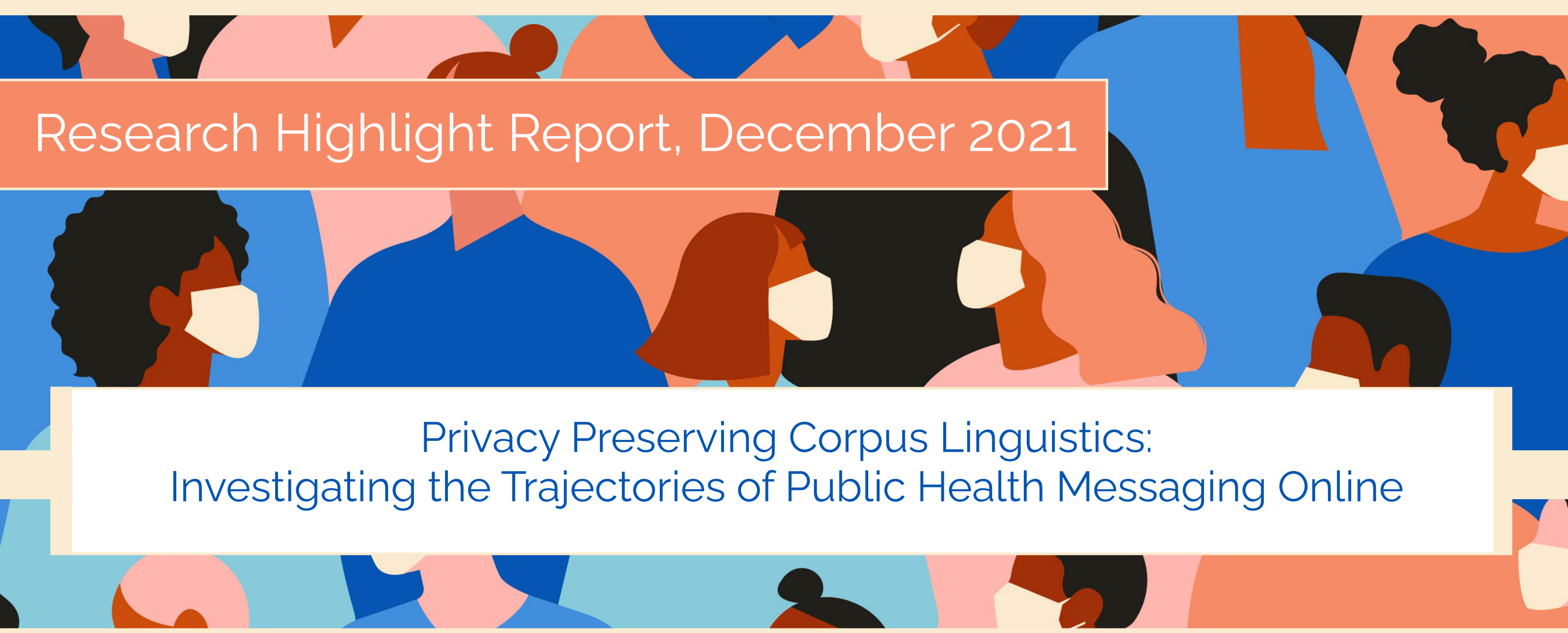


Research Highlight Report, December 2021

Privacy Preserving Corpus Linguistics: Investigating the Trajectories of Public Health Messaging Online

Emma McClaughlin, Elena Nichele, Svenja Adolphs, Pepita Barnard, Jeremie Clos,
Dawn Knight, Derek McAuley, Miriam Aydt, Tino Tom, Alexandra Lang

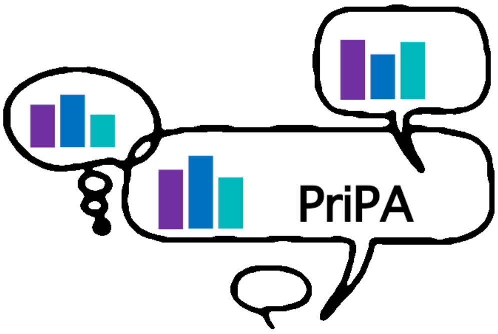
The Coronavirus Discourses project supports public health partners Public Health Wales, Public Health England, and NHS Education for Scotland in addressing key challenges that the coronavirus pandemic presents in terms of understanding the flow and impact of public health messages in public and private communications.

