Emergency Risk Communication: A Structural Topic Modelling Analysis of the UK government's COVID-19 Press Briefings

Ying Wang (Karlstad University)

Abstract
The ongoing coronavirus outbreak has caused a public health emergency of international concern. During public health emergencies, effective risk communication plays an indispensable part in a country's emergency response. This paper explores the use of Structural Topic Modelling, a machine learning technique that automatically identifies key topics and their content in textual data, in analysing emergency risk communication (ERC) practice at the state level. The data is from the UK government's COVID-19 press briefings televised between March 2020 and June 2021, totalling approximately 1 million words. The study identifies the prominent topics covered in those briefings as well as their distribution over time, which in turn reflect the UK government's priorities in handling the public health emergency. Close scrutiny of the use of a selection of key words in context sheds further light on the government's ERC practice from a linguistic point of view.

Keywords: COVID-19; emergency risk communication; Structural Topic Modelling; corpus linguistics

1. Introduction

The rapid spread of the novel coronavirus (SARS-CoV-2), a highly infectious disease, has caused an epidemic of acute respiratory syndrome (COVID-19) since the beginning of 2020 (Dryhurst et al. 2020). On 30 January 2020, the World Health Organization (WHO) declared the coronavirus outbreak a public health emergency of international concern. During public health emergencies, effective risk communication plays an indispensable part in a country's emergency response as it is crucial for people to 'understand, trust and use' the information provided through various channels about what health risks they face and what actions they can take to protect themselves, their families and communities (WHO

Wang, Ying. 2022. 'Emergency Risk Communication: A Structural Topic Modelling of the UK Government's COVID-19 Press Briefings.' *Nordic Journal of English Studies* 21(2): 226–251.

2017: 1). The importance of emergency risk communication (ERC) is increasingly recognised in the field of public health. However, much attention has been given to emergency language services, i.e., how to overcome linguistic barriers in health services in multilingual communities (Dreisbach and Mendoza-Dreisbach 2020; Li et al. 2020). The ERC practice at the government level, which supposedly reaches more people than healthcare services do, remains a gap to tack into from a linguistic perspective. From 16 March 2020 to 21 February 2022, the UK government held regular COVID-19 press briefings, where government ministers were joined by medical and scientific experts, updating the public on the latest data on coronavirus and addressing questions from the media and members of the public. These briefings provide a valuable resource for analysing ERC practice at the state level.

Over the past few decades, corpus linguistics has developed into a particularly prolific research approach for empirical investigations of patterns of language variation and use (Biber 2010). Some basic corpus analysis methods such as collocation and keyword analysis have been fruitfully exploited to study words and word-based patterns, which help to distinguish different texts and styles (Scott and Tribble 2006). Structural Topic Modelling (STM) is a machine learning method developed by Roberts and colleagues to automatically discover latent topics from textual data and to estimate probabilities that clusters of words constitute the key topics (Roberts et al. 2014; Roberts et al. 2019). The technique has been widely used in social science research for exploring the actual thematic content of large-scale textual discourse such as open-ended survey responses (Roberts et al. 2014) and texts produced by individual and organisational actors in the climate change countermovement (Farrell 2015). The potential of the technique has attracted increasing attention in linguistic research in the last few years (e.g., Liu and Lei 2018; Brookes and McEnery 2019; Busso et al. 2022). The present study pursues this line of research in exploring the potential of STM in the study of ERC discourse.

Using the STM technique, the aim of this paper is to identify and examine the main topics of the UK government's COVID-19 briefings televised between March 2020 and June 2021 as well as the significant lexical and grammatical patterns in the UK government's ERC strategies.

2. Background

2.1 Emergency risk communication (ERC)

Risk communication is defined as 'the real-time exchange of information, advice and opinions between experts, community leaders, or officials and the people who are at risk' (WHO 2017: ix). The goal of risk communication during epidemics and pandemics is to promote risk mitigation behaviours among people at risk. In order to achieve this goal, it is crucial that people understand the information and advice provided and are convinced that it is relevant, trust-worthy, and acceptable.

WHO outlines a number of essential elements of ERC, echoed also by the European Centre for Disease Prevention and Control (ECDC 2017). They include addressing people's needs and concerns, building trust, and engaging with communities as well as clear, accurate and consistent messaging, which should promote specific actions people can realistically take to protect their health. Achieving those objectives requires certain linguistic practices; as Li et al. (2020) contend, language plays an important role in the success of a country's emergency preparation and response. However, while attention has been drawn to the importance of providing language services to overcome language barriers in the dissemination of information in a multilingual society, little has been done to examine specifically the role of language use in ERC practice. The present study is an attempt to fill that gap.

2.2 Structural topic modelling (STM)

STM is a statistical approach for analysing textual data that allows researchers to discover and explore topics in the data and estimate their relationship to document metadata (e.g., date, speaker/writer) (Roberts et al. 2019). Within this framework, topics are defined as clusters of words that co-occur according to certain probabilistic patterns across a collection of documents (Blei 2012) and that represent semantically interpretable themes (Roberts et al. 2014). Unlike the corpus linguistic concept of collocation, which focuses on co-occurrence of words in close proximity, STM operates on a larger scale, typically an entire text. In other words, it is possible for the computer to infer a relationship between any words that frequently co-occur within the same text, regardless of the distance between them (Brookes and McEnery 2019). STM is similar to keyword analysis, another corpus linguistics tool, in the sense that both can be used to identify keywords in a collection of texts and thereby discover the

'aboutness' of the corpus. However, unlike the keywords approach, which requires a reference corpus or wordlist (Culpeper 2002), STM operates on a single corpus and is therefore free from the constraints imposed by the choice of data for comparison.

