Corpus-based Analysis of MH17 Online Dutch News Articles

Wan Najmiyyah Wan Md Adnan
Faculty of Social Sciences and Humanities
Universiti Teknologi Malaysia
Jalan Sultan Yahya Petra
54100 Kuala Lumpur, Malaysia
Academy of Language Studies,
Universiti Teknologi MARA Terengganu,
23000 Dungun, Terengganu, Malaysia
+60126337163
wannajmiyyah@uitm.edu.my

Sarimah Shamsudin
Faculty of Social Sciences and Humanities
Universiti Teknologi Malaysia
Jalan Sultan Yahya Petra
54100 Kuala Lumpur, Malaysia
+60127163301
ssarimah.kl@utm.my

ABSTRACT
This paper features the step-by-step corpus analysis of texts taken from an online English-based Dutch news portal as part of our corpus project on media representations of MH17 aviation disaster news discourse. The main objective of this paper is to look into the frequency and concordance analysis of the words used in real-world contexts of online news articles on aviation news. Firstly, we introduce the MH17 online Dutch News as a corpus used in the study and how the corpus is developed. Next, we report examples of findings on the text analysis of the corpus by focusing on the frequency of vocabulary lists used in the corpus, and the concordance analysis of the words 'Dutch' and 'Netherlands'. All examples are taken from MH17 online articles from Dutch News.nl. Hence, with the step-by-step guideline on how to conduct corpus analysis, this would be able to shed some light on new opportunities and perspectives in corpus linguistics research.

CCS Concepts
• Computing methodologies→Artificial intelligence→Natural language processing→Discourse, dialogue and pragmatics

Keywords
Online news articles; Corpus analysis; Aviation disasters news discourse

1. INTRODUCTION
The main study of our corpus project investigates the media representation of MH17 aviation disaster news discourse from 2014 to 2016. MH17 aviation tragedy refers to the tragic incident of a passenger plane (MH17) from Amsterdam to Kuala Lumpur that was shot down while flying at Ukraine-Russia border on 18th July 2014. It crashed in Torez, Ukraine and all on board were killed. Since the main research focuses on MH17 disaster, we decided to explore the use of online news articles from a country affected by the disaster. This study therefore concentrates on an online English news portal from the Netherlands. The Netherlands is selected as it represents one of the countries affected by the MH17 aviation disaster. Furthermore, 68 percent or 193 of the passengers out of 238 were of Dutch nationality.

This project uses a user-friendly corpus analysis software application, in line with previous work associated with investigating the characteristics of online newspaper corpora. This research project is hoped to be able to shed some light on how simple corpus analysis techniques could provide robust analysis on large number of texts. Hence, this paper reports on the analysis of a Dutch English online news from 2014 to 2016. For this online newspaper issue, 60 relevant articles were identified. Each article was relatively short as it contained only important information with regard to MH17 aviation disaster.

Reuters Institute and University of Oxford's Digital News Report stated that the Netherlands media view their reported news very highly and they believe that their news media project very trustworthy materials [1]. According to the survey, The Netherlands (59 percent) ranked third in terms of trust in media, behind Finland (62 percent) and Portugal (62 percent). The media are said to have high reputations in getting trusts from the public, be it in the newspapers, TV news, or radio news. Since the public news media set very high quality standard in the Netherlands media scene, commercial media outlets also have to follow suit if they want their news to be consumed by the users [1]. Interestingly, the issue of 'fake news' condones very low scores in the Netherlands media landscape. Hence it is clear that news awareness in the Netherlands is very high. The digital media consumption in the Netherlands has also grown over the years.

Traditional news discourse usually feature the sources' actual words – or paraphrase of something that the sources said [2]. Thus, this supports the idea that news is constructed based on the writers' voices and it becomes essential in deciding the orientation of the news story. Language comes into play in this as its usage and composition in the layout and organisation of news discourse have powerful impacts in determining how the news reaches its targeted audience [3].

In the advent of technology, the news media have become easily accessible to the public. News is becoming prompt and publicly available, and audiences have become an integral part in producing news that would attract clicks from the audience [4].

