

Using ATLAS.ti to interpret Keyword Co-occurrence Analysis: A case study on the representation of *vaccin\** across pseudoscience and conspiracy websites.

Yuze Sha and Isobelle Clarke  
Lancaster University

### **Abstract**

In this study we use ATLAS.ti to interpret the results of a Keyword Co-occurrence Analysis (KCA) of fake vaccination news. Specifically, KCA is used to uncover the most dominant patterns of co-occurring keywords across a corpus of 37,676 texts from 235 pseudoscience and conspiracy websites that mention *vaccin\**. KCA enables researchers to examine linguistic patterns of fake news from multiple angles, including discourse, register, style, and attitude. Yet, the interpretation of KCA can be time-consuming, especially when texts are long. Consequently, in this study, we leverage ATLAS.ti's Code Co-occurrence Analysis functionality, which streamlines and accelerates the interpretation of KCA results by providing access to extended concordances that highlight the patterns of keyword co-occurrence.

Taking the second most prominent dimension as a demonstration, we interpret this pattern of keyword variation across our *vaccin\** corpus as distinguishing texts that are questioning the COVID-19 pandemic, especially in relation to higher power control, from texts that are discussing childhood vaccines, especially with respect to the dangers they pose. The implications of these linguistic repertoires in relation to fake news and anti-science strategies are discussed.

**Keywords:** Keyword Co-occurrence Analysis, ATLAS.ti, anti-vaccine conspiracies, fake news, corpus-assisted discourse analysis

### **1. Introduction: Fake news, anti-vaccination discourse, anti-vaccination websites**

The rapid advancement of online communication technologies has expanded the public's daily access to a myriad of information sources. This influx of information can negatively impact the public's capacity to make rational decisions (Van Zandt 2004). This challenge is further impeded by the presence of fake news (Zhang and Ghorbani 2020). Fake news is deliberately

fabricated content that mimics the form of news media but lacks adherence to journalistic processes or intentions (Lazer et al. 2018).

In recent times, fake news has permeated various spheres, including politics (e.g., Subramanian 2017) and science, particularly concerning vaccination, a public health measure credited with preventing 4-5 million deaths annually (WHO 2024). Studies have highlighted the detrimental effects of vaccine misinformation, including the resurgence of vaccine-preventable diseases (e.g., measles) in many countries (Hotez, 2020). Consequently, ongoing research efforts aim to delineate the discursive characteristics of anti-vaccination discourse (e.g., Bean 2011; Hardaker et al. 2023), especially on social media (e.g., Maci 2019; Orlandi et al. 2022).

Although recent anti-vaccination studies have focused on social media, the impact of anti-vaccination websites remains substantial. These platforms act as primary sources of much information quoted across anti-vaccination online communities and social media posts. Studies (e.g., Betsch et al. 2012; Finney Rutten 2019; Fox 2011) have shown that individuals, especially patients and caregivers, consult the internet for health-related information, especially vaccination information. The Pew Internet & American Life Project (Fox 2011) found that eighty percent of Internet users seek health information online (Kata 2012). Among these seekers, a substantial seventy percent report that their findings on such health information websites influenced their treatment decisions.

With the capacity of websites to influence health decisions, studies have sought to understand anti-vaccination websites' content and persuasiveness (e.g., Bean 2011; Kata 2012; Moran et al. 2016; Sak et al. 2015). For example, in a content analysis of 480 websites, Moran et al. (2016) uncovered that 66.9 percent of the websites used pseudoscience as a persuasive strategy, such as confusing correlation for causation. 59.2 percent of websites referred to expert opinions to give weight to their statements and persuade their readers. In another study, Bean (2011) drew on the findings from Davies et al. (2002), Kata (2010) and Wolfe et al. (2002) who explored themes across anti-vaccination websites, to assess if the themes had evolved. Specifically, Bean (2011) used content analysis to analyse 25 anti-vaccination websites for recurring and changing emphases in content, design and credibility. The content features were summarised into four categories: safety and effectiveness, civil liberties, alternative treatments, and conspiracy theories/search for truth. Compared to findings from Davies et al. (2002), Kata (2010) and Wolfe et al. (2002), Bean (2011) found that whilst much had remained the same, there were some new themes in response to new emerging health trends and threats, such as the H1N1 outbreak. This study highlights the importance of revisiting the anti-vaccination

websites in the present study, around a decade after these studies, especially following the COVID-19 pandemic.

Like Bean (2011), many studies investigating anti-vaccination websites have employed content analysis, using human coders allocated with pre-defined code sets from earlier studies (e.g., Sak et al. 2015), or integrating these schemes with either a qualitative examination of data samples (e.g., Moran et al. 2016) or the emerging themes through an iterative examination process (e.g., Bean 2011). Whilst using human coders offers distinct advantages, such as uncovering subtle thematic variations, it also risks affecting the results' objectivity. Additionally, the process can be time-consuming, especially for large datasets, which may limit the scope of the analysis.

To address this, in the present study we applied the corpus-assisted discourse analytical approach, Keyword Co-occurrence Analysis (KCA) to anti-vaccination website texts to uncover groups of keywords that co-occur across them, which we systematically explore for themes, discourses, registers, styles, and attitudes.

## **2. Keyword Co-occurrence Analysis**

Keyword Co-occurrence Analysis (KCA) is aimed at uncovering the dominant patterns of keyword co-occurrence across the texts of a corpus (Clarke et al. 2021; Clarke et al. 2022). Keywords are terms appearing with unusual frequency compared to a reference corpus. Keywords are instrumental in highlighting the aboutness of the dataset, such as discourses (Baker 2004) and register (McEnery 2016). Yet one challenge when it comes to keyword studies is aggregation – the keywords in the keyword list may all point to the discourses, but prising apart the discourses is a task for the analyst (see Clarke et al. 2021 for a detailed discussion). In previous keyword studies, to interpret the keyword results, researchers often manually categorise keywords into semantic or thematic groups based on a close reading of corresponding concordances (e.g., Brookes 2022). While manual analysis offers depth, the categories created and the keywords assigned to the categories are susceptible to compromise, especially when corpora are large and when keywords occur frequently (Clarke et al. 2021). Instead, KCA uses a multivariate statistical technique, called Multiple Correspondence Analysis (MCA) to group the keywords based on their frequent co-occurrence across a corpus, aiming to deliver rich, multi-dimensional insights. KCA is based on the notion of linguistic co-occurrence – frequent patterns of co-occurring linguistic features are not random, but instead point to at least one shared communicative function (Biber 1988). Prior research employing

KCA has illuminated that patterns of keyword co-occurrence not only point to discourses and functions, but also sub-registers (Clarke et al. 2021), argumentative repertoires, and manipulative disinformation strategies (Clarke 2023). These applications of KCA have shown its capacity to account for the multiple senses, topics, (sub)registers, functions, and discourses that keyword co-occurrence can express.

KCA involves the following four broad steps: (1) compute keywords using a traditional keyword analysis (i.e., comparing the relative frequencies of the words in a target corpus to those in a reference corpus using a particular statistic of one's choice, e.g., log-likelihood, log ratio, difference coefficient), (2) analyse each text in the corpus for the occurrence of these keywords and record in a categorical data matrix, (3) subject the data matrix to MCA to reveal dimensions comprising the most common patterns of co-occurring keywords, and finally (4) interpret these dimensions of keyword co-occurrence, guided by the principles of linguistic co-occurrence (Biber 1988) and the indicative nature of keywords in discourse (Baker 2006).

Despite the method's strengths, the interpretation of dimensions in any dimension reduction method, such as MCA, is difficult, especially in the context of KCA where the variables are linguistic features, and the goal is to select a short, descriptive label that captures the crux of the dimension and the opposition of many features (Friginal and Hardy 2019). In previous KCA studies, analysts read texts most associated with each dimension and explored each keyword associated with the dimension in these texts to understand the relevant keywords' contexts and uses. After labelling the co-occurrence pattern, they attempt to falsify it against less associated texts following the same approach. Although effective, the interpretation process can be laborious, especially when texts are long, and dimensions comprise numerous keywords.

To address this, we explored technological solutions to expedite the interpretation process and found ATLAS.ti's code co-occurrence function to be complimentary for KCA. In the rest of the paper, we present Dimension 2 from a KCA of texts mentioning vaccination from pseudoscience and conspiracy websites to demonstrate how to use ATLAS.ti for analysing KCA results. The reason to skip Dimension 1 is because Dimension 1's results oppose long texts with short texts (see Clarke and Grieve 2019 for a more detailed description).

### 3. Methodology

#### 3.1. Vaccine sub-corpus of the Pseudoscience and Conspiracy Sources corpus

The data for this study comes from a larger project investigating different branches of anti-science (see Clarke 2023). The general corpus for this project comprises texts (all content on a single webpage – i.e., article and comments) from 235 websites labelled as “conspiracy-pseudoscience” by mediabiasfactcheck.com, which is a comprehensive and continuously updated resource of online media sites which have been rated for various levels of bias. The corpus was filtered by retaining texts according to “seed” words and phrases associated with the anti-science branches relevant to the larger project. The present study drew on the vaccination sub-corpus, which was filtered according to the seed words “vax” and/or “vaccin\*”, which spans 21 years (from 2000 to 2021). Duplicated texts were removed from the corpus using a Python script to avoid skewing the data. Table 1 presents the composition of the corpus before and after deduplication.

Table 1. Composition of Vaccination Sub-corpus

Vaccination sub-corpus	Number of texts	Number of words (tokens)
Before de-duplication	52111	62,449,596
After de-duplication	37921	31,941,747

Table 1 shows that nearly half of the anti-vaccination content is duplicated, demonstrating, like climate denial literature (Dunlap and Jacques 2013), that anti-vaccination content is recycled and reposted across other websites whenever convenient.