STM and its implementation in the statistical software *R* (Roberts et al. 2019) can be used not only to identify topics in a corpus, but also to outline topic content (i.e., words that are most characteristic of each topic) and topical prevalence (i.e., how much of a document is associated with a topic). The technique has been exploited and proved fruitful in linguistic research in recent years. Murakami et al. (2017), for instance, explore the model in a corpus of academic English and demonstrate that the topics retrieved provide rich insights into the nature of the corpus as regards the distribution of prominent topics in different parts of research papers and over time. The technique has also successfully assisted discourse studies involving various types of data, such as Liu & Lei (2018) on key discourse strategies in Hillary Clinton's and Donald Trump's presidential campaigns, Skalicky et al. (2020) on the linguistic features of humorous deception in news stories, and Busso et al. (2022) on a forensic linguistic analysis of an historic collection of abuse letters sent anonymously.

While this machine learning technique has the advantage of inferring the key topics of and their content using algorithms (Roberts et al. 2014), it is up to the researcher to decide how the topics should be interpreted once they have been generated (Busso et al. 2022). Therefore, a qualitative approach is required to provide a more nuanced understanding of the data. In fact, as will be demonstrated in this article, some of the limitations of the technique in linguistic research brought up by Brookes and McEnery (2019), such as disregarding words' context of use, excluding closed class words from the analysis, and reducing morphological variants of words to their common base form, can be partly overcome through a qualitative analysis, aided by traditional corpus linguistics tools such as collocation and concordance.[1]

As can be seen from this brief review, linguistic studies using STM are still in their early days, and the usefulness of the technique is not without debate. This is one of the reasons why more studies exploring its use should be welcome. As will be seen, the method is particularly useful

[1] See also Busso et al. (2022) for a rebuttal against the other reservations Brookes and McEnery (2019) had about the usefulness of STM in linguistic research.

for an initial exploration of the data involved in the present study, namely the UK government's briefings given during the COVID-19 pandemic.

3. Data and procedure

The UK government's COVID-19 briefings can be accessed from the BBC's YouTube account under the playlist entitled 'Coronavirus (COVID-19): Government Briefings'. Automatically-generated transcripts are available for most of the briefings. The corpus used for the present study consists of transcripts of 150 briefings of approximately 100 hours (6,027 minutes), with a total of 996,040 words. The automatically-generated transcripts were manually checked for mistakes (e.g., *covet* for *covid*). The briefings span 15 months from 16 March 2020 (the first one) to June 30, 2021. The briefings were given on a daily basis in the first two months or so and then became less regular as the country went through the first wave of the pandemic. Typically, a briefing consists of two parts: opening remarks/presentations of data and a Q&A session. In total, the opening remarks account for 2,058 minutes and the Q&A sessions 3,969 minutes. In this study, the two types of data were examined together in order to offer an overall picture of the key topics covered in these briefings. The bulk of the data were contributed by 37 speakers, of whom 21 are government officials and 16 are medical and scientific experts. The Q&A sessions contain questions from the public and journalists, which make up a tiny fraction of the corpus.

The analysis was conducted using the statistical software *R* with the Structural Topic Model *stm* package (Roberts et al. 2019). The dates on which the briefings were televised were used as metadata for the corresponding texts in the corpus. Before modelling, the texts were pre-processed using the *textProcessor* function from *stm* to reduce words to their root form and to remove stop words (e.g., *the, is, at*), punctuation, and annotation tags.

The second step was to select the number of topics for topic modelling, as the number of topics will greatly affect how good the final topics are (Silge and Robinson 2017). Residual checks and held-out likelihood estimation are two methods that help understand how different topic models perform at various numbers of topics, and can be used to select the best number of topics automatically (Roberts et al. 2019). More specifically, held-out likelihood estimates the probability of the model generating unseen held-out data and shows how well the model generalises

(Wallach et al. 2009). A better model will give rise to a higher probability of held-out. Residual checks measure overdispersion of the variance of the multinomial within the data. If the residuals are overdispersed, it means that more topics may be needed to absorb some of the extra variance (Taddy 2012). While no foolproof method has been developed to offer a right answer for the number of topics that is appropriate for any given corpus, both these measures are considered as useful indicators (Roberts et al. 2019) and were therefore performed in the present study, using the function *searchK* from *stm*, to help choose the number of topics.

The results are shown in Figure 1. As can be seen, the held-out likelihood is highest at 20, and the residuals are lowest around 20 to 40. These diagnostics indicate that a statistically sound number of topics (*K*) would be 20.

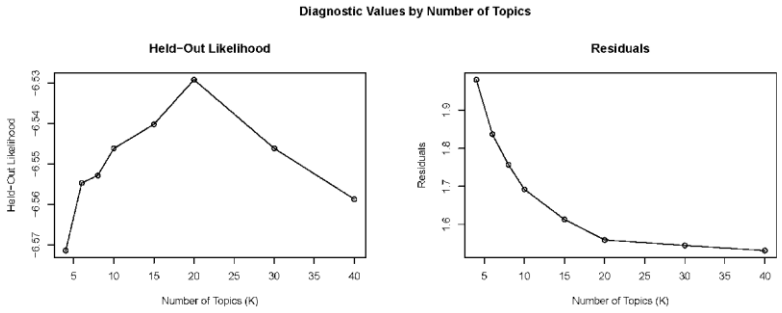

Figure 1. Model diagnostics by number of topics

After $K = 20$ was set, the *stm* function was used to estimate the structural topic model. To explore the words associated with each topic, the *labelTopics* function was used. The function returns different types of word profiles, including highest probability, FREX, Lift, and Score. Highest probability words are most common ones for each topic, inferred directly from topic-word distribution parameter β. FREX ranks words by their overall frequency and how exclusive they are to the topic. Lift and Score give higher weight to words that appear less frequently in other topics.[2] There are some overlaps across topics particularly in terms of

[2] Lift weights words by dividing their raw frequency in the topic by their frequency in other topics, while Score uses the log frequency for the calculation (Roberts et al. 2019).

highest probability words, which are expected given the interrelatedness of all thematic subjects. The top words of the other three measures, however, seem more informative in distinguishing themes. All the measures were therefore considered in deciding a thematic label for each topic (see Figure 2).

4. Results and discussion

In this section, the expected proportion of the corpus that belongs to the 20 topics identified will be presented and discussed first (4.1), followed by a close analysis of a selection of key topics in relation to prominent language features (4.2–4.3) and their distribution over time (4.4).