ICEMT 2019, July 22–25, 2019, Nagoya, Japan

ACM ISBN 978-1-4503-7210-7/19/07…$15.00
https://doi.org/10.1145/3345120.3345191

One of the advantages of online news discourse is that audience are able to gain instant access to news from online news portal. Some portals are from long-established news organisations, while others are with doubtful reliability as blogs and alternative news can easily be provided by the audience themselves [3].

1.1 Aviation Disaster News Discourse

While news discourse is popular in many corpus studies, studies related to aviation disaster news discourse are still lacking and need further exploration.

Cheng's study, for example, explored the pragmatic analysis of the word 'if' in news discourse about a collision between a US surveillance plane and a Chinese F-8 jet fighter on April 2001 [5]. The study analysed 94 news stories from cnn.com and 15 news stories from chinaonline.com during the 12-day period between the air collision and the return of the US crew members. Interestingly, it was found that the marker 'if' was used as a face-saving tactic during the disaster, regardless of the original intention of using it to implicate 'uncertainty'. There was also another study by Choi that analysed the earlier articles of the same aviation tragedy which were printed immediately after the disaster took place [6]. It was identified that the reporting of the aviation disaster news differed depending on the ideology of each newspaper, but usually it works as a representation of the government's voice.

Another study that examined linguistic tools of alienation and empowerment in the Chinese official press narratives during aviation disaster also found that Chinese media portrayed the nationalist tone and assertiveness in the media discourse, and it also worked as a tool to convey a sense of social harmony and patriotism towards the Chinese victims [7]. A different study on public discourse related to the September 11 aviation disaster in 2001 by Nimmer found that blaming and scapegoating were acceptable practices in the news [8]. He then concluded that by instilling public fear and anxiety in the news, it is easier for those in power to dominate support from the public.

Hence, we can clearly identify that even though all the reviewed studies were discussing aviation news discourse, the focus had mostly been towards terrorism discourse [8] and political discourse [5-7]. For example, studies on representations of air collision between US surveillance plane and China jet fighter in 2001 put more emphasis only on the analysis of the governments involved, and not much attention were given on other actors involved in the news [5; 6]. Therefore, the first step towards analysing the aviation disaster news discourse would be through the frequency-based corpus analysis, as suggested in this study.

1.2 MH17 Aviation Disaster News

When discussing online news articles on MH17 aviation disaster news, it is found that there were very limited studies done on the topic, especially when it comes to media communication and language studies. In one study for example, Sienkiewicz investigated on the issue of digital labour in the media investigation of MH17 in Ukraine [9]. Digital labour here refers to the unconventional approach taken by the Ukranians online, in which they release video evidence on social media in counteracting Russia's use of global broadcasting. Meanwhile, Schubert investigated a corpus of 218 online TV transcripts to identify the presence of anonymous sources in news discourse of MH17 [10]. The study analysed the utterances of the unidentified sources by using both qualitative and quantitative analysis. The qualitative data analysis adds the use of speech act and adjacency pairs of the newscasters in order to identify newsworthiness, while the quantitative data analysis inspected the number and distribution of the anonymous sources.

There were also contradicting findings on the role of the news writers when reporting the news. While most research highlighted the important role of news writers in determining what to write in the news reporting, there were also contradicting claims that news writers do not have control or power on what they write, but the events and people behind the events themselves (like the government) were in the power position [11-13].

The growth of user-friendly corpus analysis software also implies that investigating the features of corpus regardless of its size has become easier. The use of corpus linguistics for one, could be complementary to any linguistics analysis. The use of corpus and concordancing software can offer both quantitative and qualitative analysis on texts [14]. The use of frequency and other statistical measures in finding word occurrences provide quantitative point of view. Qualitative analysis can also be conducted through examining collocations, describing lexical and semantic patterns, and identifying discourse functions.

2. ANALYSING THE CORPUS

The methodology implemented in this study is the use of corpus linguistics as an aid to text analysis. Nevertheless, we still need to conduct manual interpretation of the results. This paper demonstrates that the use of corpus analysis in linguistics research can be used to explore representation issues, explore how they are evaluated and at the end, form the media representation of the aviation disaster news discourse. In the process of analysing the corpus, this study adapted stages of analysing corpus from Sinclair [15]. Figure 1 below shows Sinclair's stages for analysing corpus [15].

