A COVID-19 Corpus Creation for Bengali: In the Context of Language Study

Prasanta Mandal[1*], Apurbalal Senapati[2]

[1] Department of Computer Science and Engineering, Govt. College of Engineering and Textile Technology, 12, William Carey Road, Serampore, Hooghly-712201, West Bengal, India.
[2] Department of Computer Science and Engineering, Central Institute of Technology Kokrajhar, Kokrajhar-783370, Assam, India.

*Corresponding author
doi: https://doi.org/10.21467/proceedings.115.9

ABSTRACT

A corpus is a large collection of machine-readable texts, ideally, that should be representative of a Language. Corpus plays an important role in several natural language processing (NLP) and linguistic research. The corpus development itself is a substantial contribution to the resource building of language processing. The corpora play an important role in linguistic study as well as in several NLP tasks like Part-Of-Speech (POS) tagging, Parsing, Semantic tagging, in the parallel corpora, etc. There are numerous corpora in the literature of different languages and most of them are created for a specific purpose. Hence it is obvious that a researcher cannot use any corpus for their particular task. This paper also focuses on an automated technique to create a COVID-19 corpus dedicated to the research in linguistic aspects because of the pandemic situation.

Keywords: Corpus, COVID-19, Language, Automated.

1 Introduction

The term corpus originates from Latin and it means body. The word form corpus is singular and its plural word form is corpora. Generally, a corpus is a large collection of texts of a specific language [1]. In general, a corpus can be in the form of written text, spoken utterances, audio, video, or images as well as any combination of them. Most importantly, a corpus should be stored in a machine-readable format, so that it will be suitable for updating, modification, and the computer processing [2-3]. In the corpus linguistics disciplines, the corpora can be used for various linguistics investigations such as lexicology, lexicography, grammar, stylistics, sociolinguistics, as well as in diachronic and contrastive studies [1-3]. In several cases, the corpus contains other than textual data, so the task can be extended beyond text processing. Other than the linguistic aspect the corpus can be used in the development of the education system, in medical sciences, in several applications of machine learning, machine translation, etc. [2]. From history, it shows that a corpus can be used as a resource for studying changes in the human culture over time [5].

The corpus is not just the collection of texts, but it must satisfy some properties [1, 4]. Some of these features are outlined below:

- **Quality**: The data source must be authentic and the source should be mentioned.
- **Representation**: An ideal corpus is the representation of a language. So it should contain the data which is capable of representing the language diversity. The representativeness in the corpus is ensured

Proceedings DOI: 10.21467/proceedings.115; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-947843-5-7

by considering two factors – balancing (corpus data is to be selected covering various genres, domains, and media) and proper sampling.

- **Accessibility**: The texts should be formatted properly so that it can be retrieved easily as per the user's requirement.
- **Augmentation**: There should be the provision to update and modify the corpus while needed.
- **Documentation**: The information regarding corpus data should be maintained separately. It will be helpful for corpus management.

During a corpus creation phase, several issues are to be considered and it depends on the type of the corpus like text or speech [2, 4]. For example, the different factors regarding the text corpus development process are to be considered like, encoding scheme, size, targeted user, time and space, source, and sampling technique. The rest of the paper is organized as follows: Section 2 describes the related work, our contribution is presented in section 3, section 4 explores the methodology used in our corpus development, section 5 gives the statistics of the corpus in detail. The coverage and validation of corpus data are included in section 6, and finally, section 7 describes the conclusions part of our work.

2 Related work

In the past, many new corpora have been developed as well as some of the existing corpora have been upgraded for effective use in the area of NLP research and applications. Some of such important corpora are given below. Koeva et al., 2012 [2] described several statistics-related issues (such as corpus size, balance and representativeness, extended metadata and linguistic annotation, etc.) to be considered for developing a corpus suitable to use in NLP and applied them for up-gradation of the Bulgarian National Corpus (BulNC). Sarkar et al., 2007 [6] created a Bangla corpus by collecting data from online and offline documents in an automated way. Chungku et al., 2011 [7] developed a text corpus in Dzongkha (national and official language of Bhutan) by collecting data from the websites, print media, and printed documents. They also annotated the corpus using automatic POS tagger. Suchomel et al., 2012 [8] developed large scale corpora for six different languages American Spanish, Czech, Japanese, Russian, Tajik Persian, and Turkish by using an efficient web crawler namely SpiderLing. Ljubešić et al., 2014 [9] presented the development process of three corpora of Bosnian, Croatian, and Serbian by using SpiderLing crawler. Ljubešić et al., 2014 [10] also developed two licensed monitor Twitter corpora one for Croatian and Serbian and the other for Slovene by using Python based tool TweetCaT. They also tried to differentiate between Croatian and Serbian Twitter users.