#### 3.2. Generation of keywords and MCA

Keywords were computed in Sketch Engine by comparing the vaccination corpus to the English 2020 web corpus (enTenTen20) using the simple maths method (N=100) (Kilgariff 2009) and capping the number of keywords to the top 1000 results (Kilgariff et al. 2014). We further reduced this list according to the keywords that were dispersed across more than 5% of the texts in the vaccination corpus, resulting in 177 keywords. Each text was then computationally analysed for the presence or absence of these 177 keywords, and this was recorded in a categorical data matrix. This matrix was then subjected to MCA in R using the ‘FactoMineR’ package (Husson et al. 2024). MCA produced a series of dimensions detailing

the most common patterns of co-occurring keywords across the corpus and which texts display those patterns (see Clarke et al. 2021 for a more detailed discussion). Specifically, the MCA assigned each text and each category of a keyword (e.g. *presence* of RNA, *absence* of RNA) a coordinate and contribution score for each dimension. Categories of keywords with contributions above the average contribution score on a dimension are the most important contributors to the dimension. All contributions for a particular dimension add up to 100, so the average contribution is 0.28 ( $100/(177 \text{ keywords, each with 2 categories, namely presence and absence}) = 100/354$ ). Coordinates indicate the nature of the association between the keywords in terms of proximity, where keywords that co-occur often across the texts of the corpus will have coordinates closer to each other on one side of an axis. Keywords with strong contributions and positive coordinates co-occur often together in many texts, while keywords with strong contributions and negative coordinates co-occur often together in a different set of texts with each set rarely or never co-occurring with the other set. Thus, a dimension represents a pattern of keyword variation. We interpreted these MCA results in ATLAS.ti by (1) creating subcorpora comprising the texts most associated with each dimension (2) creating codes aligned with the keywords most associated with each dimension, and (3) using the code co-occurrence function to observe paragraphs in the texts where the keywords co-occur. This facilitated a more systematic and expedited visualisation of keyword co-occurrence in texts by pointing to paragraphs where the keywords most strongly associated with each side of the dimension co-occur rather than searching each text one keyword at a time.

### **3.3. Corpus construction on ATLAS.ti**

#### **3.3.1. Creating the subcorpora**

To build our subcorpora in ATLAS.ti, we selected the top 50 texts most associated with the positive and the negative side of each dimension. These 100 texts were then imported into ATLAS.ti and we used the “Group” function to categorise them into two subcorpora based on their associative polarity (e.g., [Dimension 2\_positive] & [Dimension 2\_negative], see Figure 1). These texts represent the most prototypical texts of the discourse (or shared function, etc.), tending to include many, if not all, of the keywords most strongly associated with the particular pole of the dimension.

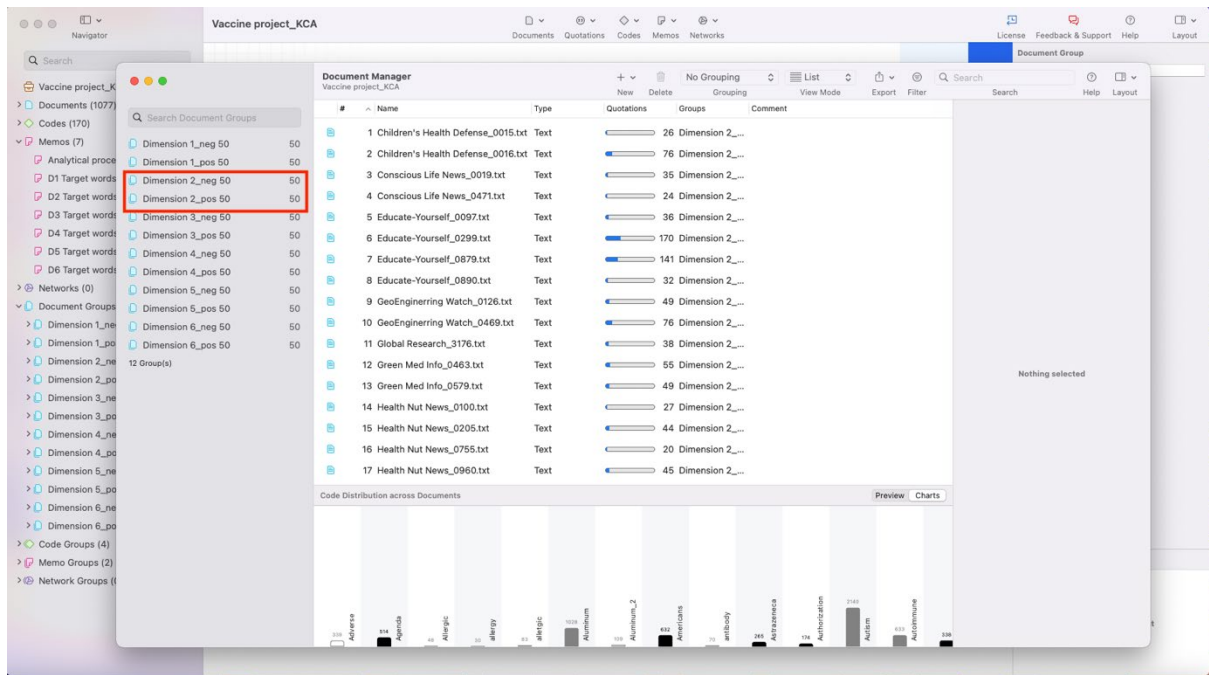


Figure 1. User interface of ATLAS.ti

### 3.3.2. Creating the codes

We then created codes based on the keywords most associated with each pole of the Dimension from the MCA results. Table 2 shows the keywords that are contributing above the average contribution (ctr) for Dimension 2 and their respective coordinate (coord).

Table 2. The keywords most strongly contributing to positive and negative Dimension 2 (\_P for Presence; \_A for Absence)

	Dim.2 coord	Dim.2 ctr		Dim.2 coord	Dim.2 ctr
sars-cov-2_P	1.676	1.869	rubella_P	-1.822	1.31
lockdowns_P	1.564	2.018	pertussis_P	-1.725	1.372
plandemic_P	1.555	0.993	mumps_P	-1.63	1.27
wuhan_P	1.542	2.518	tetanus_P	-1.552	0.948
pcr_P	1.532	1.042	mmr_P	-1.481	1.647
distancing_P	1.527	1.4	aluminum_P	-1.343	1.179
fauci_P	1.492	2.149	pediatrics_P	-1.318	0.779
variant_P	1.461	0.878	hepatitis_P	-1.317	1.031
lockdown_P	1.392	1.666	infant_P	-1.307	0.987
ma_P	1.372	0.817	mercury_P	-1.208	1.094
spike_P	1.292	0.981	autism_P	-1.197	2.077
mna_P	1.264	1.545	childhood_P	-1.142	1.546
moderna_P	1.264	1.487	pediatric_P	-1.124	0.553
mask_P	1.224	1.412	measles_P	-1.123	1.223
quarantine_P	1.191	0.716	gardasil_P	-1.119	0.839
coronavirus_P	1.174	4.289	nvic_P	-1.11	0.546
authorization_P	1.168	0.706	parental_P	-1.072	0.514
pandemic_P	1.099	3.65	hpv_P	-1.052	0.933
passport_P	1.087	0.478	disorder_P	-0.988	0.7
biden_P	1.061	1.017	merck_P	-0.942	0.624
astrazeneca_P	1.012	0.593	neurological_P	-0.921	0.558
covid-19_P	1.003	4.514	toxicity_P	-0.909	0.358
covid_P	0.979	3.183	exemption_P	-0.876	0.4
pfizer_P	0.952	1.287	polio_P	-0.875	0.508
jab_P	0.927	0.813	autoimmune_P	-0.866	0.481
tyranny_P	0.916	0.643	chronic_P	-0.82	0.678
gates_P	0.884	1.082	child_P	-0.699	1.203
trump_P	0.843	1.047	immunization_P	-0.567	0.401
fake_P	0.8	0.627	syndrome_P	-0.565	0.335
conspiracy_P	0.8	0.569	brain_P	-0.558	0.521
elderly_P	0.75	0.421	injury_P	-0.533	0.439
infected_P	0.721	0.723	covid-19_A	-0.435	1.96
experimental_P	0.716	0.839	children_P	-0.41	0.846
virus_P	0.714	2.031	coronavirus_A	-0.312	1.14
copyright_P	0.689	0.313	covid_A	-0.283	0.921
propaganda_P	0.628	0.45	pandemic_A	-0.282	0.936
agenda_P	0.585	0.424	virus_A	-0.262	0.746
deadly_P	0.576	0.426			
outbreak_P	0.568	0.445			
viral_P	0.563	0.394			
americans_P	0.492	0.562			
infection_P	0.463	0.449			
children_A	0.211	0.435			

We employed the Text Search feature (see Figures 2 & 3) of ATLAS.ti to pinpoint and code paragraphs within the texts most associated with Dimension 2 that contained the target keywords.