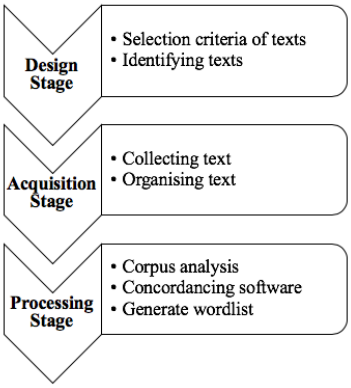

Figure 1. Stages for analysing corpus [15]

The first step in analysing a corpus is the design stage, in which we identify the texts based on the agreed selection criteria. DutchNews.nl is chosen because it is easily available online and provides coverage of Dutch news for English-speaking community. Sixty articles were identified from Dutchnews.nl that are related to MH17 from 2014 to 2016.

The second step after the design stage is the acquisition stage in which we collect the texts and organise them by using coding. For example, 17DN01, 17DN02, and 17DN03. 17 refers to MH17, DN refers to Dutch News and 01 refers to article number one.

Lastly, the processing stage refers to the analysis part of the corpus, in which concordancing software is used to generate wordlists.

For this research, we used a concordancing software programme called WordSmith 7.0 to conduct the text analysis. WordSmith 7.0 is considered as one of the most comprehensive and established software for corpus linguistics [16]. Its website also offers step-by-step tutorial on how to use the software. The software needs to be purchased and downloaded from its website. As with other text analysis software, WordSmith 7.0 requires the corpus to be in the form of txt files in order to use it for corpus analysis. Figure 2 below shows the main page of WordSmith 7.0.

Figure 2. Main page of WordSmith 7.0

2.1 Frequency-based vocabulary list

Frequency-based vocabulary list refers to the frequency of words used in the corpus and they are ranked based on the most frequent to the least frequent words. The list can also be classified into different wordlists (grammar-based or content-based) based on what the researchers are looking for. It was found that the total tokens for Dutchnews.nl online articles are recorded at 19,716 tokens. Token here refers to the total word count in a corpus [14].

2.1.1 Frequency keyword list

In creating the frequency keyword list, we analysed 60 collected online news articles from Dutchnews.nl. After cleaning up unnecessary data and eliminating unrelated words, the cleaned corpus of txt files were uploaded into the software. After running the corpus through the keyword analysis using the WordSmith 7.0 software, Figure 3 shows the first page of the actual frequency keyword list produced. Using the keyword list analysis, it was found that the ten most frequent words used in the corpus are the words 'the', 'to', 'of', 'in', 'and', 'a', 'is', 'on', 'Dutch' and 'for'. Table 1 shows the list of ten most frequent grammar words and its frequency percentages for online news articles from DutchNews.nl related to the MH17 disasters. On the other hand, Table 2 shows the list of ten most frequent content words in the corpus.

2.1.2 Concordance list

Not only frequency words can be analysed based on its occurrences (as stated in Table 1 and Table 2), but we can also design our analysis based on target words in the corpus. The use of concordance list can show the dispersion of a selected target word and how it is being used in the texts. For this, we used an example of the most frequent content word 'Dutch' to see the representation of the word in Dutchnews.nl. It can be considered as an important word as its occurrence in the corpus is 223 times.

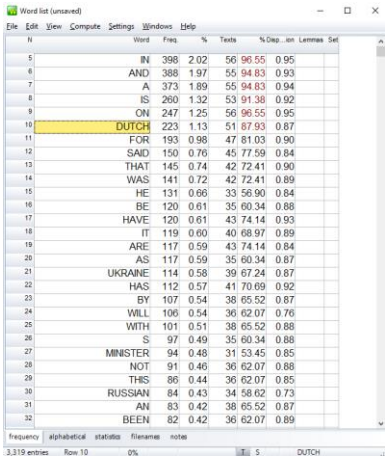

Figure 3. Keyword analysis using WordSmith 7.0

Table 1. Frequency and percentage for most frequent grammar words

Frequency	Percentage (%)	Grammar Words
1454	7.37	THE
564	2.86	TO
542	2.75	OF
398	2.02	IN
388	1.97	AND
373	1.89	A
260	1.32	IS
247	1.25	ON
193	0.98	FOR
145	0.74	THAT