Adhikary et al., 2016 [11] described the structure and functionality of the web crawler CorpoMate which can be used for building a corpus from online resources. Al-Khatib et al., 2016 [12] introduced a corpus of manually annotated news editorial for the sake of argument mining. Wagner Filho et al., 2018 [13] represented the development procedure of a (freely available) large Web corpus brWaC for Brazilian Portuguese language, having 145 million sentences and 2.7 billion tokens. Nowshin et al., 2018 [14] proposed a technique to develop a Bangla to English parallel corpus by applying crowd-sourcing approach. Salam et al., 2019 [15] tried to develop the Bangladeshi National Corpus (BDNC).

In the literature, it is observed that various corpora have been designed for different languages. But there is no corpus (in Bengali) dedicated to COVID-19 pandemic situation. So in this paper, we are going to propose a methodology to develop a COVID-19 corpus for Bengali language in an automated way.

Proceedings DOI: 10.21467/proceedings.115
ISBN: 978-81-947843-5-7

Series: AIJR Proceedings
ISSN: 2582-3922

3 COVID-19 Corpus for Bengali

Here we defined our corpus as a COVID-19 corpus because this corpus is dedicated to the terms related to the COVID-19 pandemic. During the pandemic, all the focuses were related to the COVID-19 situation in different aspects and hence all the mainstream print media covered all such topics with the highest priority. Several new terms are coined and some are adopted from other languages. Some other medical-related terms are being used frequently. So, it can be assumed that, if we can consider all the COVID-19 related news articles it will cover all the terms related to the COVID-19. Based on that assumption our corpus is created by collecting the COVID-19 related news article from a leading Bengali newspaper "Anandabazar Patrika" in the time-span of 17th January, 2020 to 13th March, 2021. This paper has reported the 1st COVID-19 related news article on 17th January, 2020 and continuing till date.

4 Methodology

To create the COVID-19 corpus, we have collected the data from the news articles (in the time-span of 17th January, 2020 to 13th March, 2021) published in the well-known Bengali newspaper "Anandabazar Patrika". After downloading the news articles from the website, we have done some cleaning in the raw data and prepared the text suitable to include in our COVID-19 corpus. We have used the following steps (shown in the Figure 1) sequentially for that work.

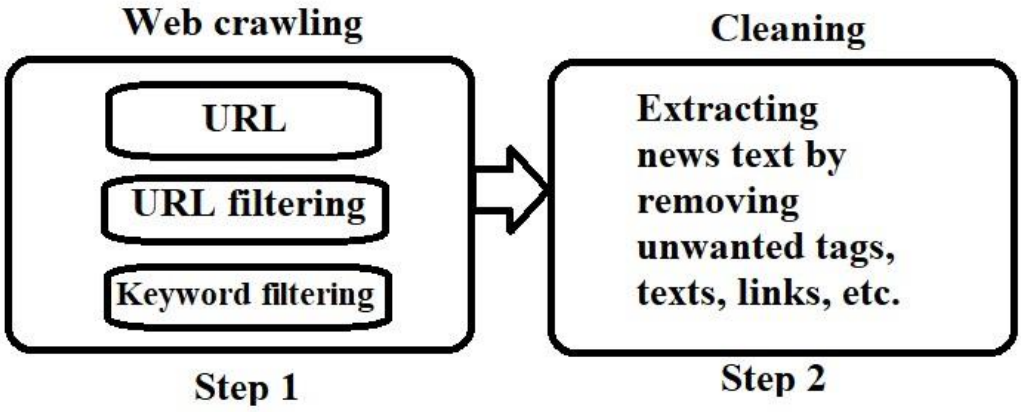

Figure 1: *Block diagram of the system used in COVID-19 corpus data preparation*

4.1 Downloading or Web crawling

- To perform this step, first we have written a web crawler using the Python library Beautiful Soup. If we pass the URL (Uniform Resource Locator) of newspaper page with multiple news articles' links to the web crawler and it returns all the URLs present in that page. Using this web crawler, we have prepared a list of URLs to download the earlier dated news articles.
- As we want to download all the corona related news articles, so we have used a filter to pick up only the URLs related to the corona news. To write the filter, we have prepared a list of corona related keywords {"corona", "coronavirus", "corona-virus", "novel-coronavirus", "covid", "covid19", "covid-19", "lockdown", "pandemic", "containment", "vaccine", "vaccination", "comorbidity", "antidote", "quarantine", "community-transmission", "covaxin", "covishield"}. The filter selects and returns only the URLs containing at least one of the above enlisted keywords.