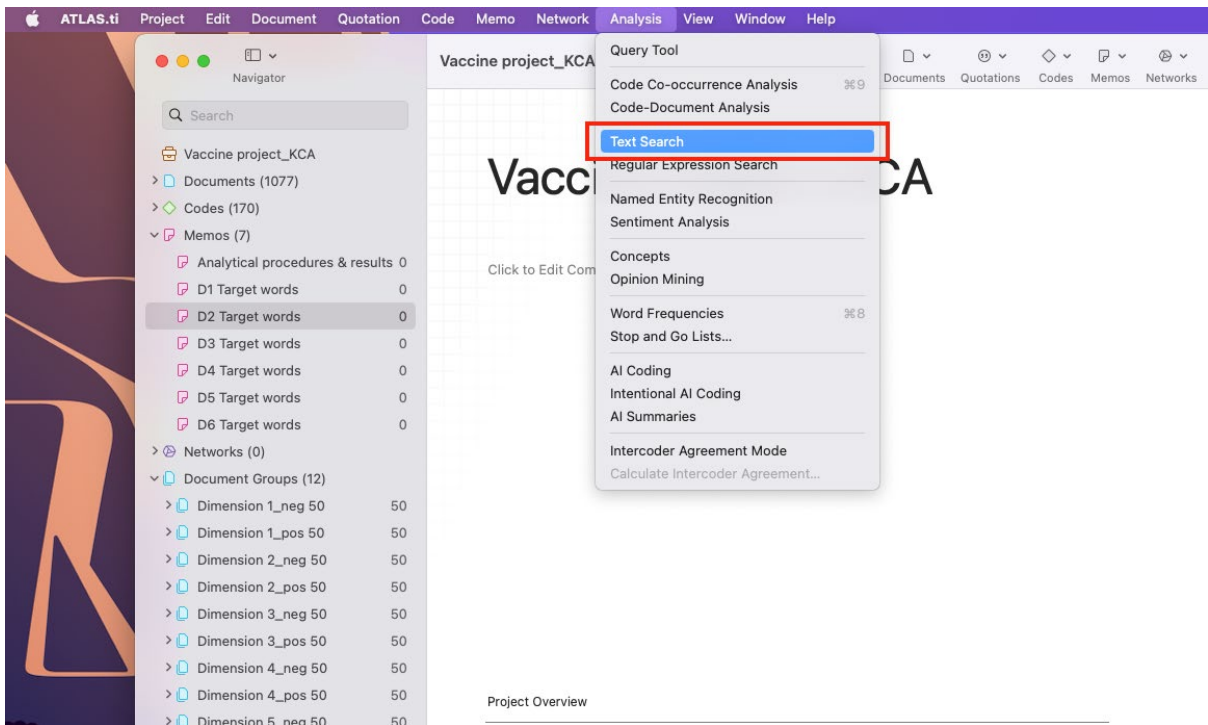


Figure 2. Text Search Functionality on ATLAS.ti

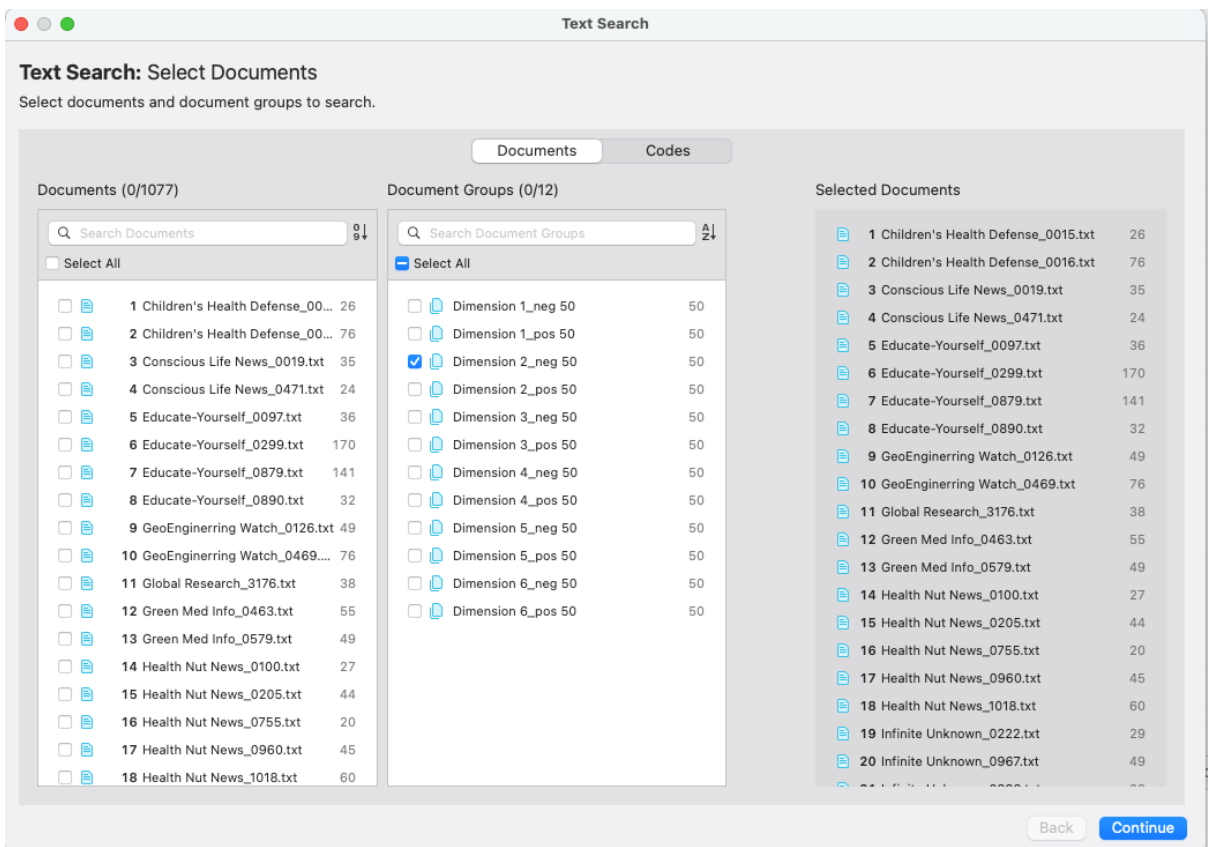


Figure 3. Selecting target document (groups)

Subsequently, we entered the target keyword for coding and set the query's scope. Different from previous studies (e.g. Clarke et al. 2021), our interpretation of the dimensions of

keyword co-occurrence concentrated on how the keywords co-occurred in individual paragraphs (see Figures 4 & 5) rather than entire texts for the purpose of accelerating the interpretation process. This approach enables us to isolate specific segments within the most strongly associated texts where the keywords associated with a particular side of a dimension appear together, facilitating a more detailed examination of the factors contributing to their co-occurrence.

