest was in the 10 key topics that analysts decided on for the entire corpus. The other steps were to encourage analysts to think in a similar way to how the topic modelling algorithm computes topics.

reducing risk, whereas in Analysis 4, the label was more neutral. The same can be said for Topic 5, where in Analysis 3 it was labelled as “Environmental impact on local communities”, yet in Analysis 4 it was “Impact on local communities”. Both analysts were presented with the same list of words, yet one chose to highlight the “environmental impact”, whilst the other left the type of impact open to various options. Perhaps they thought it was unclear, or perhaps they thought it was implied; yet based on the simple labelling by eyeballing, we cannot say for sure. Using Method B, there is of course no way to verify those assumptions. And based on this alone, we cannot claim that convergence in analysts’ opinion equals a valid interpretation of the data. In research, we tend to argue that high inter-analyst agreement equates to reliability; but that does not by extension equate to validity.

In Method C, where topic labels were assigned with the aid of concordance analysis, there was more divergence in the analysts’ labelling (see Table 5). This time, 4 topic labels were judged to be exactly or almost the same; 1 was judged to have some degree of similarity; and 5 were judged to be quite or completely different.

In Method C, analysts examined 100 randomised concordance lines for each of the 10 words in each of the 10 models, thus equipping the analyst with much more contextual data than they received in Method B. Having more context to work with led to increased divergence in opinion. In fact, half of the topic comparisons were judged to be quite or completely different – a surprising result, given that we expect the systematicity of concordance analysis (and the systematicity of inter-analyst coding) to increase reliability. One reason for this difference could have been due to the concordance lines being randomised. We decided to randomise the lines to mimic a real concordance analysis as closely as possible. This may have meant that there was slight variation in the lines that each analyst was presented with (and thus based their labelling on), but given the rather small size of the corpus which meant that frequencies of the topic words were mostly around or below 100, it was quite likely that both analysts had very similar sets of concordance lines to read and thus, this was deemed to not be an issue.

A second reason for the low inter-rater reliability score could be in the caveat pointed out in Section 2.2.3: incorrectly assuming that all 100 instances of a word type belongs to just a single topic, whereas in reality it may have been the case that different uses of a particular node word actually belonged to a different topic. Whilst it is possible that this may have led to lower agreement (in that analysts were presented with lines potentially encompassing multiple topics), there were only four words repeated across multiple topics, and it is thus unlikely to have had a major effect. Again, whether this has any bearing on the *actual* discourse being constructed in the texts is an open empirical question. As in Method B, this is an interpretation based on the available evidence, rather than the texts themselves.

Table 5
Comparison of topic labels assigned by concordance analysis (Method C).

Topic number	Analysis 5	Analysis 6	Similarity score
1	Corporate ethics	Corporate practices	3
2	Responsible sourcing of materials	Sustainable offerings	3
3	Employee career development	Employee career and product development	1
4	Advancements in healthcare	Prioritising health and medicine	1
5	Business/community integration	Sustaining communities and the environment	2
6	Business prosperity	Securing our future: supporting people and planet	3
7	Global impact	Collaborative protection for every individual	3
8	Environment and climate change	Corporate ethics	3
9	Advancements food and farming	Cultivating sustainability: regenerative farming	2
10	Financial support and wellbeing	Employee and customer support	1
Mean similarity score:			2.2

In Method D, topic labels were produced based on a close reading. As shown in Table 6, and 4 topics were judged to be exactly or almost the same; 4 topics were judged to have some degree of similarity; and only 2 topics were judged to be quite or completely different. What that means is that there was at least some degree of similarity in 8 out of the 10 topics; this amounts to a high proportion, especially given the large amount of data that our analysts had to read through. Again, similar to Methods B and C, what we find is that some topics have a slightly different focus, even if they may appear to be similar on the surface.

Comparing analyst responses within the same condition gives us some insights into how analysts interpret data without being assisted by machines. There are essentially two polar opposites here: there is hardly any context available to the researcher in Method B (other than the list of co-occurring words and their wider knowledge), whereas there is plenty of context available to them in Method D. Yet what we find is that the similarity metric is uncannily similar: 1.7 in Method B, and 1.8 in Method D. And, by extension, Method A’s similarity score of 1.3 means that it performed ‘best’ out of all four methods (i.e., ChatGPT and Claude produced a highly similar result). What this suggests is that regardless of whether automated LLM-assistance, eyeballing or close reading is employed, the similarity score between analysts is likely to be similar. Whilst this has no bearing on the quality of the analysis (and thus the most important topics being identified), it is an important implication for inter-researcher reliability in that when analysts receive the same instructions to identify topics in the same data set (at least one that is as homogenous as this), they are likely to arrive at similar conclusions at the two ends of the spectrum (both when there is very little co-text, and when there is much of it available).