Table 2. Frequency and percentage for most frequent content words

Frequency	Percentage (%)	Content Words
223	1.13	DUTCH
150	0.76	SAID
120	0.61	UKRAINE
114	0.58	MINISTER
112	0.57	NOT
94	0.48	RUSSIAN
91	0.46	NETHERLANDS
84	0.43	RUTTE
77	0.39	CRASH
71	0.36	PEOPLE

To create the list, we clicked the Concord button in the WordSmith 7.0 software, and searched for the word 'Dutch' as the search-word. The concordance then listed all the examples of 'Dutch' which also shows the words and words separator that comes before and after it. Figure 4 shows the snapshot of the concordance list for the word 'Dutch'.

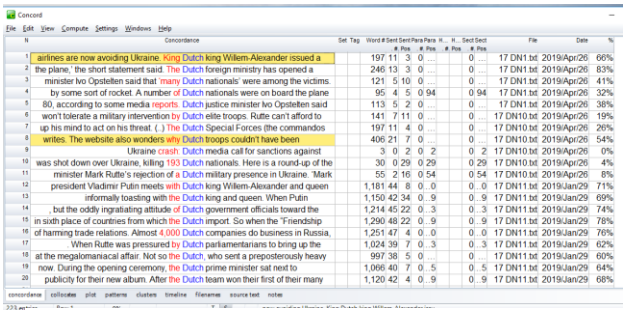

Figure 4. Concordance list for 'Dutch'

From Figure 4, the words that occurred before the word 'Dutch' are mostly grammar words like prepositions ('with', 'of' and 'by') and articles ('a', 'the') and those that occurred after the word 'Dutch' are mostly nouns ('military presence', 'ministers', 'nationals', 'King') in which the word 'Dutch' mostly functioned as an adjective to describe the noun.

Similar analysis was also conducted on another related content word, 'Netherlands', as we want to find out how the words 'Dutch' and 'Netherlands' are represented in the corpus. Hence, Figure 5 shows the concordance list for the word 'Netherlands'.

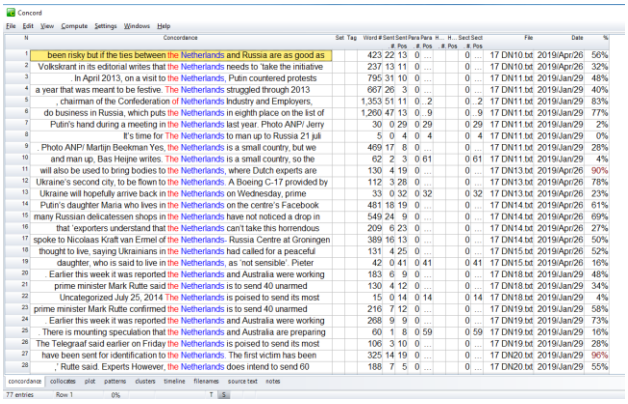

Figure 5. Concordance list for 'Netherlands'

From the analysis, the words that occur before 'Netherlands' is the article 'the' and those that occurred after the word 'Netherlands' are mostly verbs such as 'is', 'does', 'struggled', and 'needs' and the conjunction 'and'.

3. DISCUSSION

Based on the analysis, it was found that there are slight differences as in how the words 'Dutch' and 'Netherlands' are represented in the corpus. The lexical 'Dutch' is commonly used when referring to locals and other smaller groups of actors that mostly represent the people of Dutch such as 'the Dutch King', 'Dutch government' and 'Dutch nationals'. On the other hand, the word 'Netherlands' in the news is always associated to actions and decisions taken by the country or shows its ties with other countries. Thus, although there seems to be a slightly different pattern in the usage of the words 'Dutch' and 'Netherlands' in the corpus, both words show positive representation especially in displaying nationalism in the news. These findings are found to be similar to previous studies which also indicated the elements of nationalism and representations of government's voices in the news [5-7]. Nevertheless, further analysis should be conducted to identify whether the use of grammar words, especially nouns and adjectives that occurred before the words 'Dutch' and 'Netherlands' contribute in giving in-depth representations of those words in the news discourse.