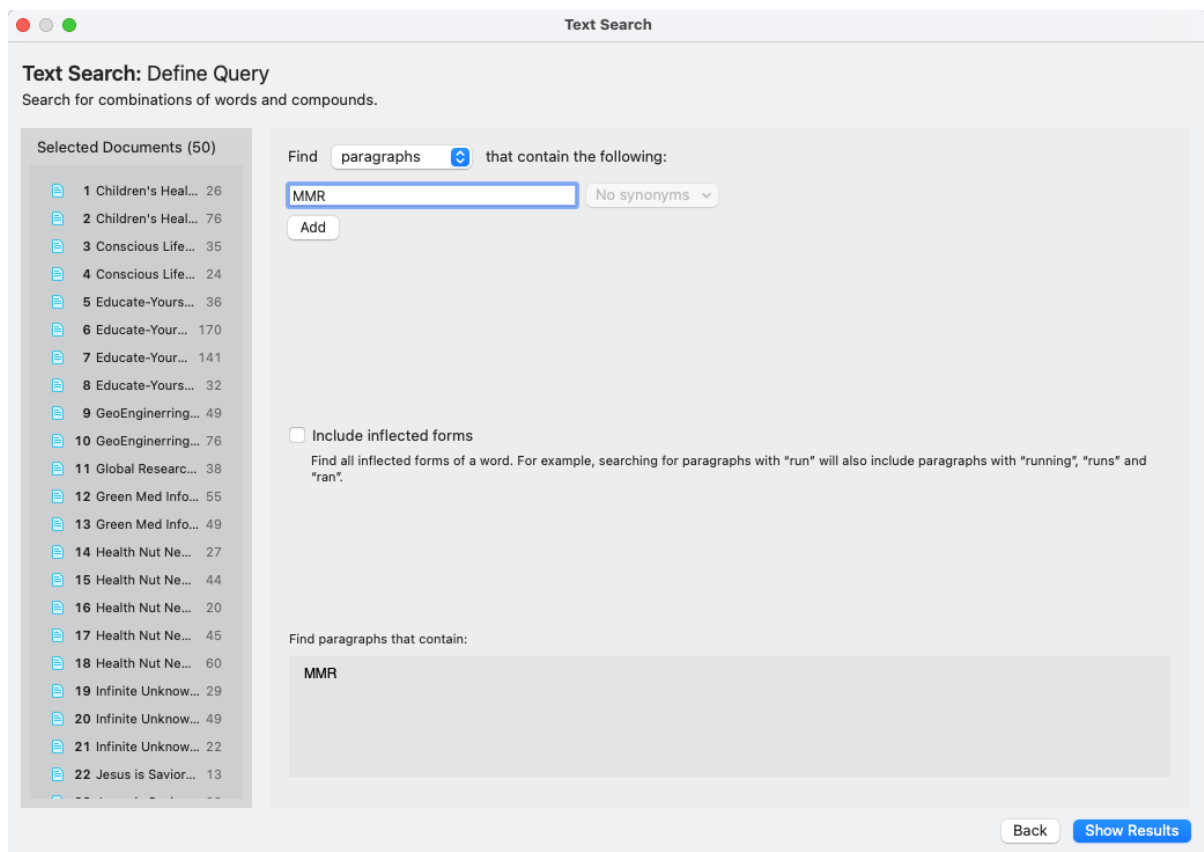


Figure 4. Defining query

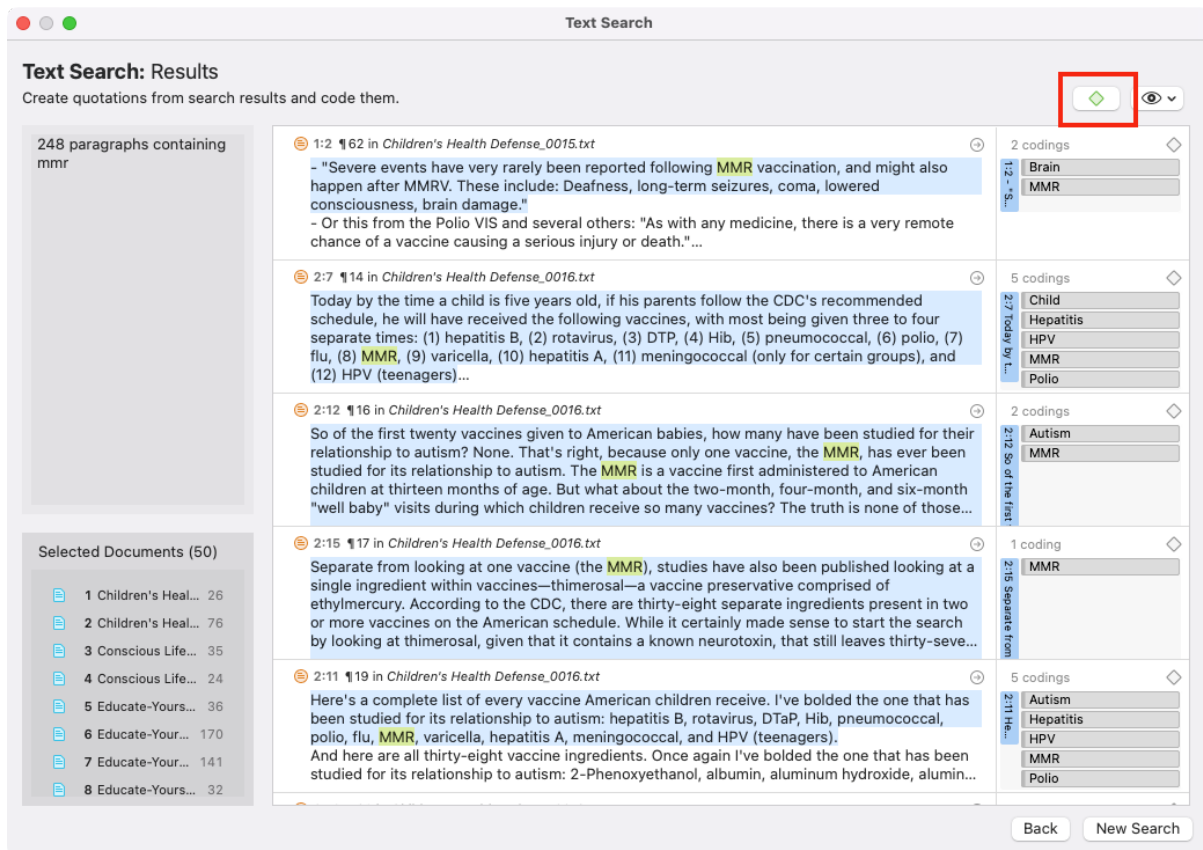


Figure 5. Results & bulk coding on ATLAS.ti

For each side, we then used ATLAS.ti's bulk code function (top right side in Figure 5) to mark every occurrence of each keyword within the top 50 texts, respectively. Once the coding process is completed, the instances of keywords within paragraphs are marked, thereby enabling the subsequent Code Co-occurrence Analysis.

### 3.3.3. Analytical framework

After constructing and annotating our corpus, our objective, as with other KCA studies, was to delineate what the patterns of keyword co-occurrence point to. To guide this interpretation, we used the analytical framework established in Clarke et al. (forthcoming), which outlines five preliminary areas of inquiry (see Table 3).

Table 3. KCA Analytical Framework

Construct	Definition	
Topic or subject matter	The subject matter/aboutness of the text.	What do the texts concern? What are the texts about?
Discourse	“[S]et[s] of meanings, metaphors, representations, images, stories, statements and so on that in some way together produce a particular version of events” (Burr, 2015: 74-75).	Are the patterns of co-occurring keywords being used in texts to focus on a particular event and/or aspect? If so, what is the event or aspect? How is vaccination being represented? What aspect of vaccination is being zoomed in on?
Register	A variety of language associated with both a particular situation of use and with pervasive linguistic features that serve important functions within that situation of use (Biber and Conrad, 2009: 33). “Registers are described for their typical lexical and grammatical characteristics [...] and also [...] for their situational contexts, for example whether they are produced in speech or writing, whether they are interactive, and what their primary communicative purposes are” (Biber and Conrad, 2009: 6). The function of linguistic features in the situational context.	How are the keywords functioning in the texts? Are the keywords characteristic of a particular language variety? What is/are the purpose(s) of the texts? Do the texts share a specific or primary communicative purpose? Are all the texts a particular register?
Style	The use of linguistic features that reflect aesthetic preferences, associated with particular authors or historical periods (Biber and Conrad, 2009: 18).	Do the keywords reflect aesthetic preferences of particular authors/historical periods?
Vaccine/Evolution/GMO attitude	The vantage point of vaccines and vaccination expressed in the texts.	Are the texts overtly pro- or anti-vaccination? Are the texts disinterested in vaccination?

The interpretation process began by using the Global Filter function (see Figure 6) in ATLAS.ti to isolate the target Dimension 2 subcorpora for examination.

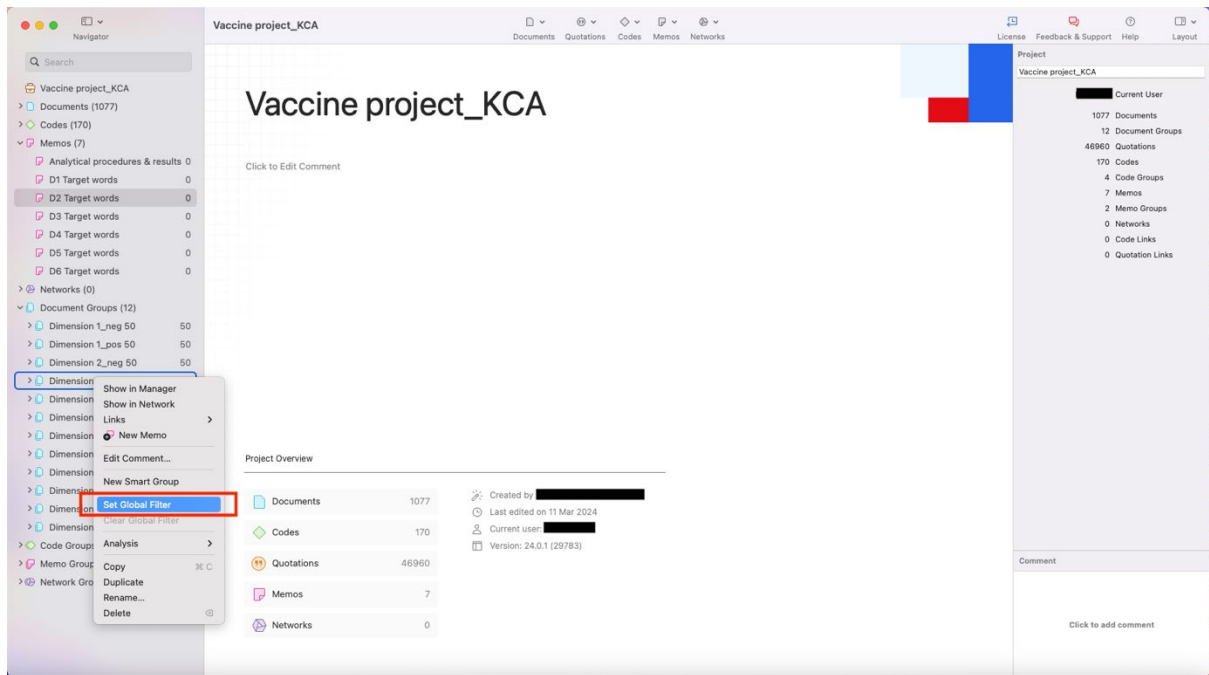


Figure 6. Setting Global Filter on ATLAS.ti

We then utilised the Code Co-occurrence Analysis function in ATLAS.ti to analyse and summarise the patterns of co-occurrence throughout the subcorpus (see Figure 7).

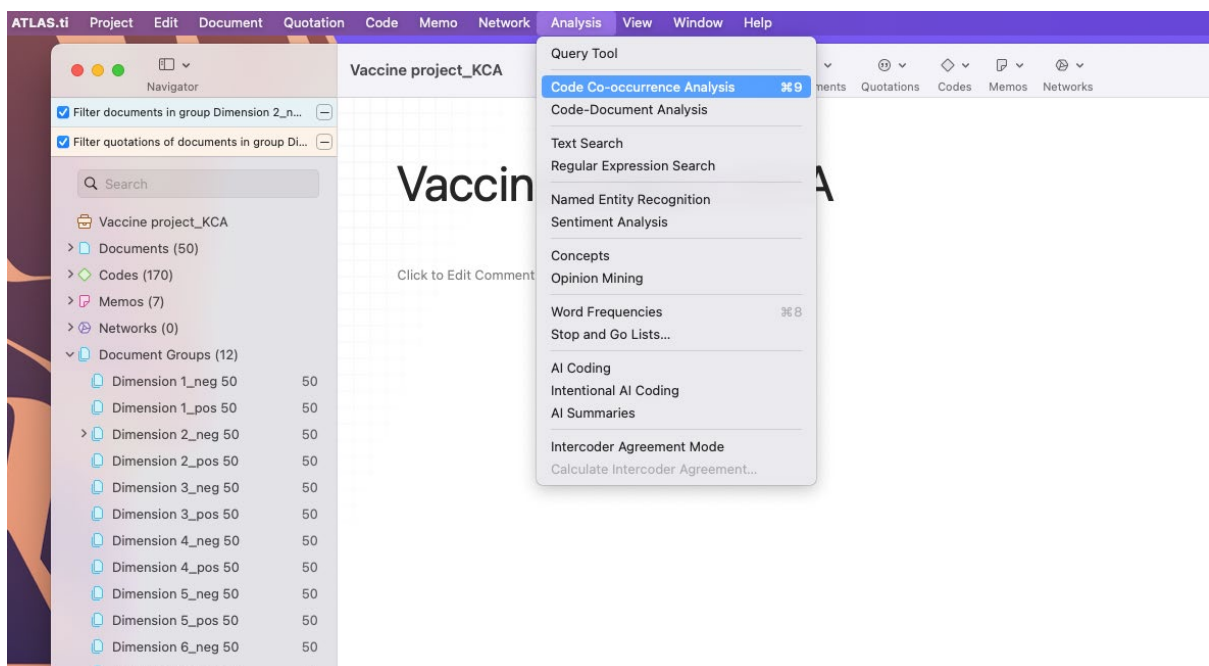


Figure 7. Code Co-occurrence Analysis on ATLAS.ti

Figure 8 displays the table of results from this analysis. The frequency with which two codes (representing keywords in this study) co-occur in the same paragraph is displayed in the middle. The intensity of colouring indicates the strength of co-occurrence within the

paragraphs of this subcorpus, with deeper colours signifying stronger associations. By selecting a specific column, the right side of the table reveals detailed concordances of these co-occurrences.

