Annals of Library and Information Studies
Vol. 68, December 2021, pp. 406-410

Resolving the confusion of the authorship attribution of a Bengali book

Siladitya Jana[a] and Satyaki Mazumder[b]

[a]Assistant Librarian, Indian Institute of Science Education and Research, Kolkata, Email: siladitya.jana@iiserkol.ac.in

[b]Assistant Professor, Department of Mathematics and Statistics (DMS), Indian Institute of Science Education and Research, Kolkata, Email: satyaki@iiserkol.ac.in

Received: 25 August 2021; accepted:10 November 2021

The present paper aims to determine whether the Bengali book *Londoner Naksa ebong France Bhraman* (*Wondrous Capers at London and Travelling in France*) was written by the geologist Pramathanath Bose (P.N. Bose). To find it out, two well-established style markers often used in authorship attribution studies; namely, function words and punctuation marks, are used here. The result shows that possibly this book was penned by the geologist P.N. Bose. As a corollary, it may also be added that this approach may be used in future authorship attribution studies involving Bengali writings.

Keywords: Authorship Attribution; Bengali Language; P.N. Bose

Introduction

In the long history of authorship attributions studies, English language texts reigned supreme. Authorship attribution works in other languages are not as common as in English. A specific case study with the authorship of a particular book, it was possibly never done before for any texts in Bengali. The present work undertakes to study a Bengali book, titled *Londoner Naksa ebong France Bhraman* (*Wondrous Capers at London and Travelling in France*)[1] whose authorship is uncertain. It was claimed that this book was written by geologist Pramathanath Bose (popularly known as P.N. Bose) (1855-1934)[2]. However, it may be added here that there is no well-established proof of this claim. The present work aims to determine whether geologist P.N. Bose was the author of this book with contentious authorship.

P.N. Bose was the first native Indian graded Geologist at the Geological Survey of India (GSI). It was an officer-level post[3]. Bose was born in a village named Gaipur. This place is about sixty kilometres away from Calcutta (now Kolkata)[4]. Bose's early education started at the village school. In 1864, Bose took admission at the Krishnagar Collegiate School (in Nadia district of West Bengal, India). He completed his Secondary (matriculation) education at the age of fifteen. But, due to the prevalent rule of the University of Calcutta, he was not allowed to sit for the Entrance Examination until he attained the age of sixteen. So, he was forced to wait for one year to appear in the examination. In 1871, he completed the Entrance Examination at the University of Calcutta. Bose completed F.A. (First Examination in Arts, equivalent to Intermediate Examination) from Krishnagar College in 1873. Then he took admission at St. Xavier's College, Calcutta, to study B.A. During this period, it was one of the most prominent colleges in India. This college was established in 1860 by Belgian Jesuits[4].

Bose sat in the "Gilchrist Prize Scholarship" examination in 1874 and stood first. In October, he took admission in B.Sc. at the University of London. He completed B.Sc. in 1878. He received the *Edward Forbes Medal and Prize* for his B.Sc. result. He took admission at the Royal School of Mines (now part of Imperial College London) in the same year. In 1879, he applied to the Secretary of India in England for a suitable job at home[4].

On 13 May 1880, he joined the Geological Survey of India (GSI) at its London office. He returned to India on 30th July 1880 and joined the Department of Geology, Government of India as Assistant Superintendent (at that time, GSI was part of the Department). For twenty-four years (1880-1903), he served GSI. During 1884-1893, he published thirteen research papers in the *Records of GSI* and one memoir (as a separate book)[4]. He introduced the study of micro-section as an aid to petrological work at GSI[5]. He found out several new sources of important

minerals in India[4]. Another important work by Bose was his identification of carbonatites as igneous rocks and their origin by re-mobilisation of limestone. However, his GSI superiors did not recognise this. It was recognised much later in the 1960s[6]. As a writer, apart from scientific works, he wrote several books and numerous articles on various socially relevant issues.

