Corpus Building Project Report

Media Perspectives on COVID-19 Vaccine in New Zealand and Australia

Group E

1. Background

As the COVID-19 pandemic continues to impact communities worldwide, vaccine play a crucial role in mitigating the spread of the virus and reducing its impact on public health. In the early stages of COVID-19 vaccine rollout, faced with the novelty of the vaccines, people had uncertainties about them. The role of news media wielded a significant influence over people's opinions regarding the COVID-19 vaccine. News and media play an important role in shaping people's levels of acceptance and understanding regarding the vaccine. Understanding media perceptions of covid 19 vaccine is essential for ensuring effective vaccine communication, building trust, and addressing vaccine hesitancy (Machingaidze and Wiysonge 1338).

The aim of this project is to build a corpus that allows researchers to study the regional comparisons about media perspective in the New Zealand and Australia regarding the COVID-19 vaccines. To achieve this, we want to select a leading News & Media website as our source of corpus. According to the result from the digital intelligence platform SimilarWeb, which specialises in web analytics, web traffic, and performance, we have selected two highly visited news and media publisher websites for collecting texts. For the New Zealand media perspective, we have chosen nzherald.co.nz, while for the Australian media perspective, we have selected abc.net.au.

2. Corpus

Our corpus consists of two subcorpora with two main considerations:

Firstly, the source countries of the corpus have comparability. When discussing the COVID-19 vaccine, New Zealand (NZ) and Australia (AU) share relatively similar backgrounds, such as having well-developed and comparable healthcare systems, emphasizing public awareness, and having lower population densities compared to other countries. While there are differences, like AU having a greater variety of vaccine types and a larger population, these differences could encompass a wider range of media viewpoints. Therefore, sourcing media perspectives from both countries provides more comprehensive data for our topic.

Secondly, the subcorpora themselves are directly comparable. They are composed of media articles from a leading website in NZ and AU, respectively. We used the same keyword ("covid 19 vaccine") and a "most relevant" filter to search and collect articles. The themes, data sources (leading websites), and domains (media articles) of both subcorpora are consistent (Reppen 13).

Moreover, each subcorpus has a relatively balanced dataset, which facilitates effective comparison and analysis.

Based on these considerations, we initially built a corpus comprising 100 documents, from which we were unable to identify meaningful and relevant word tokens for our intended research focus. Consequently, we expanded our corpus to encompass 200 documents totaling 211,057 word tokens. It consists of two comparable subcorpora: the first subcorpus is sourced from the website nzherald.co.nz, comprising 100 documents with a total of 77,419 word tokens; the second subcorpus is sourced from the website abc.net.au, consisting of 100 documents with a total of 133,638 word tokens.

By utilizing AntConc's n-grams tool with a minimum n-gram size of 1 and a maximum of 2, we successfully identified 5 word tokens pertinent to the media perspective on the COVID-19 vaccine. These words—specifically "booster," "government," "mandates," "risk," and "side effects"—had frequencies of 507, 382, 243, 258, and 32 respectively. These identified words spurred our interest for further analysis. Our aim is to unearth language patterns associated with government actions and discourses about risks and side effects within the context of the media perspective. Accordingly, we proceeded to classify them into two distinct groups: "government" "mandates" and "booster" in one, and "risk" and "side effects" in the other. This categorization serves as a foundation for conducting more comprehensive explorations of their usage within the corpus.

With these two groups defined, our initial step is to employ the "Keyword List" tool. This tool enables keyness analysis to compare language usage between the two subcorpora. We can designate NZ subcorpus as the target corpus and AU subcorpus as the reference. Considering the relatively modest corpus size, we can set a "keyword statistic threshold" of $p<0.01$ to mitigate the risk of excessive false positives and enhance the precision of our analysis. Additionally, we acknowledge the flexibility of interchanging target and reference corpora for analysis.

This comparison allows us to discern disparities in the media coverage of the COVID-19 vaccine between New Zealand and Australia. Through this comparison, we aim to discover varying expressions and language usage in the media coverage, and thus gain insights into the diverse media perspectives presented by the media in NZ and AU.

Additionally, we can make use of AntConc's various tools to analyse each of these 5 word tokens individually. We can utilize the collocation and clusters tools to understand their collocations and associations, thus analysing language patterns. We can employ the concordance tool to closely examine the context in which these words appear, gaining a deeper understanding of their contexts and usages. In conclusion, the outlined methods allow us to thoroughly explore the media perspective on the covid-19 vaccine in NZ and AU, helping us uncover language patterns.

3. Method

Employing "covid 19 vaccine" as the keyword, we retrieved the news article related to the project's topic from both nzherald.co.nz and abc.net.au. To ensure a collection of texts closely aligned with our topic, these results were then organized according to their relevance using the sort option provided by the website. Upon a review of the search results obtained from abc.net.au, it came to our attention that among the retrieved content, there were radio news articles, Chinese news articles etc. Given that our project centres around the construction of a text-based corpus, a decision was made to further refine the search results by selecting “ABC News” category in the filters. Then we applied Webscraper.io as our web scraping tool.