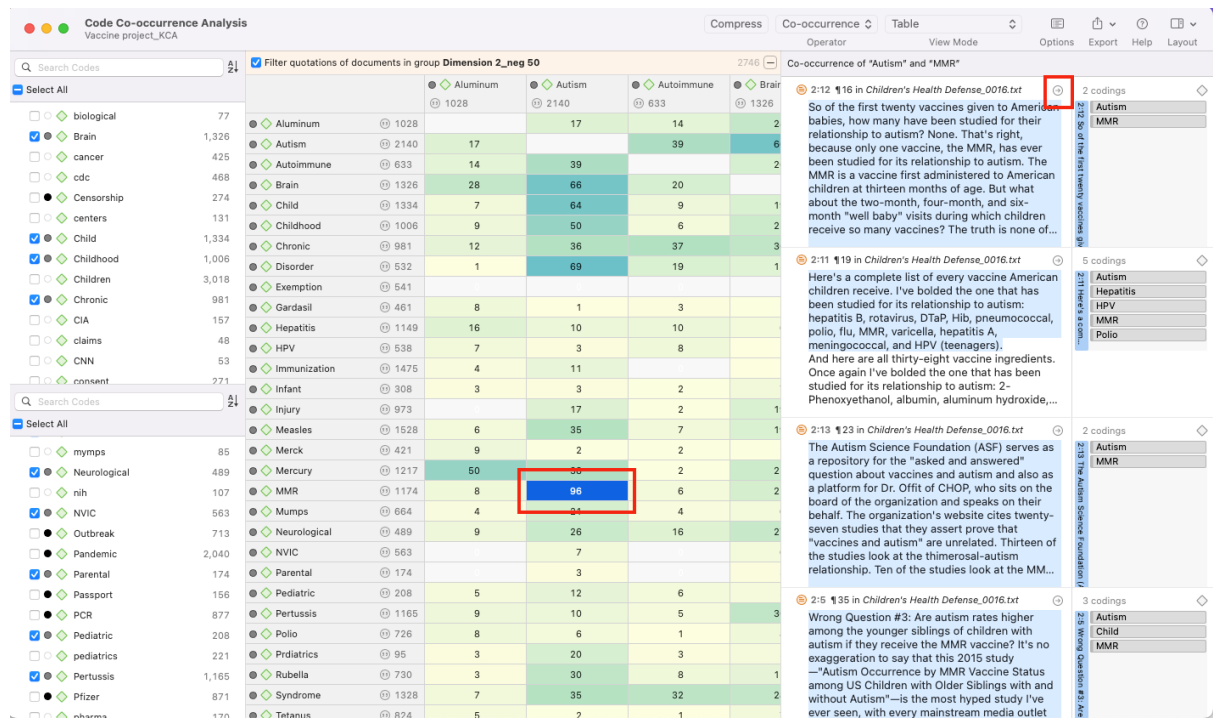


Figure 8. Code Co-occurrence Analysis Table

To identify paragraphs where more than two keywords co-occurred, we used the “Global Filter” function to initially filter concordances that have been coded with specific keywords. Subsequently, we used Code Co-occurrence Analysis to explore their co-occurrences with other keywords. For instance, to explore how the keywords associated with negative Dimension 2 (as presented in Table 2) co-occur in texts we set “Dimension 2\_neg 50” (the document group) and *MMR* (one of the target keywords) as the “Global Filter” criteria (see Figure 9). We then explored the Code Co-occurrence Analysis table to view the co-occurrence of *MMR* with *autism* (another target keyword) and all other keywords associated with the negative side of Dimension 2 (see Figure 10). We repeated this for all keywords strongly contributing to each dimension.

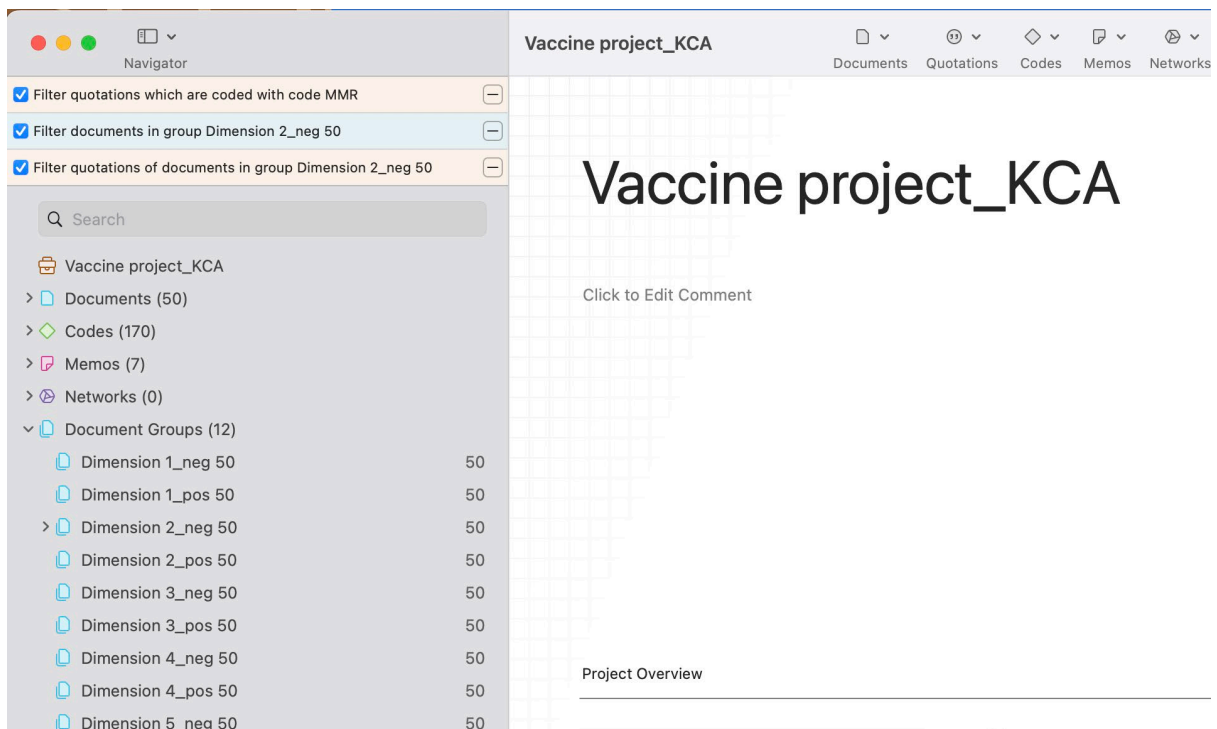


Figure 9. Co-occurrence analysis of more than two keywords (Global Filter Setting)

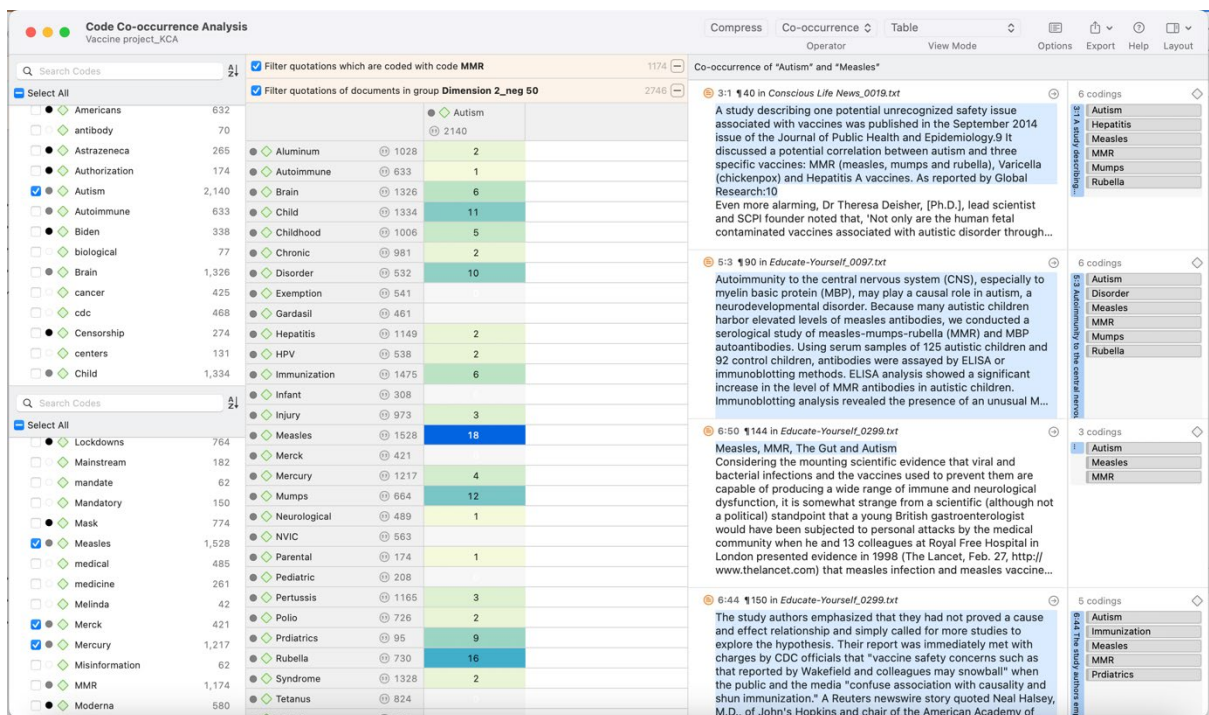


Figure 10. Co-occurrence analysis of more than two keywords (Co-occurrence Table)

#### 4. Results

We present our interpretation of Dimension 2 in this paper. It should be noted that whilst the interpretations and the concordances presented below are based on the top 50 most prototypical

texts, these patterns were observed in less strongly associated texts. After we had interpreted the top 50 texts, we sought to falsify our interpretations by exploring a random set of texts that were less strongly associated with the particular pole of the dimension. If the interpretations were falsified we refined the interpretation and repeated the process of falsification until no more refinement was needed.