Canning Library, Calcutta published a book entitled *Londoner Naksa ebong France Bhraman* (*Wondrous Capers at London and Travelling in France*) in 1291 B. S. (1884). The author of this book was mentioned as Pramatha Nath Bose. However, this book was not mentioned in Bose's own writings and his biographies by Jogesh Chandra Bagal[7] and Manoranjan Gupta[8]. We have not found any mention of this book in the autobiography written by P.N. Bose's wife Kamala Bose[9]. Sen and Dan[4], in their edited volume on the Bengali writings of Pramatha Nath Bose, have given an exhaustive bibliography of both his Bengali and English writings. However, they also have not mentioned this book as Bose's work. Even then, Ghosh[2], in his edited volume *Pramatha Nath Basu*: *Barnamay Adhyay* (*Pramatha Nath Basu*: *Eventful Life*), has included this book as Bose's. So, there is confusion about the authorship of this book. Hence, it is imperative from both the academic and intellectual point of view to determine whether this book was written by the geologist P.N. Bose. The present work is an endeavour in this direction.

Objectives of the study

- To find out whether geologist P.N. Bose was the book's author with contentious authorship *Londoner Naksa ebong France Bhraman* (=*Wondrous Capers at London and Travelling in France*);
- To check the suitability of two important style markers: function words and punctuation marks in Bengali authorship studies.

Sample preparation

Book chapters and articles were taken as samples to study Bose's works. The aim was to compare Bose's writings as available in this book with his known writings in Bengali. Dan and Sen[4] compiled his known Bengali writings in *Bangla Rachana Sankalan* (*Collection of Bengali Writings of Pramathanath Bose*).

For the work under discussion, four book chapters and six articles of 750 words each were randomly selected from the book with contentious authorship and the edited volume, respectively. Then, the samples from the book were compared with the articles from the writings style's point of view. It may be noted that all the samples used here were written in the Bengali language. As the work is with texts of literary nature, it may not be possible to get a sample of precisely 750 words. In all cases, a sample was used with a complete sentence. This may have resulted in having some samples with little more or less than the sample size of 750 words.

The reason why 750 words was chosen is because most of the chapters in the book are less than 800 words. But in some cases, we were able to find chapters with around 750 words but less than 800 words. Chapters will less than 750 words were not selected.

List of the samples

Four chapters (marked as PNB1, PNB2, PNB3 and PNB4) were taken from the book under discussion. The book is divided into two parts. The first part of the book is called *Londoner Naksa* (*Wondrous Capers at London*). The second part is entitled *France Bhraman* (*Travelling in France*). Two chapters were taken from the London part and two from the Paris part. The selected book chapters are as follows:

1. *Amar Basa o Chakrani* (*My Rented Place and the Maid Servant*)
2. *Grihakartri* (*Landlady*)
3. *Jadughar, Bijnancharcha* (*Museum and the Cultivation of Science*)
4. *Jatiya-swabhab* (*National Character*)

These four chapters were compared with six articles (marked as PNB5…PNB10) by Bose. These articles are as follows:

1. *Hindudharmer Nabajiban* (*Reemergence of Hindu religion*). *Nabajiban* 1, 466-471 and 547-551,1291 B. S. (=1884)
2. *Kencho* (*Earthworm*). *Bharati* 8:309-315, 1291 B.S. (=1884)
3. *Upay Ki* (*What is the Way Out?*). *Bharati* 12:301-308, 1295 B.S. (=1888)
4. *Bangala Bhasay Bijnan Siksha* (*Science Education Through the Bengali Language*). *Bharati* 14:341-351, 1297 B.S. (=1890)
5. *Himalaye Ekti Nihar Bahur Pase* (*Beside a Glacier in Himalaye*). *Bharati* 15:137-141, 1298 B. S. (=1891)
6. *Bharate Bilati Sabhyata* (*Impact of Western Civilisation on India*). *Bharati* 15:404-408, 1298