- The first step is to create a site map for our target website. Utilizing the Inspector within Chrome Developer tools, we can inspect the CSS selector for each section of the website which helps us to construct the site map.
- The second step is to scrap the information. By selecting "Scrape" from the dropdown menu of the sitemap, we can extract the data and subsequently export it as a CSV file.
- The third step is to transform CSV file to text file. This step will focus on retaining the last 50 text articles, which represent the most relevant ones based on the sorting by relevance conducted earlier.
- The last step is to run the jupternotebook Scrapey.ipynb from the learning materials. Then we can export the content we have scraped into a directory of text files.

During this process, a significant challenge arises in the form of selecting the appropriate selector, particularly when dealing with pagination and timestamp. On nzherald.co.nz, article timestamps are stored using two different methods: one is individually, making it easily extractable, while the

other is embedded within a title picture. Therefore, an extra timestamp selector has been incorporated to retrieve the embedded information.

Another challenge we faced during the process was the selection of the top 50 most relevant articles. Initially, we manually curated these articles from the pool of text files we had gathered from the website. However, upon closer examination, we discerned a pattern in the arrangement of articles within the CSV file. The articles were sorted in ascending order of relevance, with the most relevant article located at the end. Leveraging this observation, we modified our approach. We decided to retain only the last 100 articles within the CSV file, as these represented the most relevant ones, and subsequently converted them into individual text files. This adaptation allowed us to automate the selection process while ensuring the inclusion of the articles that are most closely aligned with our research objectives.

4. Limitation

In analysing the constructed corpus, it is important to acknowledge the inherent limitations that define its scope and relevance. While our methodology aimed to capture a comprehensive view of media perspective on COVID-19 vaccination, it is crucial to note that certain constraints and biases may have influenced the breadth and accuracy of our findings. In the following session, we will discuss these limitations, highlighting the potential implication they have on the insights derived from our corpus.

Scope limitation:

Our corpus creation strategy is centred around the most significant websites in New Zealand and Australia, aiming to adopt a focused research approach. Nevertheless, this deliberate decision does place limitations on the range of viewpoints and understandings. Due to the restricted range of information sources, there exists a possibility for bias to emerge during the interpretation of the subject matter, which may lead to the neglect of diverse perspectives.

It is vital to recognise that the selected website may exhibit distinct editorial preferences and stances. As a result, it is possible for our corpus to unintentionally acquire these biases, so potentially compromising its impartiality and comprehensiveness. Despite our conscientious planning, which encompassed the utilisation of the website's sitemap, it is possible that certain

components, such as videos, images, and other multimedia content, may have been omitted owing to technological constraints, the dynamic nature of the website's structures, or unanticipated circumstances.

Furthermore, it is imperative to acknowledge an additional constraint within the parameters of our research. A total of 100 text files were gathered for each subcorpus. It is crucial to note that the quantity of text files available may not be adequate to comprehensively support our examination of media viewpoints on COVID-19 vaccinations.

Relevant limitation:

Our corpus compilation significantly relied on the use of a particular keyword ("covid 19 vaccine") and the "most relevant" filter. This method may unknowingly exclude valuable articles that contribute substantially to the discourse but do not precisely match the keyword. It is possible to neglect nuanced discussions and alternative terminologies (Lüdeling et al. 7). The website's automated algorithm used to determine the "most relevant" articles may prioritize factors such as popularity or recency. Our interpretation of relevance may not always align with this approach, which jeopardizes the exhaustiveness of our corpus.

The determination of "relevance" is subjective and can vary between researchers. Our reliance on automated relevance filters may lead to the unintentional exclusion of relevant articles that are significant from a variety of perspectives. While selecting the top 100 articles from each website provides a snapshot, it may not capture the complete scope of discussions. This sample size may neglect diverse perspectives, controversies, and emerging trends, impeding an all-encompassing analysis.

Appendix: Acknowledgements

We would like to express my gratitude to ChatGPT for its invaluable assistance throughout the process of writing this project report. The role of AI in shaping and refining various sections of this report has been significant.

In the background section, we leveraged ChatGPT's capabilities by using the prompt "Draft a background structure for my corpus building project based on the following description." This allowed me to generate an initial draft for the background based on a precise project description. Subsequently, we developed the actual background content, guided by the foundational ideas provided by ChatGPT. The collaborative effort ensured a comprehensive and well-structured background for this report.

Furthermore, in the limitation section, we engaged ChatGPT using the prompt "What would be the potential limitation of my corpus building project based on the following project background and methodology." This enabled me to identify potential limitations associated with the corpus building project. Drawing from the project's background and methodology descriptions, we selected relevant information and crafted the limitation statements. To enhance the accuracy and coherence of these limitations, we subjected them to ChatGPT's proofreading.

This project would not have reached its current form without the intelligent guidance and insightful suggestions offered by ChatGPT. We are thankful for the support and assistance that AI has provided in making this project a reality.

References:

Lüdeling, Anke, Stefan Evert, and Marco Baroni. "Using web data for linguistic purposes." *Corpus linguistics and the Web*. Brill, 2007. 7-24.

Machingaidze, Shingai, and Charles Shey Wiysonge. "Understanding COVID-19 vaccine hesitancy." *Nature medicine* 27.8 (2021): 1338-1339.

Reppen, Randi. "Building a corpus: what are key considerations?." *The Routledge handbook of corpus linguistics*. Routledge, 2022. 13-20.