- Now we download the news articles from the selected URLs by using the web crawler and save them as separate .txt files in the day-specific folder. As we have collected the relevant news articles of 422 days (in the time-span of 17th January, 2020 to 13th March, 2021), so we have stored .txt raw news data files in 422 folders where each folder stores all the news articles' files for a particular day and each .txt file stores the raw news data of one article of that day.

4.2 Cleaning

- In this step, we take each of the .txt files containing the raw data of the news articles and perform the cleaning operation. Through the cleaning process, we remove the HTML code, headers, footers, unwanted links, and advertisement texts etc. from the raw data.
- After removing all the unnecessary things, we get the details (headline, author, place, and published date) and actual content of the news article. Now we save that content in a .txt file according to the format (shown in the Figure 2) using UTF-8 encoding.
- Actually we take all the .txt files from a day-specific folder, then we clean each of them one by one and save the cleaned text into the day-specific .txt file. As we have collected 422 days' news articles, so finally we get 422 .txt files where one file contains all the news articles of a specific day.

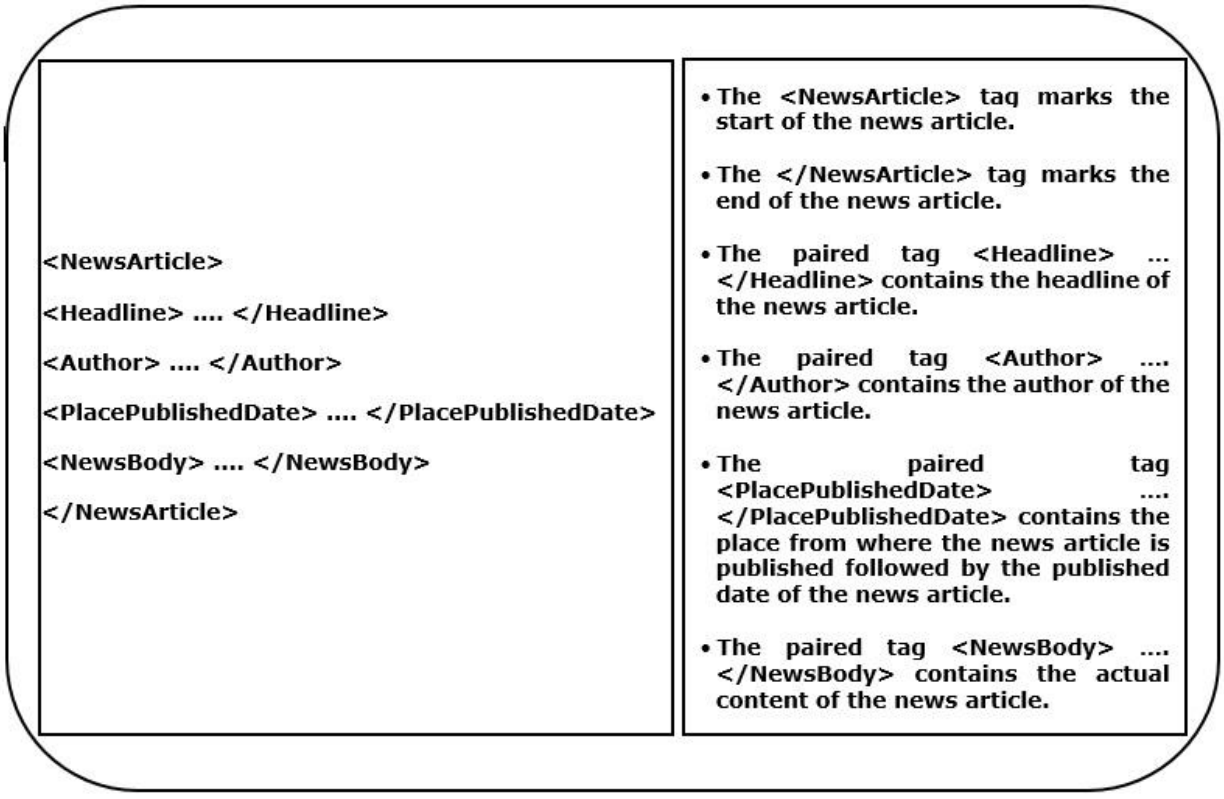

Figure 2: *Format to store each article in a .txt file*

5 Size of the corpus

We have developed our corpus by collecting all the corona related news articles (in the time-span of 17th January, 2020 to 13th March, 2021) from the well-circulated Bengali newspaper "Anandabazar Patrika". The volume summary of our corpus is shown in the Table 1.