Method C – interpreting the topic model output with the aid of concordance analysis – is in many ways the odd one out. Method C should, theoretically, be the middle-ground in terms of the amount of context available to the analyst, and thus theoretically we might expect a similarity score in the middle too. Yet that is not the case and the similarity score is 2.2: a slightly higher degree of divergence, in comparison. This appears to suggest that when the analyst is given additional tools and a little more context to help their interpretation, the door is open for increased divergence of opinion.

4.2. Comparing analysts’ responses across conditions

Whilst Section 4.1 compared analysts’ responses within the same condition, the present section looks at how those responses differed across conditions; in other words, how the labels assigned via the four methods differed. This is important, because it allows us to explore how the method might impact our interpretation of data, rather than simply how different people interpret the same data.

Two topics were shared across all four methods and all eight analyses. Different analyses referred to these topics in different ways, but generally speaking we can label them as "Healthcare" and "People". The former refers to advancements in healthcare systems, whereas the latter refers to various forms of personnel development such as training programmes, additional support, empowerment, and so on. The fact that all four methods (i.e., those labels derived via the topic model and those from our human analysts) identified these topics suggest that they are the most salient topics in the corpus. There is one further topic that was identified via all four methods, but not necessarily by all eight analyses: "Ethics". Both ChatGPT and Claude had reference to “Human Rights”; one of the Method B analysts identified "Human rights" as a topic; a Method C analyst identified "Corporate ethics"*;* and a Method D analyst identified "Business ethics". Whilst not salient enough for all eight analysts to identify (as with "Healthcare" and "People"), it was salient enough for at least one analysis in each condition to identify, again

Table 6
Comparison of topics produced in Method D. Because analysts were asked to produce their own topic labels, they were sent to us unordered; as such, we have reordered them in such a way that makes comparison easier.

Analysis 7		Analysis 8		
Topic label	10 words	Topic label	10 words	Similarity score
Environment	*Forest, agriculture, water, biodiversity, animal welfare, waste, organic, reforestation, natural resources, stewardship*	Environment and nature	*sustainability, climate, footprint, welfare, biodiversity, ecosystem(s), protecting, resources, health, water*	1
Carbon	*Emission, footprint, fossil fuels, transition, carbon neutral, netzero, scope, offsetting, greenhouse gas, reduction*	Inclusive, carbon-neutral economy	*neutral(ity), net-zero, footprint, reduction, renewable, alternative, offsetting, health(y), emissions, forests(s)*	2
Climate change	*Climate, challenge, action, protection, renewables, climate risk, mitigation, adaptation, resilience, policy*	Tackling climate change	*Address, complex, problem/ challenge/ impacts, fight (ing)/combat/ tackle, target, value chain, risk(s), health, biodiversity loss, forests/ water*	2
Products/ services	*Innovation, R&D, circular, design, smart, packaging, sourcing, life cycle, longevity*	Product sustainability, affordability and availability	*Sustainable, affordable, available, inclusive, circular, safe, developing, healthy/ nutritional, long-lasting/ durable, product life cycle*	2
Governance	*Ethics, conduct, values, audit, transparency, reporting, compliance, regulations, fairness, accountability*	Corporate governance	*Board, strategic, growth, strategic, (company) value, transparency, culture, sustainability, risk, climate*	1
People	*Employees, customers, colleagues, suppliers, training, human, rights, recruitment, retention, personnel development, patients*	Employee empowerment, development, and engagement	*Empowering, development, engagement, retain, training, opportunities, health, safety, care, promoting*	2
Health	*Nutrition, safety, healthcare, wellbeing, medical insurance, mental health, covid-19, emergency, illness, pandemic*	Health	*Employee(s)/ workers/ workforce, system, strategy, mental, physical, public, global/ human, local, soil, diet*	1
Diversity	*Inclusion, ethnicity, gender, sexual orientation, disability, women, LGBTQ+, race, equity, equal opportunities*	Diversity and inclusion	*Staff, clinical trials, communities, policy, framework, recruitment, gender, ethnic, promote/ foster, belonging*	1
Company effort	*Commitment, support, contribution, goal, target, approach, strategy, partnership, ambition, help*	Business ethics	*Fair, transparent/ transparency, responsible, complying/ compliance, values, trust, integrity, conduct, regulate/ regulations, legal/law*	3
Money	*Assets, investment, affordability, pricing, market performance, shareholder value, growth, capital, costs, pay*	Safety	*Health, well-being, culture, risks, personnel/ workforce/ employees/ patient, manage(ment), environment (al), product, quality*	3
Mean similarity score:				1.8

