Chapter 13
The Bengali Literary Collection of Rabindranath Tagore:
Search and Study of Lexical Richness

Suprabhat Das
Indian Institute of Technology Kharagpur, India

Anupam Basu
Indian Institute of Technology Kharagpur, India

Pabitra Mitra
Indian Institute of Technology Kharagpur, India

ABSTRACT

Rabindranath Tagore is one of the most prolific authors of Bengali literature. He has added a vast amount of richness in style and language to the Bengali text. The present study aims at a quantitative study of vocabulary size and lexical richness as well as effective search engine for his works. Several statistical measures of term distribution have been used to measure lexical richness. An initial attempt has been made to build a search engine, Anwesan, for Rabindra Rachanabali collection. The first complete digital Rabindra Rachanabali released by Society for Natural Language Technology Research, Kolkata, in 2010, has been used in the study. It was observed that a high lexical richness value was characteristics of most of Rabindranath Tagore's work.

INTRODUCTION

One of the most prolific writers in Bengali literature is Nobel laureate Rabindranath Tagore (May 7, 1861 - August 7, 1941). He had dominated both the Bengali and Indian philosophical and literary scene for decades. He was a social reformer, patriot and above all, a great humanitarian and philosopher. He had modernized Bengali art by changing its rigid classical forms. He was the ambassador of Indian culture to the rest of the world. For his eternal writing *Gitanjali*, he was awarded the Nobel Prize for Literature in the year of 1913, becoming the first Asian Nobel laureate.

DOI: 10.4018/978-1-4666-3970-6.ch013

He is the only litterateur who penned anthems of two countries, *Jana Gana Mana*, the Indian national anthem and *Amar Shonar Bangla*, the Bangladeshi national anthem.

Different statistical techniques in stylistic analysis of literary texts have been studied for long time. An empirical law to estimate vocabulary size from collection size, which is known as Heaps' law (Heaps, 1978), is now becomes a benchmark in the field of information retrieval, though it is not well-known in linguistics. Besides that, there are multinomial Bayesian approaches (Boender & Rinnooy Kan, 1987) and few essential but rarely followed procedures (Nation, 1993) to estimate the vocabulary size. Many lexical richness measures have also been studied and applied on English and other languages for years. Different measures of lexical richness were applied on the data from the works of three contemporary French singers (Ratkowsky & Hantrais, 1975), the volumes of the *Travaux de Linguistique Quantitative* (TLQ) series (Ratkowsky, 1988), which was initiated by Swiss publishing firm Slatkine in 1978, Biblical texts (Holmes, 1994) and sixteen works from eight English authors (Tweedie & Baayen, 1998) to study different lexical styles. The hidden connections in the medical literature have also been reported using lexical statistics (Lindsay & Gordon, 1999 May). The corpora of three playwrights, Euripides – a great tragedian of classical Athens, Aristophanes – a comic playwright of ancient Athens, and Terence – a playwright of the Roman Republic, was studied to compare the trends in vocabulary richness over time (Smith & Kelly, 2002). The number of different types in the first fifty thousand words in each text from the twelve texts of twelve different authors along with the effects of text-doubling and text-combining on measures of vocabulary richness have been reported (Hoover, 2003). Besides that, the studies of vocabulary richness have been done for child language and second language research to monitor changes in children and adults with vocabulary difficulties. Primarily type/token ratio was used to measure lexical diversity in child language research (Richards, 1987). After that, different advanced measures in child language and second language have been reported by many researchers (Bogaards & Laufer-Dvorkin, 2004; Haznedar & Gavruseva, 2008; Richards & Malvern, 2000 September). There is a large body of research works on information retrieval methods, including several commercial search engines for English speaking users. There are search engines for the literary works of Shakespeare.

There were no major works on statistical analysis as well as search engines for Bengali literary works. We have made an initial attempt on Rabindra Rachanabali collection to study vocabulary size and different lexical richness measures. Various measures of lexical richness have been computed for different genres of Rabindra Rachanabali collection and different chronological intervals. The statistical measures are also compared with the measures from another Bengali author Bankim Chandra Chattopadhyay. We also build a search engine, Anwesan, for Rabindra Rachanabali collection.

The rest of the paper is organized as follows: The details of the Rabindra Rachanabali collection along with available metadata are given in next section. How the vocabulary size is correlated with Heaps' law and the analysis of estimated vocabulary size using Heaps' law are described in the following sections. Next sections describe a brief overview of different measures of lexical richness and the analysis of lexical richness, chronological study and comparative study with another Bengali author. After that, the details of Anwesan, its architecture, components, advanced features and usage statistics have been given. In the last section, we have concluded about our evaluation result and some other features are also discussed that can be included in future works for the betterment of our research work.