Omkar Joshi

DIGI405

18th May 2024

Corpus Analysis Essay

Introduction

A corpus is a collection of texts from different resources built for a specific purpose. It is set of language text carefully assembled; it is not a randomly collected data from various sources. Corpus analysis is an opportunity to test various intuitions about natural language and challenge our ideas. It provides us quantitative measures to understand text (Björkenstam). It also allows to study language patterns, changes in natural context along with grammatical structures, collocations, semantic patterns and variation in language across different genres.

Background

While checking for interesting corpora among peer groups I found, group S has built corpora which focuses on "Mainstream media's portrayal of lockdown measures in New Zealand and the United States" (Steven, Delgado and Korng). Their aim was to examine, compare New Zealand and United States of America's media depiction regarding lockdown in Covid-19 pandemic. I found it very thought-provoking to understand media representation of lockdown, it helped to understand how two different countries approach same problem with diverse approach while considering their own situation. That's the motivation behind choosing 'Group S' corpora for this assignment.

Data and Methods

The New Zealand Herland along with The Dominion Post these two major sources are used to build corpus of New Zealand. It consists of 499 unique articles which includes total 339619-word count. In case of United States corpus two important newspaper's articles are utilized, namely The New York Times and The Washington Post. This entire corpus comprises of total 353 unique documents, consists of total 559291 words.

Group S mentioned one technical issue they faced after building United State corpus that "US sub corpus contained many word-tokens such as "https", "com", "www", "nytimes" and "html" We realize that much news from the US have references to other containing hyperlinks, which was not the case for the New Zealand news" (Steven, Delgado and Korng). To overcome this technical issue, I used Python programming and wrote two functions one for cleaning these unnecessary word tokens and hyperlinks from corpus text files and another function to read, modify changes and save those files in given directory.

Concordance and keyword analysis these two methodologies I have applied for corpus analysis. Keyword analysis gives high frequency words which helps to conveys focal point of

corpus and concordance analysis gives context of words in corpus and finds specific patterns in the text. As a pre-processing of both corpora in AntConc software, I have applied NLTK English stop list provided in course material. It helped to eliminate insignificant high frequency words. Some additional unwanted words list is mentioned below, which is further eliminated.

Additional stop list -

New Zealand Corpus	United State Corpus
said	said
per	one, many
also	would
could, would	mr

Results and Discussion

I started corpus analysis with high frequency words because those words help to understand main focus point of corpus. Words and phrases which occur most frequently can be used further to understanding corpus in depth by using collocation and concordance analysis

Top 10 high frequency words -

New Zealand Corpus		United State Corpus	
Frequency	Word	Frequency	Word
1862	new	2196	people
1825	lockdown	1973	new
1738	covid	1938	coronavirus
1599	people	1361	virus
1069	zealand	1296	health
908	health	1121	pandemic
892	level	1109	cases
844	auckland	999	state
781	cent	950	lockdown
742	year	947	city

Above high frequency words table highlights some common words like health, new and similar words like covid and coronavirus. Unusual word appeared in New Zealand corpus is level and also it has mention of country name along with important city. These words indicated that New Zealand corpus is more focused on its own country and covid-19 lockdown situation where as United State corpus seems to be focused on overall scenario of covid-19 pandemic, but any conclusion can not be drawn from this because high frequency words might give focus point of corpus but for analysing it deeply, context of these words is paramount.

Along with quantitative, qualitative analysis method also needs to be incorporated to yield of more effective analysis. A concordance analysis is one of the most important techniques which allows corpus analysis in close context (Baker). A concordance is occurrence of particular search word in corpus along with its context i.e. few words to the left or to the right. Concordance is also referred as “Key Word in Context” i.e. KWIC. As to start examining corpora I initiated with word “lockdown” as a search word for concordance analysis.

As mentioned previously in frequency analysis, word “level” appeared frequently in New Zealand corpus. Same word again appeared in concordance analysis, when search filter is applied to one word left to word “lockdown”. There was a pattern observed – “level * lockdown” for example - “level 3 lockdown” and “level 4 lockdown”. To understand it further I took reference of academic article from Lancet Regional Health, it states that alert level was designed for the policy setting of first outbreak of coronavirus in New Zealand from February to May 2020 (Kvalsvig, Wilson and Davies). Article mentions that in New Zealand there were total 6 alert levels, among which level 3 is ‘Limited local outbreak’ and level 4 is ‘Stop regional spread’ of covid-19. Once read through context of levels in text files of corpus, it talks about how New Zealanders were going through level 3 and level 4 of lockdown. Academic article confirms that what exactly mention of word ‘level’ before lockdown meant.

To understand New Zealand media’s perception about trace of covid-19 cases, I used word “tracing” as a search word with same filter as applied earlier. I found multiple appearance of common pattern “contact tracing”. After going through text files to understand context of this pattern. Many texts were mentioning about how Ministry of Health requires new technology for contact tracing and it launched mobile application and how rapid contact testing is been started. Some texts also indicate experience of people after using application. Journal article by Smart Health (Ali and Dang) authorizes that in early 2020 NZ COVID Tracer application was launched and by May 2020 around 380,000 users installed this app on there mobile or other devices. It also talks about how this application was used as a digital diary by users for logging their daily travel and moment.