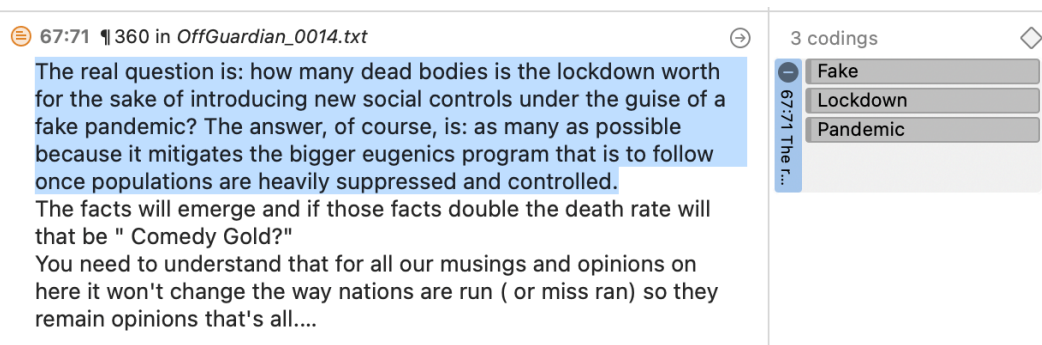
## 4.1. Positive Dimension 2

The keywords most associated with the positive side of Dimension 2 co-occur in texts to discuss the COVID-19 pandemic and question the legitimacy of government regulations related to the pandemic, including those concerning the COVID-19 vaccine.

### 4.1.1. Questioning the legitimacy of government regulations

A prominent representational discourse found across positive Dimension 2 texts concerns governmental control and regulations during COVID-19. Table 2 shows that many keywords strongly contributing to positive Dimension 2 are related to COVID-19 (*COVID*, *COVID-19*, *coronavirus*, and *sars-cov-2*), and COVID-19 related policies (*lockdown*, *mask*, and (social) *distancing*). Additionally, names of prominent political figures like *Biden*, *Trump*, and *Fauci* and keywords related to government actions, such as *agenda*, *authorization* are prevalent. These keywords are used to question the legitimacy and reasoning behind governmental interventions, including vaccination campaigns, accusing the government of a sinister agenda. Also, keywords such as *fake* and *experimental* frequently co-occur with both policy- and virus-related keywords often to suggest that the pandemic is not real, as illustrated in (1).

(1)



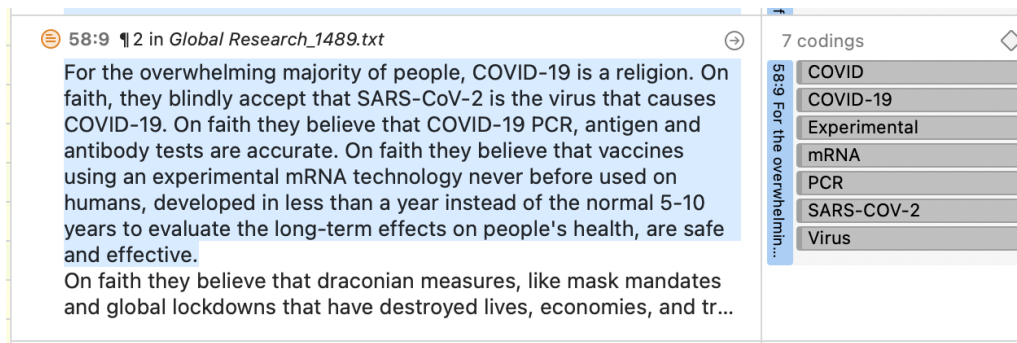
The screenshot shows a text editor window with a file named 'OffGuardian\_0014.txt'. The text in the editor is: 'The real question is: how many dead bodies is the lockdown worth for the sake of introducing new social controls under the guise of a fake pandemic? The answer, of course, is: as many as possible because it mitigates the bigger eugenics program that is to follow once populations are heavily suppressed and controlled. The facts will emerge and if those facts double the death rate will that be "Comedy Gold?" You need to understand that for all our musings and opinions on here it won't change the way nations are run (or miss ran) so they remain opinions that's all...'. A blue highlight covers the first sentence. On the right, a sidebar titled '3 codings' shows a list of tags: 'Fake', 'Lockdown', and 'Pandemic'. The 'Fake' tag is selected and highlighted in grey.



#### 4.1.2. COVID vaccination

COVID vaccination is a prominent theme. This is realised by keywords related to different COVID vaccine types (*Moderna*, *mRNA*, and *Pfizer*). Notably, these vaccine-related keywords often co-occur with the keyword *experimental* to directly describe the vaccine in phrases like “experimental mRNA technology” (see (2)), “experimental test vaccine”, or “experimental gene therapy mRNA drugs”, rather than simply referring to it as a “vaccine”. Such texts describe the vaccines as being hastily developed and question their need, safety, and efficacy. Notably, the reference to mRNA vaccine as “experimental gene therapy” is used to suggest that the vaccine is altering people’s genetic code and poses damage to individuals’ health. Referring to the vaccine as “experimental” contributes further to the discourse of governmental control as those who get the vaccine are positioned as test subjects.

(2)



The screenshot shows a text editor window with a highlighted paragraph of text. To the right of the text, a list of codings is displayed, including COVID, COVID-19, Experimental, mRNA, PCR, SARS-COV-2, and Virus.

58:9 ¶ 2 in Global\_Research\_1489.txt

For the overwhelming majority of people, COVID-19 is a religion. On faith, they blindly accept that SARS-CoV-2 is the virus that causes COVID-19. On faith they believe that COVID-19 PCR, antigen and antibody tests are accurate. On faith they believe that vaccines using an experimental mRNA technology never before used on humans, developed in less than a year instead of the normal 5-10 years to evaluate the long-term effects on people's health, are safe and effective. On faith they believe that draconian measures, like mask mandates and global lockdowns that have destroyed lives, economies, and tr...

7 codings

- COVID
- COVID-19
- Experimental
- mRNA
- PCR
- SARS-COV-2
- Virus

Despite references to different types of COVID-19 vaccines, the only vaccination reference found in the keyword list was *jab*. The keywords *vaccine* or *vaccination* were not strongly associated with positive or negative Dimension 2. Whilst this is most likely because they are used fairly equally across the texts associated with the positive and the negative sides of Dimension 2 and thus do not contribute to this pattern of variation, the strong association of *jab* alongside COVID-19 related keywords introduces meaningful connotations. Unlike the more medically oriented and neutral terms *vaccine* and *vaccination*, *jab* carries a more informal tone with violent connotations. The selection of *jab* over the other choices might also aim to cast the COVID-19 vaccination in a more negative or forceful light, contributing to amplified scepticism or reluctance towards COVID-19 vaccination initiatives.

Furthermore, this linguistic choice may serve as a mechanism of delegitimation, attempting to weaken the discourse’s connection to authoritative narratives (see (3)). By avoiding formal medical terminologies, it might serve to reduce the credibility and legitimacy of vaccination efforts and foster doubt, fear of injury, and diminish trust in scientific expertise and authority.

(3)

54:17 ¶16 in Blacklisted News\_0184.txt

With only 37% of the public sheeplike enough to get the jab, any "woke" business that continues to invoke mask mandates or require proof of vaccination will find their profits dissipate faster than Trump's lead in swing states at 3:00 am. If you were to believe the fake news media and government drones, you would believe the majority have been vaccinated and the "anti-vaxxers" were a crazy conspiracy theorist minority. As with most things being fed to you daily, this is a big fat lie. The rational, resistant (not hesitant), critical thinking MAJORITY are done with this scamdemic. Biden and his handlers are being forced to throw in the towel, for now.  
The propaganda press polluters of the truth are still selling a completely false narrative about the vaccines being the reason cases, hospita

5 codings

- Biden
- Conspiracy
- Fake
- Jab
- Mask

#### 4.1.3. Negative consequences

Another prominent discourse stresses the negative consequences of governmental controls during the COVID-19 pandemic, including COVID-19 vaccinations. This narrative is underscored by the co-occurrence of the keywords *elderly* and *deadly* with policy-related keywords, such as *lockdown(s)*, *quarantine* and *experimental* (vaccines), which are used in texts often to dispute the need for such interventions by (1) blaming the high infection and death numbers among the elderly as a direct consequence of government interventions, such as claiming that systems for elderly care collapsed due to lockdowns, or (2) accusing the COVID-19 death rates of being inflated due to the susceptibility of vulnerable populations to infections or death during the flu season, rather than as a direct consequence of COVID-19 (see (4)).