4.1. Topics identified and their proportion

Figure 2 presents the proportion of the 20 topics in the corpus.

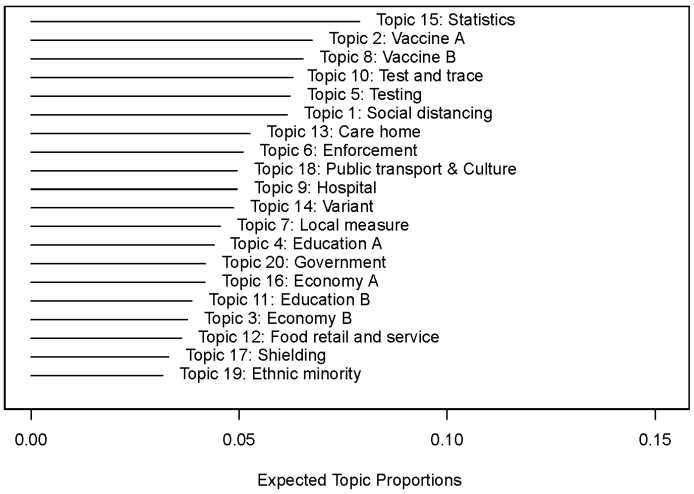

Figure 2. Proportion of top topics in descending order

Among the topics, some appear to have a common theme (e.g., Vaccine for Topics 2 and 8, Economy for Topics 16 and 3, Education for Topics 4 and 11) as they share several key words. However, the whole lists of

words, particularly those associated with the measures of FREX, Lift, and Score, seem to point to different thematic dimensions. For instance, both Topics 2 and 8 centre on vaccine, but the former (labelled as Vaccine A) features general vocabulary associated with vaccine, including medical terms (e.g., *dose*, *jab*, *shot*), distribution and uptake (e.g., *supply*, *deploy*, *rollout*, *program*, *uptake*, *invite*, *hesitate*), types of vaccine (*pfizer*, *astrazeneca*), related institutes or committees (*mhra*, *jcvi*, *oxford*), vaccination sites (*gps*, *pharmacy*, *center*), while the latter (labelled as Vaccine B) has a specific dimension focusing on the 'effects' of vaccination, not only from a medical point of view (e.g., *protect* against *variants*, boost *immunity*, reduce *transmission* and *risk* of *severe* disease and *mortality*), but also societal, in relation to *stages/timetable* of the *unlocking roadmap*, for instance. Examples (1) and (2) illustrate how the highlighted words, which occur on the list for Topic 8, centre around the dimension on the effects of vaccines in terms of restricting increased transmission caused by lockdown easing and offering overall protection. Interestingly, when protection is mentioned, it serves either as a kind of caution after positive news—although the vaccines are helpful, it takes time before we are fully protected, as in example (1), or as justification for certain steps taken in the vaccination program, such as the decision to 'go faster' by delaying the second dose and offering the first dose to as many people as possible (see example (2)).

(1) we accept that with this **highly transmissible** virus of course there will be an increase in **transmissio**n as we **unlock** vaccines will help hem it in as we go down the **ages** but it's going to take a while before we've got the full **protection** (DSB210329)[3]

(2) and our quite strong view is that we think its **protection** will be quite a lot more than 50% so therefore in net public health terms there'll be **substantially** more **protection** by going faster (DSB210105)

As shown in Figure 2, the most discussed topic in the corpus is related to the latest data on coronavirus, with words such as *number*, *test*, *death*. This

[3] The corpus citation formula "DSB210329" stands for Downing Street Briefing, broadcast on 29 March 2021.

is not surprising as most briefings open with a government minister reporting on a series of numbers about tests carried out, positive cases, and deaths, as in example (3). A more detailed presentation and analysis of the statistics is also given later by a medical or scientific expert. The latter often involves comparison of data over time and across countries, hence the use of words such as *peak*, *flatten*, *decrease*, *differ*, *exponential* (*rise/growth*), *slower*, and *comparison*. As can be seen in example (4), the comparisons often lead to predictions on the development of the situation with words such as *expect*, *think*, and *might*.

(3) let me give you an update on the latest data from the **cobra** coronavirus data file and I **can** report that through the government's ongoing monitoring and testing program as of today 559,935 people have **now** been **tested** for the virus 133,495 have **tested** positive and of those who have contracted the virus 18,100 have very sadly died (DSB200422)

(4) and here the **deaths** are **now** below the first **peak** average and coming down but they still **got** a long way to go and i **expect** them to take some weeks to come right down final **slide** please (DSB210210)

Overall it seems that the positive side of the development gets highlighted, given the presence of words such as *flatten* and *decrease* that are highly associated with this topic, whereas their antonyms seem absent from the list. The following examples illustrate the use of the word *decrease* in context. In (5), it occurs in a matter-of-fact description of the trend of hospital admissions. The positive news is further strengthened by the adverb *steadily*. In (6), *to decrease* is an aim, used as a justification for the restrictive measures to stay in place. The use of the word in (7) is similar to that in (5); however, this time, the good news is followed immediately by caution (i.e., the *decrease* does not mean the disease has gone away)—the same strategy as used in (1) with *protection*.

(5) overall you can see the estimated new daily admissions with covid 19 **peaked** in early April that has been steadily **decreasing** ever since (DSB200521)

(6) it's important that all the measures that we're taking stay in place in order to allow us to maintain this level of control and to see the epidemic begin to **decrease** (DSB200416)

(7) what you can see is that we are coming back down to average levels of death for the UK now and that's due in the purple line to the **decrease** in covid deaths and the other consequences of the covid infection so it is coming back down towards baseline towards normal, but don't be fooled that this means it's gone away the disease is growing across the world (DSB200623)

Apart from Topic 15, the topics can be roughly divided into two broad categories: measures taken to tackle the disease (Vaccine, Test and Trace, Social Distancing, Local Measure, Shielding) and societal sectors of concern (Care Home, Public Transport and Culture, Hospital, Education, Government, Economy, Food Retail and Service, Ethnic Minority). In terms of proportions, it is quite clear from Figure 2 that topics related to the former are more heavily represented in the corpus than the latter. In what follows, a selection of topics in each category will be examined more closely. For the former, the most frequent ones (Vaccine A, Test and Trace, Testing, Social Distancing) were chosen. There are two topics associated with vaccine, of which only Vaccine A was picked, as it covers more aspects and is thus more representative of the topic than Vaccine B. Similar considerations have led to the choice of Care Home, Hospital, Education A and Economy B for the subsequent analysis.