suggesting some form of similarity between them.

Aside from the topics identified via all four methods, we can identify four topics that were shared across Methods B and C. This is perhaps unsurprising, given the analysts started with the same set of topics and associated co-occurring words, but it is interesting to see it evidenced nonetheless. These four topics were related to "Renewable and sustainable materials", "Employee training", "Business/community integration", and "Sustainable farming". One further topic was shared across Methods C and D, related generally to the "Environment". It is difficult to identify any further similarities, simply because the topics identified across analysts are so varied.

With that said, it is interesting to see how the responses gathered via Method A differed. And even despite ChatGPT and Claude being fed the same set of co-occurring words as our analysts in Methods B and C, the topic labels differed considerably. The analyses in Method A were so similar to each other that it was difficult to imagine exactly what each topic referred to more specifically. We were able to find 4 thematic trends across Methods B and C, but it was more difficult to put our finger on exactly which topic belonged to which theme in Method A. For example, 8 topic labels (spanning 5 different word lists) used the term *sustainable or sustainability*. What this suggests is that when humans assign topic labels (in Methods B, C and D), they attempt to make each topic as distinct as possible. They seemingly try to label topics in such a way that they bring a degree of diversity and creativity to the process; it is unlikely that they would use the same words to label topics, and instead vary them. This could be a matter of mere labelling, but equally it could signal that both the ‘topic labeller’ and the reader are imagining different discourses being built up within them. This can open up spaces for more diverse perspectives and interpretations. ChatGPT and Claude, on the other hand, do not do this, because they are not capable of evaluating semantic (dis)similarity (or, more accurately, without further prompting they do not seek to maximise semantic dissimilarity between topics, which we believe humans do). Instead, LLMs produce topic labels that are rather generic making it more difficult to accurately say what discourses can be found in texts with high proportions of that topic.

We noted in Section 4.1 that it is difficult to establish what the most important topics are within the corpus. After all, Methods A, B and C all work with only an abstracted set of data. Does that mean that, by implication, Method D is the ‘gold standard’ in finding the true discursive representations? It may be, but that is difficult to test. But rather than thinking about this in terms of there being a ‘true’ discourse that is out there and waiting to be found by the perfect method, we argue it is perhaps better to think of different methods highlighting different parts of the data. Think of it like each method acting as a magnifying glass on different areas. We build up the whole picture of the data through different methods and approaches.

What we do see is that our Method D analysts were able to identify topics that were nowhere to be found via the other methods. Both Method D analysts each identified topics related to diversity – one labelling it simply as "Diversity", and the other as "Diversity and inclusion". This is important, because whilst the topic modelling algorithm did not pick up on this topic, it was still considered salient enough by our human analysts to be labelled. In other words: for the human, frequency is not equal to importance, and even if something important is mentioned relatively infrequently, we may pick up on it where the computer did not.

5. Discussion and conclusion

Taken together, what can we learn about how humans assign topic labels, and thus interpret texts, in comparison with machines or when they are machine-assisted? To what extent do interpretations differ between analysts, and to what extent do different methods highlight different results?

