==
Language in India www.languageinindia.comISSN 1930-2940 Vol. 18:2 February 2018
India's Higher Education Authority UGC Approved List of Journals Serial Number 49042
==

Creation and Compilation of Hindi Newspaper Text Corpus

Vandana and Niladri Sekhar Dash
Indian Statistical Institute

==

Abstract

Developing a corpus for the study of various aspects of a language is a highly challenging task which involves effective planning and implementation of the same. The prime concern in the development of a corpus is the overall design criteria. In this chapter we aim at presenting some theoretical guidelines on the design criteria of a one million words digital corpus of Hindi Newspaper Text Corpus (HNTC) which has been developed as a part of an on-going research activity. After the determination of the planning stage a comprehensive description of the various steps involved in the development of the corpus is discussed. An overview of the developed corpus is also highlighted with detailed specifications. Since the developed corpus has to be used subsequently for various kinds of linguistic analysis, it has been documented efficiently. This chapter also tends to give importance to documentation, storage and management of the developed corpus as it requires extreme care on the part of the corpus builder. It is a highly tedious task. Proper documentation of the corpus will ensure it authenticity and retrievability. Also, it will be utilizable for a wider range of potential areas in future.

Keywords: Corpus, Compilation, Hindi, Newspaper, Documentation

1. Introduction

The development of text corpus in Indian languages began with the generation of the Kolhapur Corpus of Indian English (KCIE) which was designed by Shastri (1988) in an effort at individual level to identify the types of similarity and difference among American English, British English and Indian English. From then onwards several attempts may have been made to develop corpora for all major Indian languages at the individual level but these are not much appreciated or attested in the history of corpus generation and application in India.

The next most important milestone in this route is the TDIL (Technology Development of Indian Languages) project which was initiated in early 1990s by Department of Electronics (DoE), Ministry of Communication and Information Technology (MCIT), Govt. of India in 1991. It was launched with a mission for developing corpora in electronic form in all Indian languages included in the 8th Schedule of the Constitution of India for subsequent works of language technology (Dash 2008). The Central Institute of Indian Languages (CIIL), Mysore was entrusted with the responsibility for coordinating the corpus development task on behalf of the MCIT as well as developing required tools and systems for conversion of the corpus into Unicode format as well as for its storage, management, dissemination, and utilization by interested researchers. The CIIL has collaborated with Lancaster University, UK for these tasks (Baker, McEnery 2003).

Generation of corpora in Indic languages has certain unique challenges associated with script and text representation in these languages. When compared with English or other advanced languages, the Indian languages may be considered as resource-poor languages, as there is hardly any sophisticate tool or technology available that may be easily used to develop Indian languages corpora in digital form covering texts of various disciplines and subject domains. However, recent advancement in computer technology as well as availability of more Indian language data in electronic form have paved new ways into corpora development, processing and their utilization in various domains of descriptive linguistics, language technology, and applied linguistics (Dash, 2009).

We briefly describe the design criteria of the Hindi newspaper corpus in Section 2, and report about the design criteria for collecting the texts for the newspaper corpus in Section 3; we present the steps involved in the development of the corpus. Section 4 focuses on the issues of proper documentation and storage of the corpus. Finally, in Section 5, we conclude with a focus on importance of the developed corpus and also highlighting the future research and possible directions of corpus-related activities in this country.

2. Design Criteria of Hindi Newspaper Corpus

The research work started with a careful planning stage where the design principles for the corpus were decided keeping in mind the purpose for which the corpus was being developed (Dash, 2005). These established a number of selection criteria as follows:

2.1 Representativeness

It contained representative texts of Hindi Newspapers in written form. The newspapers texts were available in electronic mode and were extracted from the web archives of the leading Hindi dailies in India. Since the newspapers have a lot of contents, the three major categories that are sampled are: Headlines, Full articles and Editorials.

2.2 Source Selection

Web has been taken as the major source for collection of texts for the corpus. The archives of the leading newspaper present on the web have been collected manually.