B. S. (=1891) and 16:432-437, 1299 B. S. (=1892) (Bose 2008))

The importance of function words in these kinds of studies lies in the fact that they are both topic-independent and context-free[10]. The same can be said about punctuation marks. Two style markers, namely, function words and punctuation marks are used in this work. Mosteller and Wallace[11] probably used it first in authorship attribution studies. Later, it was used by several others[12-16]. Most probably, Mascol[17,18] pioneered the use of punctuation marks as a style marker in authorship attribution works. Later, other researchers also applied it in their works[19-21]. A later work tested the efficacy of both these style markers in a single work[22].

However, these works were pertaining to English language texts. In recent years, several researchers have undertaken work on Bengali authorship attribution studies. They used several approaches in this regard. Character-level signals were used in one such approach[23]. Others used the unigram and bi-gram features of the language along with the richness of the Bengali vocabulary[24] in this regard. Other researchers used the authorship classification approach using character n-grams, feature selection for authorship attribution, feature ranking, and analysis[25]. Several other researchers in different manner[26,27] also used the classification approach.

Methodology

Six most recurrently used function words were identified and taken for the analysis of the samples. The function used words are *ki*, *o* (কি), *ar* (আর), *kintu* (কিন্), *ebong* (এবং), *je* (যে). These words along with their frequency in the selected 10 sample texts were placed in a workbook. Later, a programme was used for chi-square test in Matlab(R) and the workbook was imported in the software for calculation. The calculation was done to check whether there are similarities and dissimilarities amongst and between the different text samples. The calculation was done to find out the chi-square values (X^2) and their corresponding p values. The null hypothesis (H_0) for the present work was "from the same group", and if the p-value is less than 0.05, the null hypothesis was rejected at a 95% level of significance. After that, an analysis of the result was undertaken.

In the selected samples, it was found out that Bose used four punctuation marks frequently. These were identified and used in this work. These are danri (|; equivalent to a period (.) in the English language), comma (,), semicolon (;), and question mark (?). The process mentioned above for the function words was followed here for punctuation marks as well. Then data were collected about how these punctuation marks were used in the sampled ten texts. After that, following the process under the function words, a calculation was done to determine the chi-square values (X^2) and their corresponding p values of all the samples. The null hypothesis and the related norms are the same as above.

Findings

Function words

The first four samples (PNB1 - PNB4) are from the book with contentious authorship. The first two samples (PNB1 and PNB2) are from the first part of the book, i.e., that deals with the travel account of the London city. PNB3 and PNB4 are from the second part of the book, which deals with his travelling experience in France. The collected data are presented in Table 1.

From the first row of Table 2, it is observed that the sample texts from the book with contentious authorship are matching with each other based on the pattern of function words usage. It proves beyond doubt that these chapters are from the same book, penned by the same author. The second row shows the result of comparison between the book chapters and the articles known to be written by Bose. The result shows that these two sets of writing are like each other. From this, it may be inferred that these two sets of writings are produced by the same person, the geologist P.N. Bose. The third row shows the result of the calculation amongst the articles

Table 1 — Frequency of function words in the samples

Samples	ki (কি)	o (ও)	ar (আর)	Kintu (কিন্)	ebong (এবং)	je (যে)
PNB1 (L)	4	15	4	3	3	3
PNB2 (L)	3	12	1	10	4	13
PNB3 (F)	6	17	2	4	5	4
PNB4 (F)	3	16	2	8	0	4
PNB5	1	9	0	8	4	15
PNB6	8	7	3	7	5	6
PNB7	5	17	2	5	15	6
PNB8	3	8	2	5	8	11
PNB9	8	8	6	4	2	6
PNB10	2	6	1	4	5	12

Table 2 — Results of using function words as a style marker

Texts	Samples	Matching (%)	Not Matching (%)	Comments (based on p values)
Book Chapters	PNB1-PNB4	100	0	Similar
Book Chapters: Known Articles	PNB1-4 : PNB5-10	58	42	Similar
Known Articles	PNB5-PNB10	73	27	Similar