We find that both highly decontextualised approaches (i.e., LLM-assisted analyses and topic modelling) and highly contextualised approaches (i.e., close reading) all produce high levels of inter-analyst agreement. This is interesting, because these are methods at polar opposite ends of the scale. One potential explanation for this could be that the analysts in *both* Methods B and D engaged in a form of eyeballing. Method B analysts were indeed instructed to eyeball the topic word lists; yet Method D analysts were instructed to read the annual sustainability reports in full. In addition to asking our Method D analysts to read the full texts and produce topics (as outlined in Section 2.2.4), we also asked them the following questions: “Could you elaborate on the process of how you identified words and topics? Which strategies did you use?” One analyst said that they used "close reading (combined with skimming and scanning)", whilst the second wrote the following:

> I first skimmed through the texts, highlighting frequent words and phrases. Then I did more close reading focusing on the highlighted content. After that, I tried to identify topics by grouping the words and phrases. As the next step, I listed the 5 most frequent topics of each text as instructed. Finally, I categorised the topics to form the 10 most frequent topics in all 10 texts. During this final phase I frequently went back to the texts to pick up words that make up the topic. The process was not that straightforward as many topics seem to overlap (or rather the words could be categorised under more than one topic).

In essence, then, whilst analysts did read the texts in full, they were also employing skimming and scanning, always on the lookout for words and phrases which would allow them to identify topics; a process that could be interpreted as a form of eyeballing. Such a practice may indicate why inter-analyst agreement scores were so similar across methods.

Interestingly, it was Method C (i.e., concordance analysis) where uncertainty and differences in labelling were most widespread. Researchers can either have high inter-analyst reliability and be unable to easily verify the quality of their findings; or they can have low inter-analyst reliability, arrive at potential differing conclusions, and then decide on the optimal topic labels with the support of corpus-assisted techniques and additional contextual information as evidence. Further work would be necessary here to examine what ‘deciding on the optimal topic labels’ consists of. After all, when writing up an account of the analytical process involving two analysts, researchers tend to report that categorisations were carried out independently from each other, followed by a joint decision. But what that decision-making process looks like in practice is a whole different matter.

Naturally, however, high inter-analyst agreement says little about the quality and specificity of the topic labelling. For example, labels produced by ChatGPT and Claude received a high similarity score but there were instances of repeated word choices used in each label that make it more difficult to differentiate between the topics and therefore understand what each topic is specifically about and what kind of discourse(s) are being built up within them. To answer our second question, then, we must look towards how topic labels differed across methods. Here, we found that 2 topics were shared across Methods A, B, C, and D, and a further 4 topics were shared across Methods B and C. Methods D and A were, in some way, the odd ones out. Method A produced topics that were quite similar and quite generic that smaller topics could theoretically be created within them; yet Method D produced a topic ("Diversity") that was not identified by any other method. What this suggests is that humans, in their close reading, naturally define salience not just by word co-occurrence patterns, but by their importance to society and thus bring more of their world knowledge to the process of data interpretation. This is the evidence, if it were needed, that one cannot conduct comprehensive analyses of texts and their topics – especially those which claim to have some form of relevance to the social world, as they likely do in most applications of topic modelling within the social sciences – without retaining the human component.

Can we recommend the use of one method over another? After all, each method performs the same task, but it approaches it from different perspectives. Naturally, this depends on one’s theoretical position, research aims (whether the focus is purely on which topics are present, or whether there are wider questions about how those topics are constructed), time, and the amount of funding available (whether it is possible to read a large collection of texts in full), and so on. The human outputs are not completely dissimilar from the machines, yet it is clear all the same that there is no shortcut to easy interpretation in terms of what most the salient topics are. Gillings and Hardie (2023: 542), in conceiving a concordance-based approach to topic model interpretation, suggested that “Like close document-reading […] this concordance-based approach would drastically increase the effort required by the analysis. We expect any *useful* approach to interpretation of topic models to have that effect.” And so the proof is in the pudding: researchers are unlikely to have the time nor inclination to read through potentially hundreds of company annual sustainability reports and perform a close reading, yet nor should they accept a research design where machines do all the work, and few checks and balances are in place to keep track of issues. Our Method C, whilst having the lowest inter-analyst agreement, means that those same analysts have the tools and materials available to them to discuss differences and arrive at a joint conclusion.

Naturally, some caveats are in order. Firstly, it is highly likely that topic labels would have differed if the corpus had been divided in a different way prior to it being run through the topic modelling algorithm. We could have, for example, divided our sustainability reports by section or even paragraph (thus leading to a more granular analysis). Whilst this does not matter for comparing Methods A, B, and C, it might be that if we had split the corpus up according to section (where we might expect a section on Diversity, for example), then the topics may have more closely resembled those found via Method D. There are no best practice guidelines on how (or whether) to divide texts, though, so we opted to use full texts. Secondly, as discussed in Section 3.2, our decision to identify 10 topics made up of 10 words is open to debate, even for Method D where our human analysts perhaps would have identified either more or fewer topics, if left to their own devices. Thirdly, our 'similarity score' in Section 4.1 could perhaps have been operationalised in another way (e.g., the percentage of words which are used across analysts' topic labels), yet given our focus here was on how interpretations differed qualitatively, the metric serves only to highlight the major differences.