4.2 Key topics related to measures

Figure 3 presents the top 20 high-probability words for the four chosen topics associated with the main measures taken to tackle the epidemic in the UK.

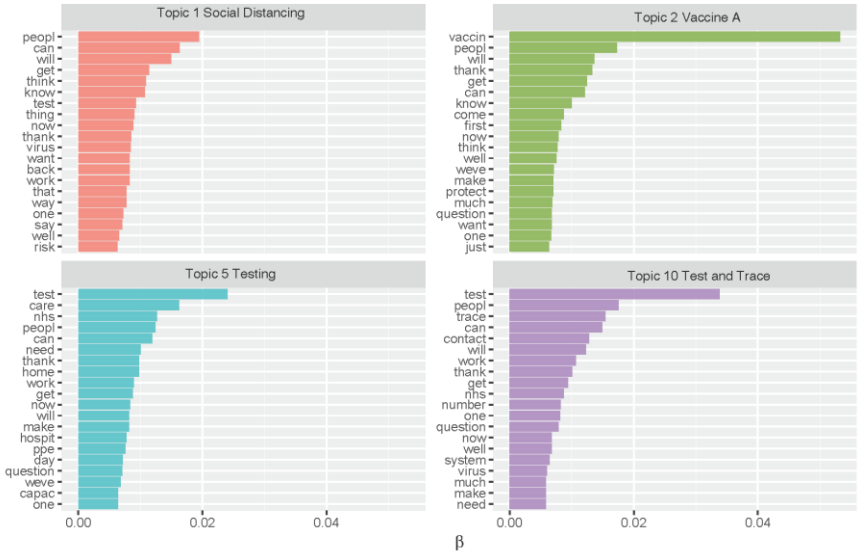

Figure 3. Top 20 word probabilities for Topics 1, 2, 5, and 10 (Beta value matrix stores the log of the word probabilities for each topic to understand what each topic is really about.)

What is clear from Figure 3 is that many words are common to all four topics, unlike in some previous studies (e.g., Busso et al. 2022; Liu and Lei 2018) where high-probability words alone can distinguish topics. Among the 20 words for each, 7 are shared by all the topics: *people*, *can*, *will*, *get*, *now*, *thank*, *one*, and 4 more are shared by three out of the four topics: *work*, *question*, *make*, *well*.

While Topics 2, 5, and 10 can be inferred from a small number of the high-probability words (*vaccine*, *protect*; *test*, *ppe*, *capacity*; *trace*, *contact*, *system*), those of Topic 1 alone are not informative enough on a common theme. As mentioned earlier, the top 20 FREX, Lift, and Score words were also consulted. Once the theme has been revealed with the help of those words, the high-probability words start to make sense. With regard to Topic 1, example (8) summarises the theme quite nicely, which is about social distancing measures (*one meter plus*, *two meters apart*, *outside your household*) and the basics of hygiene (*hands*, *face*, *space*, *outdoors*). The aim of such measures is often related to reducing the risk

of transmission, and thereby allowing people to get back to normal way of life and work, as is evident in (8) and (9). Most of the highlighted words in these two examples have high FREX, Lift, and Score values. Once the theme is clear, we can see how some high-probability words (e.g., *people, one, risk, now, back, work, way*) are related to it.

(8) having considered all the evidence while staying at two **meters** is preferable we **can now** move to **one meter** plus where it is not possible for us to stay two **meters** apart that means staying **one meter** apart plus mitigations which reduce the **risk** of transmission and these precautions could include installing screens making sure that **people face** away from each other providing **hand-washing** facilities minimizing the amount of time you spend with **people** outside your **household** and and of course being **outdoors** on public transport it already means **one meter** plus means wearing a **face** covering for mitigation as as everybody I **think now** understands (DSB200623)

(9) we want **people** to be able be confident to to go **back** to **work** in a covid **secure way** (DSB200909)

The fact that a great proportion of the high-probability words are shared by these topics is interesting in itself, however, and worth commenting on. For one thing, it means that the themes are closely related to each other. Words such as *thank* and *question* are associated with the same event type involved in the data (i.e., press conference with a Q&A session). It is noteworthy, though, that they occur on the lists for Topics 2, 5, and 10, but not for Topic 1, probably suggesting that topics such as vaccine, testing, and test and trace are more likely to prompt questions in the press conferences than the basics surrounding social distancing.

Another observation that can be made regarding Figure 3 is that many of the top high-probability words shared by more than two topics are high-frequency words such as modal verbs *will* and *can* and delexical verbs such as *get* and *make*. It may be puzzling why such words end up as high-probability words associated with these topics, a question that would prompt further investigation. As will be shown below, when the surrounding context is taken into account, these words can also be quite revealing about the characteristics of the government's ERC practice.

In order to find out why *will* and *can* feature prominently in all four topics, AntConc (Anthony 2019), a multi-purpose corpus analysis toolkit, was used to search for their collocates in the entire corpus and to examine their use in context. The top 5 verb collocates to the right (2R) as well as the top 5 collocates to the left (1L) of the modal verbs, sorted by the MI score (Mutual Information)[4], are provided in Tables 1 and 2 to give a sense of who *will/can* do what that has been emphasized in the data.