During the Coronavirus pandemic, health measures have varied across the UK, following geographical and social boundaries. We have developed a new approach that allows us to track messaging reception over time, i.e., their 'trajectories' once released to the public.

We are using a corpus linguistic approach, which involves the linguistic analysis of a digitised body of texts called a corpus (plural 'corpora'). Corpus linguistics uses specialist software to identify patterns that occur in language, including the frequency and usage of specific words.

Privacy Preserving Analytics

In this report, we outline a set of guiding principles for privacy-preserving research for researchers and professionals, which applies to a new approach we have developed, mainly relating to the development of PriPA (Privacy Preserving Analytics).



Next, we introduce the PriPA (Privacy Preserving Analytics) Extension. The PriPA extension is a digital tool designed for anyone to use on their personal devices. It safely retrieves information about individual language use for analysis. The advantage of this browser extension is that users have full control over what information they want to share.

C19COMMS

Coronavirus Discourses

The University of Nottingham in collaboration with Cardiff University is working in partnership with Public Health England, Public Health Wales, and NHS Education for Scotland to investigate linguistic evidence for effective public health messaging.

University of Nottingham
UK | CHINA | MALAYSIA

Cardiff University
Prifysgol Caerdydd

UKRI

Arts and Humanities Research Council

Guiding Principles for Privacy-Preserving Research

We have developed a set of guiding principles for privacy-preserving research. Some are drawn from existing ethical research guidelines and others have emerged from our work on developing the PriPA extension, including user testing.

1 Participants must be made aware of the motivations for carrying out the research. This information should be presented in a clearly understandable way.

2 Participants must be aware of the parameters for data collection. For our purposes this means informing the participants in advance of the data and methods we use: websites, language features (words and phrases) that form the basis of the analysis, and corpus linguistic analyses (frequencies, collocates and concordance lines) that present the data in a way that can be more readily analysed.

3 Descriptions of the features of interest must be provided in a clearly understandable way.

4 Participants should feel confident that they will not be identifiable should they decide to participate in the research.

5 Participants must have the option to view their results in full before choosing whether or not to share them with researchers.

6 Participants must have the option to exclude selected results before sharing.

7 Participants must have the option to fully delete results before they are shared.

8 No one except the participant may view data that they have decided not to share.

9 If any clearly identifiable personal information is willingly shared by the participant, it should be removed by the researchers to avoid storing unnecessary personal data.

10 Participants must have the option to download and save their results in full for their own personal use and reference. This applies regardless of whether or not they decide to share their results with researchers. The format must be accessible, and the presentation should be as intuitive as possible for participants to understand the textual representations.

11 Participants must be given the option to withdraw their results for a limited period after submission. This process must be straightforward for the participant and clear instructions must be provided by the researchers.

User Experience Testing

As part of the design process, we carried out user experience testing for the PriPA extension, gathering important insights on:

- The presentation of the PriPA extension
- Developing new privacy-preserving technologies
- Trust surrounding privacy-preserving tools
- What individuals need to know before, during and after participating in privacy-preserving corpus linguistic research

These elements are reflected in the design of the PriPA extension as outlined below.

Background to the Tool

We designed and built the PriPA browser extension to help us to understand more about:

1 Individuals' browsing activities – what information people are accessing about COVID-19 and from which (of a pre-defined list of) mainstream sources they are accessing it.

2 Individuals' language in use about COVID-19 – conversations happening on social media and other sites people login to.

Privacy-Preserving Research

Privacy-preserving research methods aim to gather insights on the research topic without compromising the privacy of the people or organisations the data came from. In practice, this means individuals can trust that their data is safe, secure, and their identity remains private.

Browser Extensions

Browser extensions are computer programs which run on Web browsers (e.g., Google Chrome). Browser extensions usually add tools or extra functions to the browser such as removing advertising from Web pages (adblockers), for example.

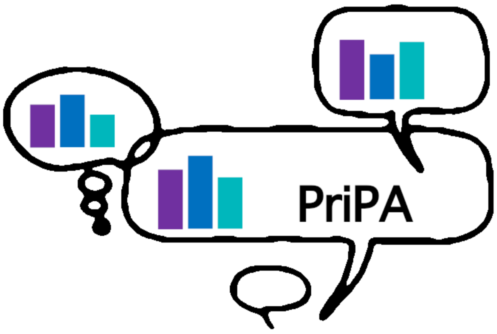
Data Protection and Ethical Research

Allowing participants to make an informed decision on which data they wish to share goes beyond what is expected by current data protection law, which allows for the (restricted) processing of personal data. We are not processing this data at all and want to normalise this approach, which comes at no cost to the research outcomes.