As more and more tools and programs are developed to wrangle large corpora, methodological triangulations such as this (cf. Baker and Egbert, 2016; Egbert and Baker, 2020) are vital to understand the efficiency and usefulness of individual approaches. Likewise, as LLMs continue to become more widely used – both by the general public and by academic researchers alike – we should not become so enamoured by them that we lose sight of key academic principles of transparency and reflexivity (see also comments made in Curry et al., 2024 and Jaworska, 2024). These principles seem more important than ever, given the lack of transparency regarding the mechanics underlying newer tools, especially those based on LLMs, which notably are not even well understood by their developers (and hence, there is a lack of protocols or manual to how to use them at present). More than ever, there is a need to test how tools and methods work in practice, and what they can do for us, also in comparison with more established approaches. In situations where researchers do not have the time or resources to engage in testing the functionalities of various approaches, it is vital that they include a methodological protocol with details regarding how the outputs were obtained (for example, the list of words retrieved using topic modelling) and what procedures were adopted to categorise outputs. In corpus linguistic research, it is considered 'gold standard' to include, for example, the exact parameters for retrieval and classification of data when undertaking a collocation or concordance analysis, and then explaining the effect of those parameters (Baker, 2023: 226). The same approach ought to be adopted in research concerned with any textual analysis, whether machine-assisted or not.

Overall, it would be useful to carry out similar quasi-experimental investigations with more analysts and more methods – including those which are either currently in development or increasingly being used across the social sciences (e.g.,. vector analysis). As linguists, we are well-placed to offer expertise on how these new methods ontologically view language, and how our methods are (or are not) appropriate to deal with such rich context-specific data.

Funding

The research presented in this paper was supported by funding from the Research and Impact Fund of the University of Reading.

CRediT authorship contribution statement

Mathew Gillings: Writing – review & editing, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Sylvia Jaworska:** Writing – review & editing, Resources, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank the analysts who so diligently interpreted the topic model according to their respective methods and thus allowed us to draw comparisons between them. We would also like to thank the two anonymous reviewers for their insightful, in-depth comments and suggestions. Any infelicities in the manuscript remain ours.

References

Baker, P., 2015. Does Britain need any more foreign doctors? Inter-analyst consistency and corpus-assisted (critical) discourse analysis. In: Groom, N., Charles, M., John, S. (Eds.), Corpora, Grammar and Discourse: In honour of Susan Hunston, pp. 283–300. John Benjamins.

Baker, P., 2023. Using Corpora in Discourse Analysis, 2nd edn. Bloomsbury.

Baker, P., Egbert, J., 2016. Triangulating Methodological Approaches in Corpus-Linguistic Research. Routledge.

Bednarek, M., 2024. Topic modelling in corpus-based discourse analysis: uses and critiques. Discourse Stud. https://doi.org/10.1177/14614456241293075.

Berber Sardinha, T., 2024. AI-generated vs. human-authored texts: a multidimensional comparison. Appl. Corpus Linguistics 4, 100083.

Blei, D., 2012. Probabilistic topic models: surveying a suite of algorithms that offer a solution to managing large document archives. Commun. ACM 55 (4), 77–84.

Brookes, G., McEnery, T., 2019. The utility of topic modelling for discourse studies: a critical evaluation. Discourse Stud 21 (1), 3–21.

Brown, G., Yule, G., 1983. Discourse Analysis. Cambridge University Press.

Cardon, P., Fleischmann, C., Aritz, J., Logemann, M., Heidewald, J., 2023. The challenges and opportunities of AI-assisted writing: developing AI literacy for the AI age. Bus. Prof. Commun. Quarterly 86 (3), 257–295.

Crosthwaite, P., Baisa, V., 2023. Generative AI and the end of corpus-assisted data-driven learning? Not so fast! Appl. Corpus Linguist. 3 (3), 100066.

Curry, N., Baker, P., Brookes, G., 2024. Generative AI for corpus approaches to discourse studies: a critical evaluation of ChatGPT. Appl. Corpus Linguist 4 (1), 100082.