2.3 Number of Newspapers

There are more than 20 online Hindi dailies published from across the country. So, we considered the websites of 4 prominent dailies based on their readership and published from different parts of the country. The list of Newspapers based on their readership is given below (Table 2.1) (Figures are compiled by media Research Users Council (MRUC) in the Indian Readership Survey (IRS) 2014)

Newspaper	Location	Readership(in millions)
Dainik Jagran	Various cities and states	16.631
Amar Ujala	Various cities and states	7.808
Prabhat Khabar	Jharkhand, Bihar & West Bengal	2.988
Navbharat Times	Delhi, Mumbai & Lucknow	2.736

Table.2.1: List of Newspapers with their readership

The major reasons for the selection of these newspapers were:

[1] These four newspapers are one of the widely circulated and leading Hindi newspapers in India and the websites of these newspapers were easily accessible and subsequently the archives were easily available for these four newspapers.

[2] The texts available from the archive sections of the online editions of these dailies were already present in a Unicode compatible form. So, it required comparatively less manual labour as it is readable by computers and does not require scanning a document and making it useful for research purpose.

2.4 Size of the Corpus

Corpus size is incredibly important in terms of the richness of the corpus data. It was decided that the size of the corpus for the present study would be within one (1) million words; else it would difficult for our further research works.

2.5 Time Span

For the generation of corpus, a fixed time span of 10 years (2007-2016) was decided. The reasons for this selection are mentioned below:

[1] As the main purpose behind the development of this corpus was to study the structure and properties of Hindi language as used in the present-day Hindi newspapers, this time span was considered sufficient for gathering insights about the present state of the language.

[2] Possibility of finding documents in electronic form (in digital archives) was easier which solved problems of manpower, time and cost.

2.6 Types of News Articles

Newspaper contains different types of articles. We decided to collect news articles from the 13 different genres, namely the followings:

[1] National news

[2] Political News

[3] International news

[4] Business News

[5] Sports news

[6] Regional news

[7] Education news

[8] Entertainment news

[9] Lifestyle and culture news

[10] Health news

[11] Science & Technology News

[12] Weather news

[13] Editorials

3. Development of the Corpus

Once the basic design principle for creation of the corpus was the planned, the development procedure of the corpus began. We adopted various methods and strategies for developing the corpus. The primary stages were the followings:

3.1 Data Collection and Sampling

The collection of data for building a well-balanced corpus had to be fairly methodological, as the purpose for which the corpus was developed should be served in all ways. So, a specific strategy for collection of data was adopted in which we decided to follow the basic random sampling technique. We decided to randomly pick any two Sundays from every month of each four newspapers and collected data from each of the 13 genres (as mentioned in the design criteria) covering the time span of 10 years as per the availability of data in the archive section of each online newspaper selected. In this way we tried to ensure balance in text representation in the corpus as almost all possible samples from each genre were represented in the corpus. The diagram below (Fig. 2.1) gives an idea about the way the Hindi news text data had been collected for the development of the corpus:

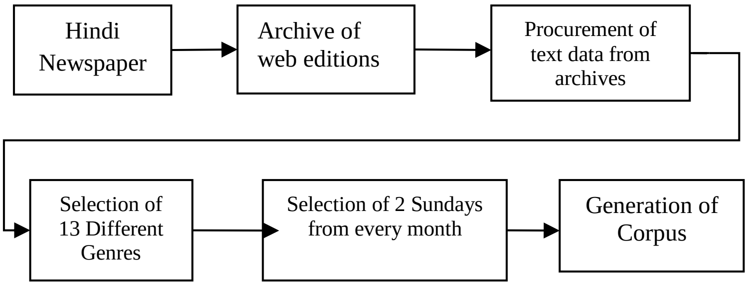

Hindi Newspaper
Archive of web editions
Procurement of text data from archives
Selection of 13 Different Genres
Selection of 2 Sundays from every month
Generation of Corpus

Fig.2.1: Method of Data collection

In case of editorials this method had not been followed. While collecting data from editorials we performed some purposeful sampling processes keeping in mind our research objective.