Table 1. Top 5 verb collocates of *will* and *can* (2 to the right)

will	*depend, enable, determine, continue, receive*
can	*assure, announce, afford, confirm, tell*

The top 5 verb collocates of *will* and *can* suggest two thematic threads running through the topics: assurance and commitment. It is fairly clear from the top right collocates of *can* that the modal verb is used mainly to offer assurance. As examples (10) and (11) demonstrate, the assurance is often connected to the government's efforts: increasing testing capability in (10) and delivering vaccination in (11). Examples (12) and (13) show that the modal verb *will* can also be associated with assurance through promising or expecting commitment: of the NHS to keep prioritising primary care in the vaccination program in (12) and of the public to respect social distancing rules in (13).

(10) I **can assure** you that the testing capability we have built in the last few weeks is world leading in its scale and sophistication (DSB200501)

(11) And I **can tell** you that this afternoon with pfizer and oxtra- oxford astrazeneca combined as as of this afternoon we've now vaccinated over 1.1 million people in England and over 1.3 million across the UK (DSB210105)

[4] MI is a measure of the strength of association between two words that takes into account the number of times they occur together and the number of times they occur separately in a corpus (see Hunston 2002 for an introduction to statistical measures used to determine collocational strength).

(12) and primary care are front and center stage in all of the immunization programs that the nhs has ever delivered and **will continue** to be so (DSB201116)

(13) we know that the vast majority of people **will continue** to act responsibly to control the spread of this virus (DSB200522)

Table 2. Top 5 left collocates of *will* and *can* (one to the left)

will	*we, that, it, there, they*
can	*we, you, I, they, people*

With regard to the subjects of the modal verbs, as seen in Table 2, those of *can* (*we*, *you*, and *people*) suggest a strong sense of engagement by including the audience (i.e., the public) in the utterances. Indeed, even in the case of *we*, which has both inclusive and non-inclusive uses, it is often the inclusive use that applies when it co-occurs with *can* in the corpus, emphasising what *we* (meaning everyone) can do to get through the crisis, particularly in relation to social distancing (with *can* ranked second on the high-probability list), as in example (14). In (15), three out of the four main measures are touched upon in relation to what the public *can* do (social distancing, taking a test, self-isolating). We can also see that the theme of 'assurance' is mixed with the sense of engagement in the last sentence: *we can* achieve the goal—to keep the transmission down, keep schools open and keep the economy moving—by following the measures outlined, which *we can* do. The co-occurrence of *together* with *we can* in (15), as well as *all* in (14), further strengthens the idea that everyone is in this together.

(14) but there are other things **we can all do** good hand hygiene being really aware of social distancing wearing a face mask on public transport (DSB200618)

(15) it's down to all of us ready to uh to to follow the guidance hands face space do the we haven't said that yet in this press conference hands face space do the essential stuff that you know uh makes sense take a test if you have symptoms self-isolate if you're contacted by by nhs test and trace that's the way **we can do it**

> together **we can get the r down**, keep kids in school, keep our schools open and keep our economy moving and bounce back more strongly (DSB201022)

Less common is the non-inclusive use of *we* when co-occurring with *can.* In the following example, *we* clearly refers to the government and *can* stresses upon its ability to act quickly if necessary.

> (16) and the instructions to people are clear if you get symptoms isolate immediately and get a test if you are contacted by NHS test and trace instructing you to isolate you must it is your civic duty so you avoid unknowingly spreading the virus and you help to break the chain of transmission this will be voluntary at first because we trust everyone to do the right thing but **we can** quickly make it mandatory if that's what it takes (DSB200527)

The subjects of *will* are not as inclusive as those of *can* suggest. As example (17) illustrates, *will* is often used to give assurance as to the positive effects of a given measure, among other things. When *we* co-occurs with *will*, it is mostly used in a non-inclusive way, referring to the government and its commitment, as exemplified in (18).

> (17) as I said at the very start the government moves to another combination of measures uh including track and trace uh **that will** help us keep the infection under control going forward (DSB200501)

> (18) **we will** be prioritizing uh care homes uh for those tests because it is a a truly uh tragic and very very difficult situation for for many people (DSB201020)

Occasionally, inclusive *we* is involved, not only to give assurance, but also to encourage people to 'follow the rules' as in*:*

> (19) the more people follow the rules then the faster **we will** all be through it (DSB200405)

Moving on to the delexical verbs shared by the topics, in the following examples (20)–(23), we can see that *get*, often in combination with a deverbal noun, forms clear and simple instructions or calls for action: *get a test*, *get that jab*, *don't get back into the old habits*. Each is associated with a different topic. Although (20) and (21) seem to be both about testing, *get a test* in the former comes together with *positive* and *self-isolate*, which are the key words for the Test and Trace procedure (Topic 10); the latter is an example of Topic 5 (Testing) where *get a test* is mentioned in connection with testing capacity. Examples (22) and (23) are related to Social Distancing (Topic 1) and Vaccine A (Topic 2), respectively.

(20) and why it's so important to to **self-isolate** if you if you **get a positive test** because that if that doesn't happen then uh obviously tests that a lot of the force of **test and trace** uh is is lost (DSB201022)

(21) at the beginning of last month at this podium I set a goal that anyone who needs a test should **get a test** and that as a nation we should achieve a hundred thousand tests per day by the end of the month (DSB200501)

(22) we must act and the most important thing for all of us is to remember the basics first wash your hands regularly and for 20 seconds **don't get back into the old habits** (DSB200909)

(23) and the way to ensure that this happens is to **get that jab** when your turn comes so let's **get the jab done** (DSB210318)

The other high-frequency verbs such as *come* and *make* occurring on the high-probability list for Topic 2 are also used in the same way. As in (24), word combinations with these high-frequency verbs (*come forward*, *get your first/second jab*, *make that appointment*) express strong, unambivalent messages as regards what the public should do to help control the virus.

(24) the most important thing to remember is please **come forward** and **get your second jab** if you are over 40 you know **make that**

> **appointment** in eight weeks not 12 weeks and then of course those who now are eligible for their first dose anyone over the age of 18 to **come forward** and **get your first dose** as well because that also offers you quite high levels of protection uh not against the infection as much as double dose but certainly against serious illness and hospitalization (DSB219623)

4.3 Key topics related to societal sectors

Figure 4 presents the top 20 high-probability words for the four chosen topics associated with societal sectors that seem to be the subject of attention in the data.