Table 3 — Frequency of punctuation marks in the samples

Samples	period	comma	semicolon	question marks
PNB1	63	52	8	1
PNB2	59	72	16	6
PNB3	47	64	12	2
PNB4	49	73	22	5
PNB5	54	67	2	1
PNB6	55	51	12	8
PNB7	55	50	6	3
PNB8	55	63	2	0
PNB9	63	54	3	4
PNB10	60	49	7	1

Table 4 — Results of using punctuation marks as a style marker

Texts	Samples	Matching (%)	Not Matching (%)	Comments (based on p values)
Book Chapters	PNB1-PNB4	83	17	Similar
Book Chapters: Known Articles	PNB1-4 : PNB5-10	54	46	Similar
Known Articles	PNB5-10	87	13	Similar

known to be written by Bose. The result corroborates that fact.

Punctuation marks

The collected data for punctuation marks are presented in Table 3.

The results of this calculation are available in Table 4.

From the perspective of the book chapters, it may be observed that they are matching amongst themselves to a high degree (Table 4). It shows that those are written by the same author. On the other hand, from the second row, we come to know that the book chapter whose author is uncertain and the articles whose author, i.e., Bose is certain are matching to a large extent. From it, we may infer that possibly, both sets of samples are authored by the same person. Lastly, the result of the third row of the Table 4 confirms that Bose penned these articles.

Conclusion

Our work shows that the style of writing is the same in both the book *Londoner Naksa ebang France Bhraman* (*Wondrous Capers at London and Travelling in France*) and the other known writings of the geologist P.N. Bose. This result is obtained using two well-established style markers: function words and punctuation marks in authorship attribution studies.

Hence, it may be concluded that in all probability, the geologist P.N. Bose was the author of the book with contentious authorship *Londoner Naksa ebang France Bhraman* (*Wondrous Capers at London and Travelling in France*). From this, it may also be said that the technique used in this study may be used for finding the authorship of Bengali texts with contentious authorship in future studies.

Acknowledgements

Acknowledgments are due to Sri Subir K. Sen (since deceased), Prof. B. K. Sen, formerly with the University of Malaya, Malaysia and Sri Ashish Lahiri, eminent science writer.

References

1 Bose P N, Pramatha Nath Basu*:* Barnamay Adhyay (ed.Ghosh, Debiprasad), Drisi; Kolkata, 2007

2 Ghosh D, Katatuku Pramathanath, In *Pramatha Nath Basu*: *Barnamay Adhyay* (ed. Ghosh, Debiprasad), Drisi, Kolkata,2007, p. 4.

3 Roy S, Pramatha Nath Bose-the First Indian Graded Geologist. *Science and Culture*, 78 (2012) 37-39.

4 Dan D K and Sen S K, Pramatha Nath Basu: Sankshipta Jibani, *In Bangla Rachana Sankalan* (eds. Dan D K and Sen S K), *Pramatha Nath Bose*, (Presidency Library; Kolkata), 2012, p. 115-124. (In Bengali)

5 Saha M, Foreword, *In* Bagal, Jogesh Chandra, *Pramathanath Bose*, (P.N. Bose Centenary Committee, New Delhi), 1955, p. xiii-xviii.

6 Heinrich, E W, *The Geology of Carbonatites,* (Krieger, Florida), 1980.

7 Bagal, J C, *Pramathanath Bose*. (P.N. Bose Centenary Committee, New Delhi), 1955.

8 Gupta M, *Pramatha Nath Bose*. (Bangiya Bijnan Parisahd, Kolkata), 1962. (In Bengali)

9 Bose K, *Swargiya Saddhi Kamala Debir Atmajibani* (Mrs. P. K. Sen, Kolkata), (n.d.) (In Bengali)

10 Stamatatos E, A survey of modern authorship attribution methods. *Journal of the American Society for Information Science & Technology*, 60 (2009) 538–556.