3.2 Amount of Data Collection

The news articles were extracted from archives of each year depending on their availability. Due to this reason, the amount of data collected based on years and genres varied across the corpus.

The amount of data collected based on year is represented through the bar diagram (Fig. 2.2) in which the x-axis represents the year of data collection and y-axis represents the amount of data collected.

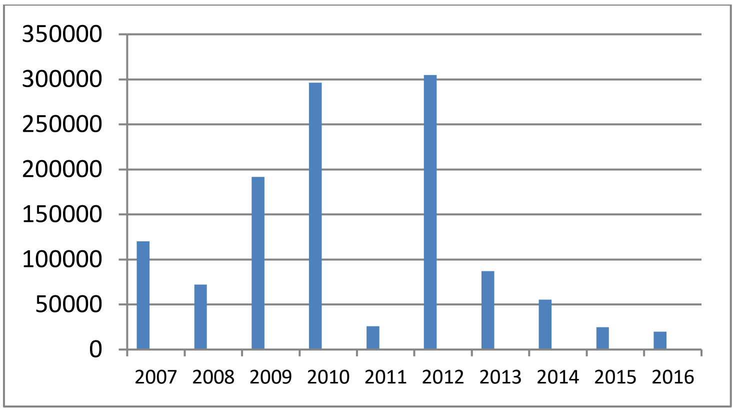


Fig. 2.2: Year-wise Data Collection

From the above graph we can see that the year 2012 shows the highest amount of data collection as the data collection process started in that year and the web editions were easily available for the newspaper '*Dainik Jagran*' from which the data was extracted. However , for the same newspaper when we tried to extract the data from the archived web edition for the year 2011 we managed to get a minimum amount of data. On the other hand for year 2010 and 2009 respectively large amount of data was collected as for these two years the data was taken from the newspaper '*Amar Ujala*' for which the archive of the web edition was maintained systematically. Also, the year 2007 and 2008 we could manage to retrieve a decent amount of data from the web archive of the newspaper '*Navbharat Times*'. For the year 2013 and 2014 we tried to get data from the newspaper '*Prabhat Khabar*' which allowed to extract data in a restricted way as for most of the times it was not available in a text format but in a PDF form which could not be converted into text. Therefore, for these two years also the amount of data collected was not very high. And, for the year 2015 and 2016 we only focused on collecting the editorials so, we only collected a selected amount of data from all these newspapers.

Similarly, the amount of data collected based on genre has been shown through a pie-chart (Fig. 2.3) representing the amount of data collected from all the thirteen genres.