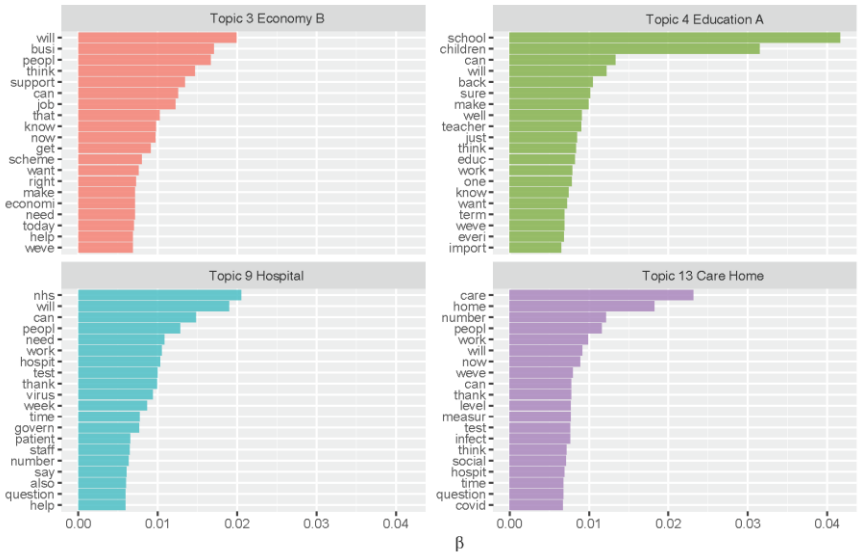

Figure 4. Top 20 word probabilities for Topics 3, 4, 9, and 13

There are fewer overlapping items across the topics relating to societal sectors than was found for those relating to measures (4.2); only the two modal verbs (*will*, *can*), which are also shared by the four topics related to measures, feature across all the topics here. The high-probability words are more informative as regards the common theme, e.g., *business*, *job*, (*furlough*) *scheme*, *economic* for Topic 3 (Economy); *school*, *children*,

back (to *school*), *teacher*, *education*, *term* for Topic 4 (Education); *nhs*, *hospital*, *patient*, *staff*, *virus* for Topic 9 (Hospital); *care*, *home*, *infection*, *social* for Topic 13 (Care Home).

Apart from these topic-related words, each topic features one or more words focusing on 'work' put into the sector in question as well as justification (e.g., *support*, *help*, *make sure*, *measure*, *test*, *right*, *important*), particularly in Topic 3, where we see three of them: *support*, *help*, and *right*. Examples (25) to (28) clearly demonstrate the central theme of the government's discourse regarding economy: to help and support businesses and workers of all kinds, and that it is *right* to do so.

(25) the Chancellor today has announced an extension of the furlough scheme which will **help** all businesses (DSB200512)

(26) in just a minute Rishi is going to explain how we're going to **help** workers of all kinds to get through this crisis **supporting** you directly in a way that government has never done before (DSB200320)

(27) over the last couple of months we have been working with industry on a plan to **support and help** people taking second jobs particularly those who are furloughed (DSB200519)

(28) it is quite a remarkable thing that we've done but it's **right** and we're protecting people's jobs we're protecting people's lives and peopl''s livelihoods and I think it's entirely **right** that we've done it (DSB200511)

For Education, as shown (29) to (31), it is *making sure* that the education is available for those in need during lockdown and that schools reopen safely as quickly as they possibly can. The word *important* often goes hand in hand with the promise to justify the decision.

(29) schools are open for them and we're working to **make sure** those who should attend do so (DSB200419)

(30) we recognize it's really **important** for the schools to be able to tailor their plan for their children to **make sure** it delivers the

maximum impact so those children catch up and really succeed (DSB200619)

(31) I think everybody understands that trying to get schools open in a way that is safe and only in a way that is safe that is really **important** for children's education (DSB200515)

The word *help* is also featured on the list of high-probability words for Topic 9 (Hospital). However, here it does not come from the government, but from the National Health Service (NHS). As (32) and (33) show, the key message is to encourage people who need medical help that is not COVID-related to also *come forward and seek help*.

(32) for anybody uh who needs uh **help** particularly emergency **help** then please come to us uh in the nhs (DSB200508)

(33) our message is that the NHS is open **help us to help you** so if you're worried about chest pains for instance maybe you might be having a heart attack or a stroke or you feel a lump and you're worried about cancer or you're a parent concerned about your child please **come forward and seek help** as you always would (DSB200427)

When it comes to Topic 13 (Care Home), the word *measure* on the high-probability list, together with two words on the FREX list (*control*, *monitor*) suggests the main efforts put into protecting care home residents as in (34) to (36):

(34) we're talking about beginning of April test capacity was limited to test a number of index cases to clarify that it was an outbreak of coronavirus then the whole care home would be treated as if the symptomatic cases were all coronavirus and and various **measures** are put in place (DSB200428)

(35) we've focused on the need to **control the spread of infection** in social care settings (DSB200415)

(36) but making sure that care homes have NHS internet access good internet access the digital tablets are absolutely revolutionary because it means that we can keep our residents and staff safe while we look after our care home residents uh through remote **monitoring** (DSB200515)

The word *test* occurs on the lists for Topics 9 (Hospital) and 10 (Care Home), which again shows the interrelatedness of the themes. As shown in (37) and (38), testing is also an effort applied to control the spread of the infection among hospital and care home staff. We see in the examples that *test* also goes together with the verb *get*. But unlike in (15), where *get a test* is a call for action, here the focus is to ensure sufficient testing capacity. Again, this shows that although some items occur on the lists of high-probability words for different topics due to the interrelation of these themes, a more systematic examination of the words and their use in context, which falls outside the scope of the present study, is needed to shed more light on the real substance of each topic.