Dell'Acqua, F., McFowland III, E., Mollick, E., Lifshitz-Assaf, H., Kellogg, K.C., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K.R. (2023). Navigating the jagged technological frontier: field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Working Paper,* No. 24-013.

Doshi, R.H., Bajaj, S.S., Krumholz, H.M., 2023. ChatGPT: temptations of progress. Am. J. Bioeth. 23 (4), 6–8.

Egbert, J., Baker, P., 2020. Using Corpus Methods to Triangulate Linguistic Analysis. Routledge.

Fairclough, N., 1992. Discourse and Social Change. Polity Press.

Fetters, M.D., Molina-Azorin, J.F., 2017. The *Journal of Mixed Methods Research* starts a new decade: principles for bringing in the new and divesting of the old language of the field. J. Mix. Methods Res. 11 (1), 3–10.

Gillings, M., Hardie, A., 2023. The interpretation of topic models for scholarly analysis: an evaluation and critique of current practice. Digital Scholarship in the Humanities 38 (2), 530–543.

Gillings, M., Mautner, G., Baker, P., 2023. Corpus-Assisted Discourse Studies. Cambridge University Press.

Gillings, M., Kohn, T., Mautner, G., 2024. The rise of large language models: challenges for Critical Discourse Studies. Crit. Discourse Stud. https://doi.org/10.1080/17405904.2024.2373733.

Gillings, M., Mautner, G., 2024. Concordancing for CADS: practical challenges and theoretical implications. Int. J. Corpus Linguist. 29 (1), 34–58.

Green, D., O'Callaghan, D., Cunningham, P., 2014. How many topics? Stability analysis for topic models. In: Calders, T., Eposito, F., Hüllermeier, E., Meo, R. (Eds.), Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2014. Springer, Heidelberg, pp. 498–513.

Jaworksa, S., Stenka, R., Parlakkaya, E., 2024. Management by keywords: a corpus-based investigation into the discourse of six capitals in best practice integrated reporting. Int. J. Corpus Linguist. 29 (3), 331–360.

Jaworska, S., 2018. Doing well by talking good? A topic modelling-assisted discourse study of corporate social responsibility. Appl. Linguist. 39 (3), 373–399.

Jaworska, S., 2024. Back to the future: topic modelling and beyond. Discourse Studies. https://doi.org/10.1177/14614456241293970.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V., 2014. The Sketch Engine: ten years on. Lexicography 1, 7–36.

Kintsch, W., 1998. Comprehension: A Paradigm For Cognition. Cambridge University Press.

Lin, P., 2023. ChatGPT: friend or foe (to corpus linguists)? Applied Corpus Linguistics 3 (3), 100065.

Marchi, A., Taylor, C., 2009. If on a winter's night two researchers… a challenge to assumptions of soundness of interpretation. Crit. Approaches to Discourse Anal. Across Discip. 3 (1), 1–20.

Mautner, G., 2015. Checks and balances: how corpus linguistics can contribute to CDA. In: Wodak, R., Meyer, M. (Eds.), Methods of Critical Discourse Studies, 3rd edn. SAGE, pp. 154–179.

McCallum, A. (2002). *MALLET: a machine learning for language toolkit.* http://mallet.cs.umass.edu. Accessed 14th February 2023.

McEnery, T., Brezina, V., 2022. Fundamental Principles of Corpus Linguistics. Cambridge University Press.

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

Pollach, I., 2012. Taming Textual Data: the contribution of corpus linguistics to computer-aided text analysis. Organ. Res. Methods 15 (2), 263–287.

Scott, M., 2006. The importance of key words for LSP. In: Macià, E.A., Soler Cervera, A., Rueda Ramos, C. (Eds.), *Information Technology in Languages for Specific Purposes: Issues and Prospects*. Springer, New York, NY, pp. 231–243.

Stubbs, M., 2001. Words and Phrases: Corpus Studies of Lexical Semantics. Blackwell.

van Dijk, T.A., 1993. Principles of critical discourse analysis. Disco. Society 4 (2), 249–283.

Wodak, R., Meyer, M., 2015. Methods of Critical Discourse Studies, 3rd edn. SAGE.

Yu, D., Li, L., Su, H., Fuoli, M., 2024. Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: the case of apologies. Int. J. Corpus Linguist.