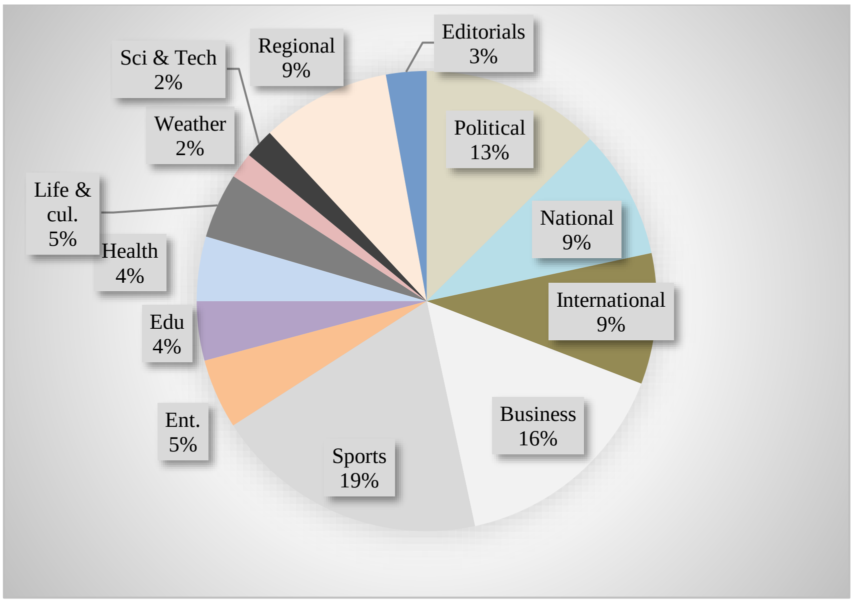


Fig. 2.3: Amount of Data Collected Genre-wise

From the above Fig. 2.3 we can observe that the data collection is not uniformly distributed across the genres. One of the major reasons contributing to this disproportion in amount of data collected is the availability of data in the web archives of the newspapers. So, the news genres that occupy the major chunk of the corpus are because of their easy availability. Additionally, there can also be some factors contributing to the high occurrence of these news items in the newspapers. Here, we see that the sports and the business genre have the highest amount of occurrence in the corpus i.e. 19% and 16% respectively. However, collectively political, national and international news cover the major component i.e. 31% of the newspaper as they are considered to be the most informative section of the newspaper with maximum readership. For example, the sports news does occur in a higher amount because it features all the information from national and international sports events to local sports events sometimes school/college tournaments with readers from all age group. Business news forms one major section as it sometimes overlaps with the national and international news items. Local and regional news also compromises about 9% of the corpus as the readers can be more interested in news events that are happening in the closet geographical proximity. While all other news genres can be very reader specific and targeted to specific readers only. For instance, the education, science and technology and entertainment news will mainly have youth and students as its readers. While the lifestyle and culture and health news can attract potential women readers along with adults. Weather news can be of importance to a specific community along the other readers who might have a glance of it. Although, editorials do form an important part of the readership but in our corpus its occurrence is low because it has been collected at a later stage with specific research questions in mind. Although, the sectional readership preference might vary from newspaper to newspaper, but these figures give a glimpse into the amount of news present in each section in a newspaper.

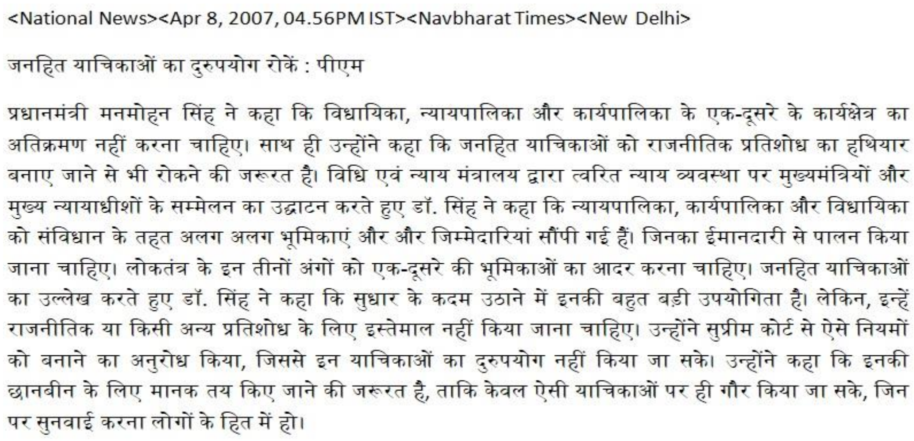
3.3 Pre-processing of the Corpus

3.3.1 Text Normalization

Upon completion of data collection, cleaning and proof reading of all data were needed to ensure the accuracy of data before they were being used for the construction of corpus. The electronic text may contain various typographical errors. So, these typographical errors had to be removed. Images and pictures were removed from the news articles as it was not required. The proper spacing between each word was checked.

3.3.2 Creation of Metadata

The last stage in the development of our corpus was the creation of metadata i.e. providing detailed descriptive information to each text. Metadata is generally defined as 'data about data'. So, descriptive header information was added to each text, giving information specific to each text, such as the name of newspaper, its year of publication, date of publication, name of the place from where the newspaper was published, name of the correspondent and whatever information is available about the text. A screenshot of a sample text (Fig. 2.4) is shown below.