What language can PriPA analyse?

The PriPA extension analyses language from specific predefined websites (e.g., news websites, public health websites or social media sites) in real time as the user browses those websites.

For our project, this data is more useful than simply analysing all of the content of a news website, for example, because it helps us understand the kinds of language that people are actually seeing. We are interested in the representation of official public health guidance in the news – what information participants have been exposed to – and the reinterpretation of official health guidance by individuals as the pandemic progresses.



How Does the PriPA Extension Work?

At its core, the PriPA extension can be used to study a range of topics. Participants will install the browser extension from the Google Chrome Web Store to join a study. For people contributing data to the Coronavirus Discourses project, PriPA will focus on words and phrases surrounding Coronavirus and related issues identified by our public health partners (language surrounding specific social groups and geographical borders).

BAME	ETHNIC(ITY)	VACCINATE(D)	IMMUNITY	HOPE(S)
BLACK(S)	OF COLOUR	SHOT(S)	IMMUNISE(D)	HOPING
ASIAN(S)	VACCINE(S)	JAB(S)	IMMUNISATION	SAFE(TY)
MINORITY/IES	VACCINATION(S)	IMMUNE	DOSE(S)	COMMUNITY/IES

Examples of keywords used in the Coronavirus Discourses study

After participants install the extension in their browser and sign up to the project, PriPA downloads a file containing the details of the data to be collected (word frequencies, collocates, concordance lines), the language to be analysed (e.g., words relating to vaccines) and a list of websites where the analysis will be run.

Word Frequencies

The analysis of how often words feature in a language dataset. In this case, language seen by participants on their computer screens.

These basic statistics are used to calculate collocations in our study.

Collocates

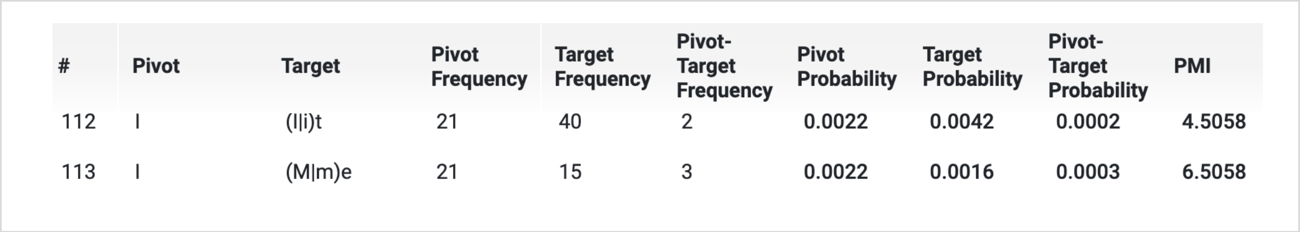
Words that most frequently co-occur in language use (e.g., salt + pepper).

We are interested in how combinations of words reveal patterns of meaning. For example, the word 'substance' appears to be neutral but it in reality it often co-occurs with words like 'harmful' and 'toxic', so substance takes on a negative meaning from its context. Other words might take on positive meaning from their context. Linguists call this phenomenon 'semantic prosody'.

Concordance Lines

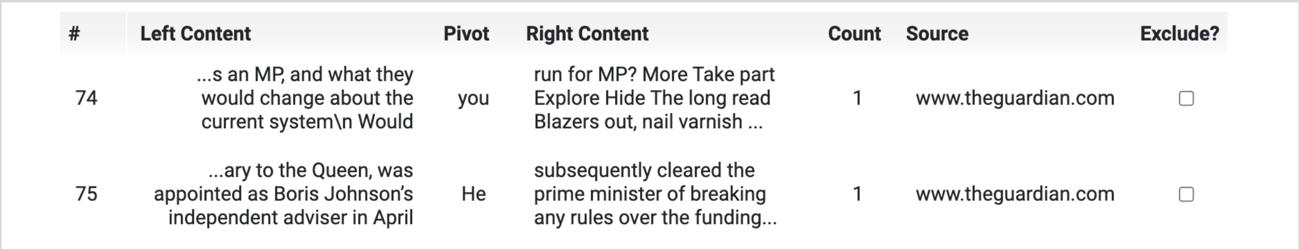
A line of text showing the linguistic context for a particular word (i.e., the words that (commonly) occur before and after a word).