(37) we're making sure that now NHS staff **get tested** including when they're asymptomatic to make sure that we understand whether whether the people who are working in hospitals have got the virus and using the tests for surveys (DSB200427)

(38) there has been a huge local effort both through local government our directors of public health our health protection teams but increasingly with support from national endeavors to **get the tests to** the home to make sure staff **get the testing** they need this will continue for some considerable time until we are convinced that we have got under this epidemic in the care homes which is of course of most concern (DSB200429)

4.4 Distribution of key topics over time

Figures 5 and 6 plot the chronological distribution of the two categories of topics discussed above (see 4.2 and 4.3). The theta value indicates the association between the topic and the corresponding document, which is defined by the date when the briefing took place (the covariate of date). The higher the theta value, the greater association between the topic and the corresponding document. The press briefings were given daily during

the first wave of the pandemic (16 March–5 June 2020) and then on a less regular basis: this is why the points are more tightly packed in the far left of the graphs.

As can be seen in Figure 5, the four topics related to measures are prioritized in different periods of the pandemic. Testing features prominently in the briefings during the first wave of the pandemic (April to May 2020). Test and Trace entered the scene towards the end of the first wave and seems to be the dominating topic until July 2020. Social Distancing and the basics were not put in a prominent position until sometime after Test and Trace came into the picture, but the topic was clearly emphasized to a great extent between the first and the second wave, during the summer of 2020, when the briefings were less frequent. During the second wave, Vaccine was clearly the most discussed measure. The other three measures still remained in the picture, being brought up from time to time. It seems that Social Distancing is more evenly distributed than the other two: while Vaccine seems to be given most weight during the second wave, the importance of Social Distancing and the basics have also been constantly reiterated.

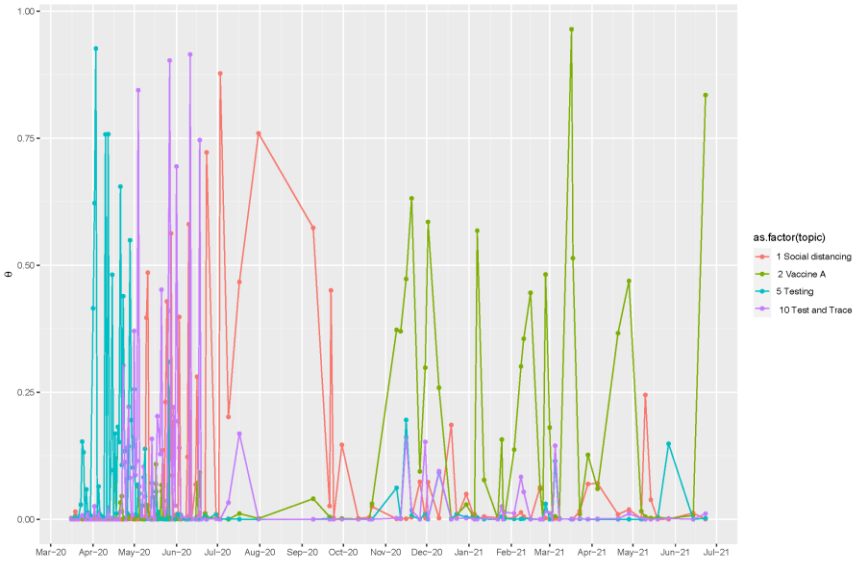

Figure 5. Topic distribution (1, 2, 5, 10) over time

With regard to the four topics related to societal sectors, in terms of proportions, Topics 13 (Care Home) and 9 (Hospital) are more sizable than Topics 4 (Education) and 3 (Economy) (see Figure 2). Figure 6 shows that the first two topics seem to have attracted much more attention in the first wave of the epidemic than in the second one. Education and Economy appear less regular in their distribution, but overall it seems that Education was brought up more frequently during the first wave and Economy during the second. In addition, both seem to increase either after or before the peak of each wave. None of these topics really stands out during the second peak (December-March), making the domination of the topic of Vaccine during this period (see Figure 5) even more pronounced.

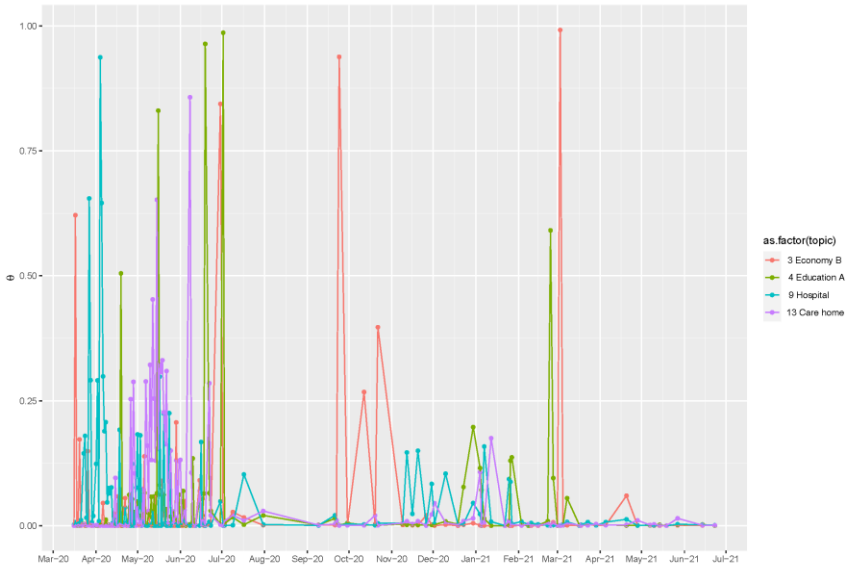

Figure 6. Topic distribution (3, 4, 9, 13) over time

5. Conclusion

Using the technique of Structural Topic Modelling (STM), the core topics in the UK government's press briefings on COVID-19 were identified. They seem to centre on two main strands of interest: measures undertaken to suppress the spread of transmission and issues related to societal sectors

of concern. Overall, the topics of the former strand are more prevalent than those associated with the latter. These topics reflect the government's priorities in handling the crisis: giving information/advice and addressing their concerns and needs in relation to not only how public health sectors are handling the situation but also other sectors such as education and economy that are badly affected by the pandemic. The results also show that the distribution of topics changes over time, with different topics being prioritised during different periods of the pandemic, seemingly in keeping with the increasing understanding of the disease and the development of medical and scientific work in tackling the pandemic.