<National News><Apr 8, 2007, 04.56PM IST><Navbharat Times><New Delhi>

जनहित याचिकाओं का दुरुपयोग रोकें : पीएम

प्रधानमंत्री मनमोहन सिंह ने कहा कि विधायिका, न्यायपालिका और कार्यपालिका के एक-दूसरे के कार्यक्षेत्र का अतिक्रमण नहीं करना चाहिए। साथ ही उन्होंने कहा कि जनहित याचिकाओं को राजनीतिक प्रतिशोध का हथियार बनाए जाने से भी रोकने की जरूरत है। विधि एवं न्याय मंत्रालय द्वारा त्वरित न्याय व्यवस्था पर मुख्यमंत्रियों और मुख्य न्यायाधीशों के सम्मेलन का उद्घाटन करते हुए डॉ. सिंह ने कहा कि न्यायपालिका, कार्यपालिका और विधायिका को संविधान के तहत अलग अलग भूमिकाएं और और जिम्मेदारियां सौंपी गई हैं। जिनका ईमानदारी से पालन किया जाना चाहिए। लोकतंत्र के इन तीनों अंगों को एक-दूसरे की भूमिकाओं का आदर करना चाहिए। जनहित याचिकाओं का उल्लेख करते हुए डॉ. सिंह ने कहा कि सुधार के कदम उठाने में इनकी बहुत बड़ी उपयोगिता है। लेकिन, इन्हें राजनीतिक या किसी अन्य प्रतिशोध के लिए इस्तेमाल नहीं किया जाना चाहिए। उन्होंने सुप्रीम कोर्ट से ऐसे नियमों को बनाने का अनुरोध किया, जिससे इन याचिकाओं का दुरुपयोग नहीं किया जा सके। उन्होंने कहा कि इनकी छानबीन के लिए मानक तय किए जाने की जरूरत है, ताकि केवल ऐसी याचिकाओं पर ही गौर किया जा सके, जिन पर सुनवाई करना लोगों के हित में हो।

Fig. 2.4"Screenshot of Metadata

In the above Fig. 2.4 we can see that the news item has been given a proper description in terms of its genre, year, date, time and place of publication. Also, the name of newspaper from where the news article has been extracted in also mentioned in the metadata. In case, where we have additional information as in the name of the news reporter or the author the metadata contains this extra information also. By storing the news articles with all these descriptive information, it becomes much easier to retrieve specific news articles for future use.

3.4 Overview of the Developed Corpus

After following the design criteria and its effective implementation at all the stages of the development of the corpus we finally have the one-million word corpus of the Hindi newspaper texts. The Table 2.2 below gives the overview of the developed corpus. (Khan, Sobhan 2012)

-

Table2.2: Overview of the Corpus

Source	**No. of Words**	**Time Span**
Navbharat Times	1,20,030	Jan -Dec 2007
Navbharat Times	72,029	Jan -Nov 2008
Amar Ujala	1,91,506	Jan - Dec2009
Amar Ujala	2,96,187	Jan - Aug 2010
Dainik Jagran	25,841	Sep - Dec2011
Dainik Jagran	3,04,880	Jan – Dec 2012
Amar Ujala	2338	Jan & Sep 2013
Dainik Jagran	80,583	May – Dec 2013
Prabhat Khabar	4091	January 2013
Amar Ujala	924	January 2014
Dainik Jagran	54,583	Jan – June 2014
Prabhat Khabar, Amar Ujala	25,000	Mar –Dec 2015
Prabhat Khabar, Dainik Jagran	20,000	Jan-Dec 2016
All above four dailies	**11,20,338**	**Jan 2007-June 2016**

After the appropriate storage of the all the files in a proper format and encoding we have a database of around 11, 20,338 words which meets the requirements of our target of developing a million-word corpus of Hindi Newspapers

4. Documentation of the Corpus

Once the corpus had been developed it required utmost care and had to be handled efficiently so that the corpus could be used in the future by other potential users. (Wynne, 2005). To ensure this, various steps were taken for the proper documentation of the corpus. The major steps involved in this process were:

4.1 Appropriate Storage of the Corpus

The development of digital resource like the present corpus has become relatively easier with more electronic texts being freely available and accessible but to ensure its use in the future itshould be properly stored depending on the needs of the research. The present corpus was stored in the following ways:

[1] The raw data has been stored in .txt format in utf-8 encoding. If the corpus was made of the files in any other format such as Microsoft Word, then they could not be processed by most of the corpus analysis tools. Also, storing the data in this form would help to use the data in future.