Proceedings DOI: 10.21467/proceedings.115
ISBN: 978-81-947843-5-7

Series: AIJR Proceedings
ISSN: 2582-3922

Table 1: *Volume summary of our corpus*

Sl. No.	Heading	Value
1	Total number of files	422
2	Total number of news articles	13,024
3	Total number of sentences	4,35,003
4	Total number of tokens	48,15,666
5	Total size of the corpus	82 MB

6 Coverage and validation

The confusion matrix is used to evaluate the performance of a classifier (both binary as well as multiclass). The general form of the confusion matrix for a binary classification problem is shown in the Table 2.

Table 2: *General form of the confusion matrix*

n = total no. of examples		**Predicted**		
		Yes	**No**	
Actual	**Yes**	TP	FN	Total no. of actual positive examples
	No	FP	TN	Total no. of actual negative examples
		Total no. of predicted positive examples	Total no. of predicted negative examples	

The confusion matrix is associated with four terms: TP (True Positive), FN (False Negative), FP (False Positive), and TN (True Negative). We use the confusion matrix to validate the performance of our Python based crawler used in downloading the news articles. Therefore, the confusion matrix is used to validate our corpus data where each article is considered as an example. In our case, the individual meaning of each of the four terms (TP, FN, FP, and TN) is interpreted in the Table 3.

Table 3: *Our own interpretation of a confusion matrix's associated terms*

Term	**Meaning**	
	Corona Related Article (Actual)	**Downloaded In Our Corpus (Predicted)**
TP	Yes	Yes
FN	Yes	No
FP	No	Yes
TN	No	No

The confusion matrix as per our own interpretation is shown in the Table 4.

Table 4: *Our own interpretation of confusion matrix used to evaluate downloader or corpus data*

n = total no. of news articles		**Downloaded In Our Corpus**		
		Yes	**No**	
Corona Related Article	**Yes**	TP	FN	Total no. of corona related news articles
	No	FP	TN	Total no. of non-corona related news articles
		Total no. of downloaded news articles	Total no. of not downloaded news articles	

Actually, we attempted to download all the corona related news articles (in the time-span of 17th January, 2020 to 13th March, 2021). Now for 3 days (05th March, 2021; 6th March, 2021; and 7th March, 2021), we have evaluated the performance of the downloader by using our own interpretation of confusion matrix (shown in the Table 5, Table 6, and Table 7 respectively).

Table 5: *Performance evaluation of the downloader for 5th March, 2021*

n = total no. of news articles		**Downloaded In Our Corpus**		Date: 05/03/2021
		Yes	**No**	
Corona Related Article	**Yes**	**8** (TP)	**0** (FN)	Total no. of corona related news articles = 8
	No	**2** (FP)	**37** (TN)	Total no. of non-corona related news articles = 39
		Total no. of downloaded news articles = 10	Total no. of not downloaded news articles = 37	

In the above table (Table 5), we note 2 news article as False Positive (FP). As the URLs of these 2 news articles contain some of the corona related keywords (which have been used to filter the URLs of news articles), so these 2 news articles are downloaded in our corpus, but actually they are not corona related news articles.

Table 6: *Performance evaluation of the downloader for 6th March, 2021*

n = total no. of news articles		**Downloaded In Our Corpus**		Date: 06/03/2021
		Yes	**No**	
Corona Related Article	**Yes**	**9** (TP)	**0** (FN)	Total no. of corona related news articles = 9
	No	**0** (FP)	**57** (TN)	Total no. of non-corona related news articles = 57
		Total no. of downloaded news articles = 9	Total no. of not downloaded news articles = 57	