Concordance lines give us extra context to explore patterns of meaning.



#	Pivot	Target	Pivot Frequency	Target Frequency	Pivot-Target Frequency	Pivot Probability	Target Probability	Pivot-Target Probability	PMI
112	I	(I|i)t	21	40	2	0.0022	0.0042	0.0002	4.5058
113	I	(M|m)e	21	15	3	0.0022	0.0016	0.0003	6.5058

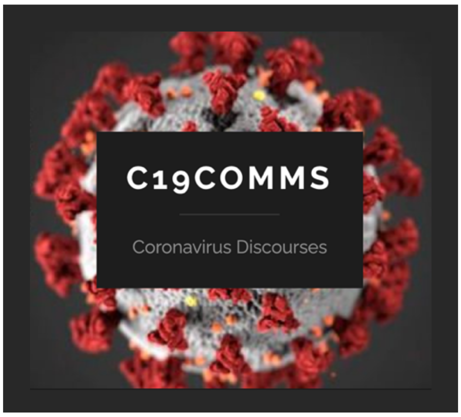
Example collocates for the word 'I' from a prototype version of PriPA – the collocates are 'it' and 'me'



#	Left Content	Pivot	Right Content	Count	Source	Exclude?
74	...s an MP, and what they would change about the current system\n Would	you	run for MP? More Take part Explore Hide The long read Blazers out, nail varnish ...	1	www.theguardian.com	☐
75	...ary to the Queen, was appointed as Boris Johnson's independent adviser in April	He	subsequently cleared the prime minister of breaking any rules over the funding...	1	www.theguardian.com	☐

Example concordance lines from theguardian.com for the words 'you' and 'he' from a prototype version of PriPA

Running the PriPA Extension and Contributing Results



After installation, the browser extension needs to be activated. PriPA will then run continuously in the background, only collecting statistics on language from the agreed websites when they are visited. For some phases of the study, this may include the processing of personal data and private conversations, but the user's privacy will be preserved as we only collect statistics on associations between words for a restricted list of words and anonymised concordance lines.

Participants will be able to view and scrutinise the analysis of their data before they submit it to the research team. Every participant can decide whether or not to submit data to the study, being fully aware of all the information that would be shared with the researchers.

Every participant can exclude specific results from the analysis before they either submit or delete their results, or they can withdraw from the study completely. If participants decide not to share their results with researchers on this study, the results are deleted and will not leave their computers.

Those interested in installing PriPA to contribute to the Coronavirus Discourses project should go to c19comms.wp.horizon.ac.uk for details on how to register their interest for the study. They will receive instructions by email if a suitable study is available at the time of registration. Alternatively, we may be able to put them on a reserve list for the next phase of the study.

Concluding Points

- The PriPA extension adheres to all the guiding principles outlined in this report. Future iterations of the browser extension will maintain or strengthen these privacy-preserving standards.

- Research carried out on the Coronavirus Discourses project using the PriPA extension will improve understanding of the trajectories of public health messages. This will help us to produce key outputs for writers of public health messages including a catalogue of specific messages and their trajectories over time and accompanying guidance to optimise the effectiveness of future messaging campaigns.

- The work reported here will further inform our investigation of the reception and evaluation of public health messaging and related measures which are of key concerns to our project partners Public Health England, Public Health Wales, and NHS Education for Scotland.

- In the future, the PriPA extension can be used as the basis for other studies investigating a whole range of topics that would benefit from gathering insights from browsing activities and private communications in a privacy-preserving way.

Banner imagery reproduced with permission from Shutterstock.
Any enquiries regarding this publication should be sent to:
Professor Svenja Adolphs (svenja.adolphs@nottingham.ac.uk)

University of Nottingham
UK | CHINA | MALAYSIA
CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD
UKRI
Arts and Humanities Research Council

The project Coronavirus Discourses: linguistic evidence for effective public health messaging is funded by the Arts and Humanities Research Council (AHRC), grant reference AH/V015125/1. It is led by the University of Nottingham in collaboration with Cardiff University, and working in partnership with Public Health England, Public Health Wales and NHS Education for Scotland.