Previous studies also indicated that there were contradicting claim of who are in power when it comes to representations of aviation news discourse. While some studies claimed that news writers are the ones in control on what they write, some other studies claimed that the events and people behind the events (also known as social actors) were in the power position [11-13]. Hence, further exploration on the representations of social actors in aviation disasters would give further support to these studies. This present study can also work as a guideline to the process of identifying social actors through the frequency analysis of the content words in the corpus.

An exploration of representation studies on aviation disaster news discourse would provide substantial contributions to the already established research on news discourse. Nevertheless, it is also worthy to note that the use of corpus methodology is not without its limitations. While the use of corpus could provide wider analysis and estimation of language use in different discourse, language study is still a subjective matter that requires in-depth techniques of language analysis that would go beyond the use of corpus methodology. Hence, it is suggested that while corpus approach is reliable and use extensive data, other approaches and different types of analysis in language should be applied as a support to the corpus technique. Compatibility of approaches in language analysis studies should be conducted as extensively as possible in order to achieve maximum results in any language research.

4. CONCLUSION

Based on our frequency and concordance analysis, we believe that this study could contribute to richer analysis of corpus representative studies, especially in the unexplored field of aviation disaster news discourse in affected countries. The use of frequency lists and concordance analysis to investigate patterns of meanings especially for lexical and semantics analysis would also provide strong evidence to the dynamics of using corpus in understanding representation studies in aviation news discourse. This kind of analysis is widely applicable and relatively easy to be conducted by early researchers, hence encouraging further exploration of different discourses in corpus-based representation studies.

5. ACKNOWLEDGMENTS

This study was funded by the Ministry of Education (Malaysia) and supported by Universiti Teknologi Malaysia under Grant Number 4F987.

6. REFERENCES

[1] Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A. L., and Nielsen, R. K. 2017. *Digital News Report 2017: A Global Assignment*. Reuter Institute, University of Oxford.

[2] Coleman, S. and Ross, K. 2010. *The Media and The Public: "Them" and "Us" in Media Discourse*. Wiley-Blackwell, Sussex, UK.

[3] Bednarek, M. and Caple, H. 2012. *News Discourse*. Continuum, London.

[4] Hoskins, A. and O'Loughlin, B. 2007. *Television & Terror: conflicting Times & The Crisis of News Discourse*. Palgrave Macmillan, New York.

[5] Cheng, M. 2002. The standoff - What is unsaid? A pragmatic analysis of the conditional marker 'if'. *Discourse & Society* 13. 3, 309-317.

[6] Choi, D., 2002. A critical discourse analysis on different representations of the same event in the media. *LAUD.* 1-17.

[7] Lams, L. 2010. Linguistic tools of empowerment and alienation in the Chinese official press: accounts about the April 2001 Sino-American diplomatic standoff. *Pragmatics.* 20, 3, 315-342.

[8] Nimmer, L. 2011. De-contextualization in the terrorism discourse: a social constructionist view. *ENDC Proceedings.* 14, 223-240.

[9] Sienkiewicz, M. 2015. Open BUK: Digital labor, media investigation and the downing of MH17. *Critical Studies in Media Communication.* 32, 3, 208-223. DOI= http://dx.doi.org/10.1080/15295036.2015.1050427.

[10] Schubert, C. 2015. Unidentified speakers in news discourse: A pragmatic approach to anonymity. *Journal of Pragmatics.* 89, 1-13. DOI= http://dx.doi.org/10.1016/j.pragma.2015.09.003.

[11] Park, S., Bier, L.M., and Palenchar, M.J. 2016. Framing a mystery: Information subsidies and media coverage of Malaysia Airlines Flight 370. *Public Relations Review*. 42, 4, 654-664. DOI= http://dx.doi.org/10.1016/j.pubrev.2016.06.004.

[12] Maros, M. and Nasharudin, S. N. S. 2016. Analysis of interaction and institutional power relations in MH370 press conferences. *Pertanika Journals.* 24, S, 169-180.

[13] Sonnevend, J. 2018. Interruptions of time: The coverage of the missing Malaysian plane MH370 and the concept of 'events' in media research. *Journalism.* 19, 1, 75-92.

[14] Baker, P. 2006. *Using Corpora in Discourse Analysis*. Bloomsbury Academic, London.

[15] Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

[16] Scott, M. 2019. *WordSmith Tools Version 7: Lexical Analysis Software.* Oxford University Press, Oxford.