Isolation was important in process of not allowing covid-19 spread, so next I searched with regards to “isolation” in New Zealand corpus. Pattern which predominantly appeared is “managed isolation”. To understand this further, I went through text files of corpus. It identifies an abbreviation “MIQ”. Journal article published by Disaster Risk Reduction, regarding experience of people living in hotel isolation and quarantine in New Zealand informs that MIQ is hotel-based managed isolation and quarantine. It is a public health intervention by New Zealand’s covid-19 border control strategy for returning citizens and permanent residents. This covid elimination strategy was first introduced on 9th April 2020. In this all citizens, permanent residents and pre-approved non-residents arriving in country by air were required to remain in MIQ hotels for 14 days (Gray, MacDonald and Puloka).

To comprehend United States lockdown situation depicted by media, I used search term “lockdown” in second corpus which represents United States condition. Three patterns most significantly got highlighted in corpus for example “anti-lockdown”, “lockdown movement”, “lockdown protest”. After going through texts relating to these patterns, it showed there were lot of social media posts and YouTube videos were against lockdown in United States. At

many hotspots in country such as Michigan, people were protesting against government policies of lockdown. This finding from corpus is confirmed by journal article published by Political Science Quarterly that, there were many protest events conducted in multiple regions in U.S. from March 2020 to March 2022 because of stringent policies of lockdown and demanded that government should consider reopening the state to allow normal business and personal activities (Pfaff, Plümper and Neumayer).

Another pattern which I found after rigorous search is that, “Austria's lockdown”, “Britain’s lockdown”, “India’s lockdown” i.e. country name followed by word lockdown. After reading through all of them, it was evident that U.S. media reported a lot about other countries lockdown situation and overall covid-19 crises in that respective country.

“People” is next search term I used for understanding situation of individuals in U.S. in pandemic era. There were lots of patterns found like “million people”, “older people” but one pattern which makes noticeable mark was “black people”. Investigating through text files in corpus for this pattern, it gives shocking insights how lockdown was used as an instrument for systematic racism by discriminating by skin colour. This awful finding is confirmed by an article published by The Royal Society for Public Health that, during pandemic Asian American people had 4.1 times higher odds of experiencing covid-19 related discrimination when compared to white people in U.S. (Oh and Waldman).

New Zealand corpus - Concordance Analysis		
Search word	Pattern	Findings
lockdown	level 3 lockdown level 4 lockdown	In New Zealand 6 level alert system was introduced to counter virus spread
tracing	contact tracing	New Zealand launched Tracer app, for tracing people to trace back spread of virus
isolation	managed isolation	New Zealand initiated MIQ i.e. hotel-based managed isolation and quarantine

United States corpus - Concordance Analysis		
Search word	Pattern	Findings
lockdown	anti-lockdown lockdown movement lockdown protest	In many parts of United States there were significant amount of protest events against covid-19 policies
	Austria's lockdown Britain’s lockdown India’s lockdown Israel’s lockdown	United States media has covered a lot of country’s lockdown situation
people	black people	In United States there were lot of discrimination based on colour in pandemic

CONCLUSION

After analysing two corpora, it is very clear that media portrayal of lockdown and overall pandemic situation is drastically different from one another. Analysing New Zealand corpus gives result as media in New Zealand was more focused on its own country and media reported tremendously about what measures government is taking to reduce impact of covid-19 on people. New initiatives such as MIQ and alert level system were part of reported policies. On other hand, United State corpus depicts about how people are unsatisfied with strict lockdown policies and were protesting against it. Media has large coverage in terms of geographic locations i.e. U.S. media was interested in reporting pandemic situation of other countries also. Media also outlined discrimination in own country based on colour. In summary New Zealand media's focus was on positive and constructive things during covid era where as United State media was more into discrimination and restart of business. This conclusion is supported by journal article which researches about new media narrative of covid-19 across 20 countries, it claims that New Zealand and Australian media's narrative was around hope and current uncertain situation in respective countries where as United States media focused on economic re-opening and tackling discrimination in pandemic (Ng, Chow and Yang).

Works Cited

Ali, Zarqa Shaheen and Hoang Dang. “Factors impacting the use of the NZ COVID Tracer application in New Zealand.” (2022). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8872797/>.

Baker, Paul. *Using Corpora in Discourse Analysis*. Bloomsbury Publishing, 2006.

Björkenstam, Kristina Nilsson. “What is a corpus and why are corpora important tools?” n.d.

Gray, Lesley, et al. “The lived experience of hotel isolation and quarantine at the Aotearoa New Zealand border for COVID-19: A qualitative descriptive study.” (2022). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9764876/>.

Kvalsvig, Amanda, et al. “Expansion of a national Covid-19 alert level system to improve.” *The Lancet Regional Health - Western Pacific* (2021): 3.

Ng, Reuben, Ting Yu Joanne Chow and Wenshu Yang. “News media narratives of Covid-19 across 20 countries: Early global convergence and later regional divergence.” (2021).

Oh, H. and K. Waldman. “Building a coalition to fight coronavirus-related discrimination against people of color.” *The Royal Society for Public Health* (2020). <https://doi.org/10.1016/j.puhe.2020.05.053>.

Pfaff, Katharina Gabriela, Thomas Plümper and Eric Neumayer. “Polarized Politics: Protest Against COVID-19 Containment Policies in the USA.” *Political Science Quarterly* 138.1 (2023): 23-46. <https://doi.org/10.1093/psquar/qqac002>.

Steven, Joe, et al. “Mainstream media's portrayal of lockdown measures in New Zealand and the United States.” 2024.