[2] All .txt files have been arranged systematically in separate folders with date as the folder name. The date format was used in this order: date/month/year respectively.

4.2 Preservation of the Corpus

The role of documentation is very crucial in the generation of a corpus from digital resources. It plays very important role in preserving the corpus for its future availability and usability. It provides accountability and authenticity to the corpus. The various procedures we followed to ensure the preservation of the corpus was:

[1] The corpus has been created in the preservation version i.e. the raw data without any annotation. And, this raw data without any annotation or extra information has been stored separately as the preservation version.

[2] We have also created backup copies of the data resource after the development of the corpus which have been stored separately on a CD-ROM and the data has also been stored in the Google drive(cloud storage) to ensure they are unaffected by any natural or man-made disasters.

5. Conclusion

In this present paper we gave a detailed description on our strategy and methodology for the development of the HNTC. As, there seems to be relatively less work done in development of corpora for all Indian languages when compared to other languages of the world. This work can contribute largely to the ongoing efforts to develop corpora for all Indian languages.Furthermore, the information extracted from this study may be used for comparison with corpora of other texts of Hindi. As, in newspaper corpus of Hindi can be compared with corpus of Hindi literary texts, corpus of Hindi scientific texts etc. Analysis along these lines has the potential to facilitate intra-linguistic interface which is an expanding area in the field of machine translation. The data can also be used for inter-language comparison and studies of the Hindi newspaper texts will texts of other languages as Bengali, Urdu, Marathi, Punjabi etc. thus, being of great help in developing parallel corpus for Indian languages. This data from this work can be used to develop a lexical database of Hindi, as in the development of word-net of Hindi. There can be various types of analysis as in sense-marking etc. which can form a fundamental element in various research works of this kind. The future direction of the present research will on the study of discourse patterns in the Hindi newspaper texts as they are a common form of written discourse. Consequently, this study will be applicable in the study of pragmatics of this language. In addition to all these this work will aim to contribute towards a general understanding of language as a human phenomenon.

==

References

Baker Paul, McEnery Tony et.al (2003) Constructing Corpora of South Asian Languages. Paper presented at Corpus Linguistics 2003, 2003-03-01, Lancaster.

Baker, Paul, Hardie A. & McEnery (Ed.). 2003. Corpus data for south Indian languages . In the proceedings of the EACL workshop on South Asian languages. Budapest.

Dash, Sekhar Niladri (2005) Corpus Linguistics and Language Technology: With Reference to Indian Languages, New Delhi: Mittal Publications

Dash, Sekhar Niladri (2008) Corpus Linguistics: An Introduction, New Delhi-Pearson Education-Longman

Dash, Sekhar Niladri (2009) Language Corpora: Past, Present and Future, New Delhi: Mittal publication

Wynne, M (editor). 2005. Developing Linguistic Corpora: a Guide to Good practice.Oxford: Oxbow Books. Available online from

http://ota.ox.ac.uk/documents/creating/dlc/

Khan, Sobhan et al. 2017. Creation and Analysis of a New Bangla Text Corpus BDNC01.In International Journal for Research in Applied Science & Engineering Technology (IJRASET). Volume 5 Issue XI, November 2017- Available at www.ijraset.com

==

Vandana Mishra
Senior Research Fellow
PhD Candidate
vandana.mishra87@gmail.com

Niladri Sekhar Dash
Associate Professor
ns_dash@yahoo.com

Linguistic Research Unit
Indian Statistical Institute
203 B.T Road
Kolkata-700108
West Bengal
India