(4)

67:64 ¶310 in OffGuardian\_0014.txt

In actuality the coronavirus is highly contagious, however, 80% of the population who contracts it are not seriously affected. Another small percentage get intense flu-like symptoms, and a very tiny percentage who are infected die. These are for the most part the elderly and those with multiple comorbidities. Of course this is horrific, but this is the same demographic who frequently get pneumonia during a virulent flu season and often die. These facts would NOT change no matter who is president. What's also undeniable is that COVID-19 like all other viruses diminish once herd immunity is established. This is the only solution.  
Having said that, when a virus is no longer perceived as a medical issue, but as mechanism for political expediency logic is thrown out the window and is replaced w

5 codings

- Coronavirus
- COVID
- COVID-19
- Elderly
- Infected

Conspiracies about the adverse effects of the COVID-19 vaccination are also promoted in positive Dimension 2 texts. For instance, (5) claims COVID-19 illnesses and deaths, especially those of “the weak and *elderly*”, are not associated with the virus, but the vaccine.

(5)

89:1 ¶28 in *The Truth Seeker\_1646.txt*

The manufacturers, leading medical journals and most governments insist these deaths are unrelated to the vaccine. In many instances, the deaths and serious illness have been attributed to coincidental infection with the virus. But evidence is mounting that for some, especially the weak and elderly, the vaccine itself is creating or worsening the very illness against which it is supposed to be protective....

"...a worrying phenomenon which appears consistently in Covid vaccine studies is a spike in purported 'infections' which occurs precisely during that three-week period, and usually immediately following the jab...The researchers raise the possibility that the jab may trigger 'symptoms likened to Covid-19 symptoms including fever' in those recently exposed to the virus... He suggests the me

3 codings

- Elderly
- Infection
- Virus

## 4.2. Negative Dimension 2

By contrast, the keywords most strongly associated with negative Dimension 2 co-occur in texts that are focused on childhood vaccinations and the hazardous substances within them, which they claim cause numerous adverse effects.

### 4.2.1. Childhood vaccination

The keywords associated with negative Dimension 2 reference children (*child, childhood*) and childhood vaccinations (*measles, mumps, and rubella, polio, pertussis, and tetanus*). Many texts also include the keyword *Merck*, a pharmaceutical company. Such texts accuse Merck of being irresponsible for not conducting long term safety tests to highlight concerns regarding the quality of vaccines (e.g., Gardasil), as illustrated in (6).

(6)

10:26 ¶85 in *GeoEngineering Watch\_0469.txt*

"There is too little long term safety and efficacy data, especially in young girls, and too little labeling information on contraindications for the CDC to recommend Gardasil for universal use, which is a signal for states to mandate it," said Fisher. "Nobody at Merck, the CDC or FDA know if the injection of Gardasil into all pre-teen girls – especially simultaneously with hepatitis B vaccine – will make some of them more likely to develop arthritis or other inflammatory autoimmune and brain disorders as teenagers and adults. With cervical cancer causing about one percent of all cancer deaths in American women due to routine pap screening, it was inappropriate for the FDA to fast track Gardasil. It is way too early to direct all young girls to get three doses of a vaccine that has not been

5 codings

- Autoimmune
- Brain
- Gardasil
- Hepatitis
- Merck

Additionally, the keyword *pediatric* is often used to cite studies from pediatric journals and associations, like *Pediatric Annals* in (7), to lend professional credibility to their claims. Importantly, while the study mentioned exists, the quote discusses the etiologies of autism, but the study does not corroborate the connection between vaccines and autism that the website asserts.

(7)

6:107 ¶ 120 in *Educate-Yourself\_0299.txt*

In 1984 in *Pediatric Annals*, Ritvo and Freeman described the medical model of autism and concluded, "The symptoms are due to neuropathology which, in turn, may have a variety of etiologies," observing that there is a high rate of abnormal EEG's, seizures, severe allergies, and significant differences in brain metabolism patterns and brain chemistry in autistic children compared to those who are not autistic.

3 codings

- Autism
- Brain
- Pediatric

#### 4.2.2. Hazardous substances

Negative Dimension 2 texts also emphasise the presence of hazardous substances in vaccinations through keywords like *aluminum*, *mercury* to assert that they can cause various health issues (*toxicity*), including injuries (*injury*), diseases (*autoimmune*, *neurological*), and disorders (*autism*). These texts question the safety of the ingredients in childhood vaccines with phrases like “vaccine-induced autism” encapsulating these concerns. Many texts dispute scientific claims that vaccines do not cause autism by suggesting that there have been limited studies investigating the impact of these aforementioned substances in other vaccines (see (8)).

(8)

18:19 ¶ 38 in *Health Nut News\_1018.txt*

"The Jain [Lewin Group] study only looked at MMR. Media reports about this study have falsely and deceptively asserted that the Jain study shows that "vaccines" in general do not cause autism. In reality, the Jain study says nothing about other vaccines. The MMR vaccine is the only vaccine that has been much studied in relation to autism, and all of the MMR-autism studies suffer from HUB. The other likely more dangerous aluminum-containing vaccines, given at younger ages, have hardly been studied at all. It is a blatant lie to claim that the science shows "vaccines" in general do not cause autism.

The science actually shows the opposite. Controlled animal experiments o...

3 codings

- Aluminum
- Autism
- MMR

A common narrative throughout negative Dimension 2 texts asserts that vaccinated children face higher risks and suffer from more health issues than their unvaccinated counterparts. An illustrative case is provided in (9), where the “Children’s Health Defense” website quotes “Dr. Daniel Neides of the Cleveland Clinic” to imply that vaccines cause children to develop neurological disorders, including Autism and ADHD.

(9)

2:1 ¶ 192 in *Children's Health Defense\_0016.txt*

Specifically, vaccinated children were found to have a fourfold higher likelihood of having autism. I'm reminded of a quote by Dr. Daniel Neides of the Cleveland Clinic who wondered if we were making trade-offs that aren't worth it. He said, "Some of the vaccines have helped reduce the incidence of childhood communicable diseases [like chickenpox and pertussis from the study above]. That's great news. But not at the expense of neurologic diseases like autism and ADHD increasing at alarming rates."

3 codings

- Autism
- Childhood
- Pertussis

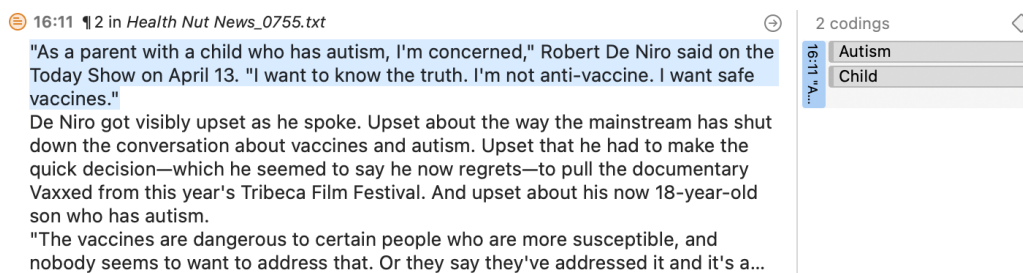
### 4.3. Addressing the remaining interpretation angles

So far, we have explored the keyword co-occurrence patterns through the lens of topic and discourse. We now turn to the remaining interpretation angles, as detailed in Table 3. From a register perspective, positive Dimension 2 is characterised by an informal, argumentative register (see (4)) through texts which question governmental policies (see (1)) and comprise colloquial references to vaccinations (e.g., *jab*) (see (3)). In contrast, negative Dimension 2 texts are more academic, featuring scientific references to substances and quotes from research studies and experts (see (7), (8) & (9)).

Regarding style, positive Dimension 2 is distinguished by political critiques of COVID-19 policies, reflecting a more provocative and contentious style. By contrast, negative Dimension 2 uses (pseudo)scientific and “evidence-based” arguments, suggesting a more analytical style.

The final aspect examines attitudes towards vaccinations. We found evidence of negative attitudes towards vaccinations on both the positive and negative sides of Dimension 2. Yet, importantly, there was also evidence of actors within- and authors of-texts outright denying being anti-vaccination, as can be seen in (10) below. Such texts nevertheless continue to call into question the safety of vaccinations, which in effect casts doubt on vaccinations and contributes to an anti-vaccination strategy. Rather than being anti-vaccination, they state that they are anti-unsafe vaccinations. This demonstrates that anti-vaccination is deemed by some as being “anti-cure” or “anti-antidote” and when this sense is evoked, those accused of being anti-vaccination will deny this label.

(10)



16:11 ¶ 2 in Health Nut News\_0755.txt

"As a parent with a child who has autism, I'm concerned," Robert De Niro said on the Today Show on April 13. "I want to know the truth. I'm not anti-vaccine. I want safe vaccines."

De Niro got visibly upset as he spoke. Upset about the way the mainstream has shut down the conversation about vaccines and autism. Upset that he had to make the quick decision—which he seemed to say he now regrets—to pull the documentary Vaxxed from this year's Tribeca Film Festival. And upset about his now 18-year-old son who has autism.

"The vaccines are dangerous to certain people who are more susceptible, and nobody seems to want to address that. Or they say they've addressed it and it's a...

2 codings

- Autism
- Child

## 5. Discussion and Conclusions

In this study, we demonstrated the application process of ATLAS.ti for interpreting the results of a Keyword Co-occurrence Analysis (KCA) of texts mentioning vaccination from websites known to promote pseudoscience and conspiracy theories.

Due to length restrictions, it was not possible to present all dimensions of keyword variation. But by delving into the second strongest pattern of keyword variation (i.e., Dimension 2), our analysis unveiled a dichotomy between discussions of COVID-19 vaccines and those on childhood vaccinations. Texts mentioning COVID-19 vaccines positioned them under the broader discourse of governmental regulations and control. Such texts were focused on questioning the need for government interventions, like lockdowns, mask wearing, and vaccinations, and promoting the conspiracy of an alternative sinister agenda. Texts delegitimised COVID-19 policies, including vaccination policy from two angles by: (1) stressing the safety of the unvaccinated by downplaying the virus's severity, and (2) highlighting the risks to the vaccinated by overstating the adverse effects of vaccines. The delegitimation is further achieved through the informal use of *jab* for *vaccine*, which could evoke concerns about safety and efficacy by distancing itself from the scientific term and register. The register of these texts is predominantly informal and argumentative, characterised by political critiques.