Proceedings DOI: 10.21467/proceedings.115
ISBN: 978-81-947843-5-7

Series: AIJR Proceedings
ISSN: 2582-3922

Table 7: *Performance evaluation of the downloader for 7th March, 2021*

n = total no. of news articles		**Downloaded In Our Corpus**		Date: 07/03/2021
		Yes	**No**	
Corona Related Article	**Yes**	**9** (TP)	**1** (FN)	Total no. of corona related news articles = 10
	No	**0** (FP)	**49** (TN)	Total no. of non-corona related news articles = 49
		Total no. of downloaded news articles = 9	Total no. of not downloaded news articles = 50	

In the above table (Table 7), we observe 1 news article as False Negative (FN). The URL of the news article does not contain any of the corona related keywords which have been used to filter the URLs of the news articles. As a result, although the content of the news article is corona related, but the news article is not downloaded in our corpus. This same validation technique can be applied for the news articles of any random date. In that case, we need to collect the relevant information (total no. of corona related news articles, total no. of non-corona related news articles) for that specific date.

7 Conclusions

From the literature it has been found that so far there is no corpus dedicated to COVID-19. So, our corpus is the first COVID-19 corpus (in Bengali) to the best of our knowledge. This corpus will help the linguists in their researches. For example, they can use our corpus to study the neologisms (words or phrases which are added to vocabulary of the language causing a relatively rapid language change) in the Bengali language due to COVID-19.

References

[1] N. Indurkhya, F. J. Damerau, *Handbook of natural language processing* vol. 2, pp. 163-177, CRC Press, 2010.

[2] S. Koeva, I. Stoyanova, S. Leseva, T. Dimitrova, R. Dekova, E. Tarpomanova, "The Bulgarian National Corpus: Theory and practice in corpus design.", *Journal of Language Modelling*, vol. 0, no. 1, pp. 65-110, 2012.

[3] R. Ali, M. A. Khan, I. Ahmad, Z. Ahmad, M. Amir, "A State-of-the-art review of corpus linguistics Journals", Jan, 2011.

[4] N. S. Dash, *Corpus linguistics: An introduction*, Pearson Education India, 2008.

[5] J. M. Twenge1, W. K. Campbell, B. Gentile, "Changes in Pronoun Use in American Books and the Rise of Individualism, 1960-2008", *Journal of Cross-Cultural Psychology*, vol. 44, no. 3, pp. 406-415, Aug, 2012, doi: 10.1177/0022022112455100

[6] A. I. Sarkar, D. S. H. Pavel, M. Khan, "Automatic Bangla corpus creation", BRAC University, 2007

[7] C. Chungku, J. Rabgay, P. Choejey, "Dzongkha Text Corpus", in *Conference on Human Language Technology for Development*, pp. 34-38, May, 2011

[8] V. Suchomel, J. Pomikálek, "Efficient Web Crawling for Large Text Corpora", *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, pp. 39-43, Apr, 2012

[9] N. Ljubešić, F. Klubička, "{bs,hr,sr}WaC – Web corpora of Bosnian, Croatian and Serbian", in *Proceedings of the 9th Web as Corpus Workshop (WaC-9) @ EACL*, pp. 29-35, Apr, 2014

[10] N. Ljubešić, D. Fišer, T. Erjavec, "TweetCaT: a tool for building Twitter corpora of smaller languages", in *Proceedings of LREC*, pp. 2279-2283

[11] A. Adhikary, S. Ahmed, "CorpoMate: A framework for building linguistic corpora from the web", *19th International Conference on Computer and Information Technology (ICCIT)*, IEEE, pp. 367-370, Dec, 2016

[12] K. Al-Khatib, H. Wachsmuth, J. Kiesel, M. Hagen, B. Stein, "A News Editorial Corpus for Mining Argumentation Strategies", *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 3433-3443, Dec, 2016

[13] J. A. Wagner Filho, R. Wilkens, M. Idiart, A. Villavicencio1, "The brWaC Corpus: A New Open Resource for Brazilian Portuguese", in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, May, 2018

[14] N. Nowshin, Z. S. Ritu, S. Ismail, "A Crowd-Source Based Corpus on Bangla to English Translation", *21st International Conference of Computer and Information Technology (ICCIT), IEEE*, pp. 1-5, Dec, 2018

[15] K. M. A. Salam, M. Rahman, M. M. S. Khan, "Developing the Bangladeshi National Corpus - a Balanced and Representative Bangla Corpus", in *2019 International Conference on Sustainable Technologies for Industry 4.0 (STI), IEEE*, pp. 1-6, Dec, 2019

Proceedings DOI: 10.21467/proceedings.115
ISBN: 978-81-947843-5-7

Series: AIJR Proceedings
ISSN: 2582-3922