The paper shows how the study of ERC can benefit from an empirical approach. The statistically-derived topics and their content help point towards significant words that may otherwise go unnoticed, particularly those high-frequency words which on the surface suggest nothing unusual. Using algorithms allows these words to emerge as significant, thereby prompting further examination. For example, words like *will, can, get, make* may not be something that one might have thought to look at closely in relation to ERC strategies. However, a close examination of the use of these words in context reveals some key strategies employed in the government's ERC practice (e.g., using simple, everyday language to deliver clear and strong messages about what actions to take, focusing on expressing assurance and commitment as well as encouraging compliance). Overall, these key words seem to have a positive prosody, with the 'risk' element, which is a central one of ERC suggested by WHO and ECDC, carefully embedded in the message.

Given the scope of the study and its exploratory nature, only a small selection of topics and words were subjected to close scrutiny. It is hoped that future research will follow up the initial findings of this study with a more systematic approach. In addition, the study only investigated one channel of communication (television). There are other channels, such as social media, that can be mobilised in ERC practice and are worthy of investigation from the same perspective. It would also be interesting to compare ERC strategies employed by different countries. This kind of comparison would be particularly relevant given that this is a new virus that has caused an unprecedented crisis for the whole world, and scientific knowledge both on the disease itself and as regards how to best tackle a pandemic of such a scale is still evolving. As a result, responses to the pandemic vary significantly across the globe, leading in turn to different

risk perceptions among the public, which play a crucial role in the success of policies adopted to suppress the transmission of such a highly infectious disease (Dryhurst et al. 2020). A comparison of ERC strategies across countries in relation to public risk perception can therefore offer important lessons for the future handling of similar crises. In that sense, this exploratory study opens up many avenues for future studies into ERC that could benefit from bridging areas of public health and linguistics research.

References

Anthony, Laurence. 2019. AntConc (Version 3.5.8) [Computer Software].

Biber, Douglas. 2010. Corpus-based and corpus-driven analyses of language variation and use. In *The Oxford handbook of linguistic analysis*, edited by Bernd Heine and Heiko Narrog, 159–192. Oxford: Oxford University Press.

Blei, David M. 2012. Probabilistic topic models. *Communications of the ACM* 55 (4): 77–84.

Brookes, Gavin, and Tony McEnery. 2019. The utility of topic modelling for discourse studies: A critical evaluation. *Discourse Studies* 21(1): 3–21.

Busso, Lucia, Marton, Petyko, Marton, Sarah Atkins, and Tim Grant. 2022. Operation Heron: Latent topic changes in an abusive letter series. *Corpora* 17(2): 225–258.

Culpeper, Jonathan. 2002. Computers, language and characterisation: An analysis of six characters in Romeo and Juliet. In *Conversation in life and in literature: Papers from the ASLA Symposium*, Association Suedoise de Linguistique Appliquee (ASLA), Vol. 15, edited by Ulla Melander-Marttala, Carin Ostman, and Merja Kytö, 11–30. Uppsala: Universitetstryckeriet.

Dreisbach, Jeconiah Louis, and Sharon Mendoza-Dreisbach. 2020. The integration of emergency language services in Covid-19 response: A call for the linguistic turn in public health. *Journal of Public Health* 43(2): 248–249.

Dryhurst, Sarah, Claudia R. Schneider, John Kerr, Alexandra L. J. Freeman, Gabriel Recchia, Anne Marthe van der Bles, David Spiegelhalter, and Sander van der Linden. 2020. Risk perceptions of COVID-19 around the world. *Journal of Risk Research* 23(7/8): 994–1006. doi: 10.1080/13669877.2020.1758193.

European Centre for Disease Prevention and Control. 2017. Public health emergency preparedness: Core competencies for EU Member States. Stockholm: ECDC.

Farrell, Justin. 2015. Corporate funding and ideological polarization about climate change. *PNAS* 113(1): 92–97.

Hunston, Susan. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Li, Yuming, Gaoqi Rao, Jie Zhang, and Jia Li. 2020. Conceptualizing national emergency language competence. *Multilingua* 39(5): 617–623.

Liu, Dilin, and Lei Lei. 2018. The appeal to political sentiment: An analysis of Donald Trump's and Hillary Clinton's speech themes and discourse strategies in the 2016 US presidential election. *Discourse, Context & Media* 25: 143–152.

Murakami, Akira, Paul Thompson, Susan Hunston, and Dominik Vajn. 2017. 'What is this corpus about?': using topic modelling to explore a specialised corpus. *Corpora* 12(2): 243–277.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. Structural Topic Models for open-ended survey responses. *American Journal of Political Science* 58(4): 1064–1082.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley. 2019. stm: R package for Structural Topic Models. *Journal of Statistical Software* 91(2): 1–40.

Scott, Mike, and Christopher Tribble. 2006. *Textual patterns: Key words and corpus analysis in language education*. Amsterdam/Philadelphia: John Benjamins.

Silge, Julia, and David Robinson. 2017. *Text mining with R: A tidy approach*. Sebastopol: O'Reilly Media.

Skalicky, Stephen, Nicholas D. Duran, and Scott A. Crossley. 2020. Please, please, just tell me: the linguistic features of humorous deception. *Dialogue & Discourse* 11(2): 128–149.

Taddy, Matt. 2012. On estimation and selection for topic models. *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics, PMLR* 22: 1184–1193.

Wallach, Hanna M., Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. *Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09)*: 1105–1112. New York: ACM.

World Health Organization. 2017. Communicating risk in public health emergencies: a WHO guideline for risk communication (ERC) policy and practice. World Health Organization. https://apps.who.int/iris/handle/10665/259807. License: CC BY-NC-SA 3.0 IGO.