Texts discussing childhood vaccines are more “academic” with frequent citations from researchers and doctors, and the use of technical terminology related to hazardous substances and associated illnesses. Yet, paradoxically, these texts also include emotional appeals, with many texts directly calling on parents to protect their children against alleged vaccine-induced diseases, disorders, and deaths.

Many of these discourses and strategies are aligned with previous research investigating anti-vaccination websites, such as Bean (2011), which noted the frequent mentions of vaccine ingredients, vaccine-induced diseases and deaths, and accusing vaccines of violating civil liberties. Yet there are some differences, especially within texts covering the COVID-19 pandemic. For example, unlike the “diseases have declined” narrative found in the websites examined in Bean (2011), the COVID-19 vaccination discussions minimise the severity of the virus by accusing the death and illness statistics as being inflated due to the elderly and the vulnerable. Also, rather than solely stressing the mandatory nature of vaccination (Bean 2011), the COVID-19 anti-vaccination discourses posit vaccinations within the framework of government control, delegitimising the vaccination alongside other policies, such as lockdown and mask wearing, amplifying the scope of its target audience who disagreed with or disliked such regulations. These differences particularly in COVID-19 vaccine discourse thus point to the adaptive nature of anti-vaccination discourses.

In this study, we have illustrated how ATLAS.ti's code co-occurrence analysis function is complimentary to KCA. Using ATLAS.ti we were able to specify the context for

codes to co-occur as paragraphs as opposed to full texts. This enabled the observation of patterns of keyword co-occurrence more systematically rather than manually searching for the keywords in the full texts associated with the dimension.

Our results have pointed to some of the ways in which fake news may mimic authentic news, such as through references to experts, genuine citations, technical terminology, and political critique (Lazer et al. 2018). But, as shown, this is coupled with additional strategies like overstating and downplaying, which can add to the challenge of distinguishing fake news. Moreover, some texts exploit vague language, prompting their readers to “fill in the gaps”. For instance, by claiming that the COVID-19 pandemic is fake and that the government interventions are not aimed at preventing the spread of the vaccine but are instead part of a vague, unspecified agenda, readers can create what that agenda is and their own reasons for that agenda. Essentially, fake news can thus be moulded by the reader, making it considerably difficult to be distinguished from real news.

The present study also reveals the influence of COVID-19 on anti-vaccination discussions. Even though our dataset spanned 21 years, COVID-19 pandemic emerged to be dominant within our corpus. Future research should therefore continue to track the evolution of anti-vaccination websites’ strategies to better equip the public to delineate fact from fiction.

## Research Funding

This research was funded by the Leverhulme Trust. Grant number ECF-2020-590.

## References

- Baker, Paul. 2004. Querying keywords: Questions of difference, frequency, and sense in keywords analysis. *Journal of English linguistics*, 32(4), 346-359.
- Baker, Paul. 2006. *Using corpora in discourse analysis*. Bloomsbury Publishing.
- Betsch, Cornelia, Noel Brewer, Pauline Brocard, Patrick Davies, Wolfgang Gaissmaier, Niels Haase, Julie Leask, Frank Renkewitz, Britta Renner, Valerie Reyna, Constanze Rossmann, Katharina Sachse, Alexander Schachinger, Michael Siegrist & Marybelle Stryk. 2012. Opportunities and challenges of Web 2.0 for vaccination decisions. *Vaccine*, 30(25), 3727-3733.
- Bean, Sandra. 2011. Emerging and continuing trends in vaccine opposition website content. *Vaccine*, 29(10), 1874-1880.
- Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge University Press.
- Biber, Douglas & Susan Conrad. 2009. *Register, Genre and Style*. Cambridge University Press.
- Brookes, Gavin. 2022. ‘Lose weight, save the NHS’: Discourses of obesity in press coverage of COVID-19. *Critical Discourse Studies*, 19(6), 629-647.
- Burr, Viven. 2015. *Social Constructionism* (3<sup>rd</sup> edition). Routledge.
- Clarke, Isobelle, Elena Semino, Zsófia Demjén, William Dance, Tara Coltman-Patel & Richard Gleave. *forthcoming*. HPV Vaccine Discourse Online: A Corpus Linguistic Approach. Routledge.

- Clarke, Isobelle, Gavin Brookes & Tony McEnery. 2022. Keywords through time: A study of representations of Islam in the British press. *International Journal of Corpus Linguistics* 27(4): 399-427.
- Clarke, Isobelle & Jack Grieve. 2019. Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018. *PloS ONE* 14(9): e0222062.
- Clarke, Isobelle, Tony McEnery & Gavin Brookes. 2021. Multiple Correspondence Analysis, newspaper discourse and subregister: A case study of discourses of Islam in the British press. *Register Studies* 3(1): 144-171.
- Clarke, Isobelle. 2023. The discourses of climate change across conspiracy and pseudoscience websites. In S. Maci, M. Demata, P. Seargeant and M. McGlashan (eds.) *The Routledge Handbook of Discourse and Disinformation*, pp. 325-341. Routledge.
- Davies, Paul, Simon Chapman & Julie Leask. 2002. Antivaccination activists on the world wide web. *Archives of disease in childhood*, 87(1), 22-25.
- Dunlap, Riley, & Peter Jacques. 2013. Climate change denial books and conservative think tanks: Exploring the connection. *American Behavioral Scientist*, 57(6), 699-731.
- Finney Rutten, Lila, Kelly Blake, Alexandra Greenberg-Worisek, Summer Allen, Richard Moser & Bradford Hesse. 2019. Online health information seeking among US adults: measuring progress toward a healthy people 2020 objective. *Public Health Reports*, 134(6), 617-625.
- Fox, Susannah. 2011. 80% of internet users look for health information online. *Pew Internet & American Life Project*.
- Friginal, Eric & Jack Hardy. 2019. From factors to dimensions: Interpreting linguistic co-occurrence patterns. In Tony Berber Sardinha & Marcua Pinto (eds) *Multi-Dimensional Analysis: Research Methods and Current Issues*, Chapter 7. Bloomsbury.
- Hardaker, Claire, Alice Deignan, Elena Semino, Tara Colman-Patel, William Dance, Zsófia Demjén, Chris Sanderson & Derek Gatherer. 2023. The Victorian anti-vaccination discourse corpus (VicVaDis): construction and exploration. *Digital Scholarship in the Humanities*, fqad075.
- Hotez, Peter. 2020. Combating antisience: Are we preparing for the 2020s? *PLoS Biol* 18(3): e3000683
- Husson, Francois, Julie Josse, Sebastien Le, & Jeremy Mazet (2024). Package 'FactoMineR'. Available from: <<https://cran.r-project.org/web/packages/FactoMineR/FactoMineR.pdf>>
- Kata, Anna. 2010. A postmodern Pandora's box: anti-vaccination misinformation on the Internet. *Vaccine*, 28(7), 1709-1716.
- Kata, Anna. 2012. Anti-vaccine activists, Web 2.0, and the postmodern paradigm—An overview of tactics and tropes used online by the anti-vaccination movement. *Vaccine*, 30(25), 3778-3789.
- Kilgariff, Adam. 2009. Simple maths for keywords. In *Proc. Corpus Linguistics*, Liverpool, UK.
- Kilgariff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubiček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý & Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, 1(1), 7-36.
- Lazer, David, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts & Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380), 1094-1096.
- Maci, Stefania Maria. 2019. Discourse strategies of fake news in the anti-vax campaign. *Lingue Culture Mediazioni-Languages Cultures Mediation (LCM Journal)*, 6(1), 15-43.
- McEnery, Tony. 2016. Keywords. In *Triangulating methodological approaches in corpus linguistic research* (pp. 20-32). Routledge.
- Moran, Meghan, Melissa Lucas, Kristen Everhart, Ashley Morgan & Erin Prickett. 2016. What makes anti-vaccine websites persuasive? A content analysis of techniques used by anti-vaccine websites to engender anti-vaccine sentiment. *Journal of Communication in Healthcare*, 9(3), 151-163.
- Orlandi, Ludovico, Gianluca Veronesi & Alessandro Zardini. 2022. Unpacking linguistic devices and discursive strategies in online social movement organizations: Evidence from anti-vaccine online communities. *Information and Organization*, 32(2), 100409.
- Sak, Gabriele, Nicola Diviani, Ahmed Allam & Peter Schulz. 2015. Comparing the quality of pro-and anti-vaccination online information: a content analysis of vaccination-related webpages. *BMC Public Health*, 16, 1-12.



- Subramanian, Samanth. 2017. Inside the Macedonian Fake-News Complex. Wired. <https://www.wired.com/2017/02/veles-macedonia-fake-news/>
- Sun, Xinmei. 2020. Anti-vaccine discourse under the COVID-19 context: A corpus-assisted discourse analysis. In: University of Nottingham.
- Tan, Andy, Chul-joo Lee & Jiyoung Chae. 2015. Exposure to health (mis) information: Lagged effects on young adults' health behaviors and potential pathways. *Journal of Communication*, 65(4), 674-698.
- Van Zandt, Timothy. 2004. Information overload in a network of targeted communication. *RAND Journal of Economics*, 542-560.
- WHO. 10 facts on immunization [Internet], WHO [cited 2024 March 23<sup>rd</sup>]. Available from: <<http://www.who.int/features/factfiles/immunization/en/>>.
- Wolfe, Robert, Lisa Sharp & Martin Lipsky. 2002. Content and design attributes of antivaccination web sites. *Jama*, 287(24), 3245-3248.
- Zhang, Xichen & Ali Ghorbani. 2020. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2), 102025.