11 Mosteller F and Wallace D L, Inference and disputed authorship: the Federalist, (Addison Wesley, Reading, MA), 1964.

12 Argamon S and Levitan S, Measuring the usefulness of function words for authorship attribution, 2005. Available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.71.6935&rep=rep1&type=pdf (Accessed on 19th August 2021).

13 Koppel M, Schlier J and Argamon S, Computational methods in authorship attribution. *Journal of the American*

Society for Information Science & Technology, 60 (2009) 9–26.

14 Miranda G A and Calle M J, Function words in authorship attribution studies, *Literary and Linguistic Computing*, 22 (2007) 49–66.

15 Stamatatos E, A survey of modern authorship attribution methods, *Journal of the American Society for Information Science & Technology,* 60 (2009) 538–556.

16 Zhao, Y, and Zobel, J, Effective and scalable authorship attribution using function words. In Proceedings of the Second AIRS Asian information Retrieval Symposium (eds. Lee G G, Yamada, A. Meng, H. and Myaeng, S.H.), (Springer, Berlin), 2005, p. 174–189.

17 Mascol C, Curves of Pauline and pseudo-Pauline style I, *Unitarian Review*, 30 (1888) 452–460.

18 Mascol C, Curves of Pauline and pseudo-Pauline style II, *Unitarian Review*, 30 (1888) 539–546.

19 Chaski C E, Empirical evaluation of language-based author identification techniques, *Forensic Linguistics*, 8 (2001) 1–65.

20 Mingzhe J and Jiang M, Text clustering on authorship attribution based on the features of punctuations usage. In *11th International Conference on Signal Processing, ICSP 2012*, 3, 2175–2178. Available at http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6492012&isnumber=6491878 (Accessed on 19th August 2021).

21 O'Donnell B, Stephen Crane's *The O'Ruddy*: a problem in authorship discrimination. In *The Computer and Literary Style* (Ed. Leed, J.) (Kent State University Press, Kent, Ohio), 1966, p. 107–115.

22 Jana S, Sister Nivedita's influence on J. C. Bose's writings, *Journal of the Association for Information Science and Technology*, 66 (2015) 645-650.

23 Khatun, A, Rahman, A, Islam, M S, and Marium-E-Jannat, Authorship Attribution in Bangla literature using Character-level CNN. In *22nd International Conference on Computer and Information Technology (ICCIT)*, 2019, p. 1-5. https://ieeexplore.ieee.org/document/9038560 (Accessed on 28th July 2020)

24 Das S and Mitra P, Author identification in Bengali literary works. In *Pattern Recognition and Machine Intelligence. PReMI 2011* (eds. Kuznetsov, S.O., Mandal, D.P., Kundu, M.K., and Pal, S.K.). *Lecture Notes in Computer Science*, 6744, (Springer, Berlin), Available at https://doi.org/10.1007/978-3-642-21786-9_37 (Accessed on 19th August 2021)

25 Phani S, Lahiri S and Biswas A, Authorship attribution in Bengali language. *In Proceedings of the 12th International Conference on Natural Language Processing*, 2015, p. 100-105. Available at https://www.aclweb.org/anthology/W15-5915.pdf (Accessed on 19th August 2021)

26 Chowdhury HA, Imon MAH and Islam MS, Authorship attribution in Bengali Literature Using fastText's Hierarchical Classifier. In *4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT)*, Dhaka, Bangladesh, 2018, p. 102-106, Available at https://ieeexplore.ieee.org/document/8628109 (Accessed on 19th August 2021)

27 Hossain MT, Rahman MM, Ismail S and Islam MS, A stylometric analysis on Bengali literature for authorship attribution. In *20th International Conference of Computer and Information Technology (ICCIT)*, Dhaka, 2017, p. 1-5, Available at https://ieeexplore.ieee.org/document/8281768 (Accessed on 19th August 2021)