Culturomics On A Bengali Newspaper Corpus

Shanta Phani
Dept of IT
BESU, Shibpur
Howrah 711103, West Bengal, India
shantaphani@gmail.com

Shibamouli Lahiri
Dept of Computer Science and Engineering
Penn State University, University Park
PA 16802, USA
shibamouli@psu.edu

Arindam Biswas
Dept of IT
BESU, Shibpur
Howrah 711103, West Bengal, India
abiswas@it.becs.ac.in

***Abstract*—We introduce culturomic studies on a leading Bengali newspaper corpus - Ananda Bazar Patrika, in the same spirit as [15]. Based on 11 years' worth of Bengali newswire text, we are able to extract trajectories of salient words that are of importance in contemporary West Bengal. To the best of our knowledge, this is the first time a culturomic trend analysis is being performed on an Indic language. As a result of our analysis, we obtain interesting insights into word usage and cultural shift in contemporary West Bengal. Moreover, we model culturomic trajectories using ARIMA and obtain word usage predictions that closely follow actual usage patterns.**

I. Introduction

Bengali culture and language enjoy a unique place in the world, shaped by millennia of assimilation, conflicts and refinement. From the very early kingdom of Shashanka (ca. 7th century AD [28]) to modern age, the culture of Bengal has been continuously enriched by many hundreds of poets, artists, musicians, litterateurs and critics. Bengali language has, at the same time, experienced profound changes. A language that started its journey as a mere offshoot of the now-defunct *Magadhi Prakrit*, has gradually become one of the richest Indic languages spoken by 230 million people across the world [26].

Today, Bengali is spoken not only by the people of Bangladesh and West Bengal, but also by many people in the Indian states of Tripura and Assam [26]. Moreover, Bengali is predominantly spoken among the first-generation Bengalee immigrants in other countries. The language continues to change and evolve. New terms are introduced, and old ones are gradually replaced. Capturing the temporal trend of such changes is a challenging task. It is further compounded by the fact that in recent years, globalization has introduced new terms and cultural dimensions that were previously unheard of.

In this paper, we attempt to address this challenge by restricting our attention to the changes West Bengal has experienced in the last decade (2001-2011). We further restrict our attention to a leading Bengali daily that enjoys wide circulation in West Bengal - the Ananda Bazar Patrika (ABP)[1]. Based on 11 years' worth of Bengali newswire text from ABP, we are able to extract some very interesting information. For example, in the same spirit as [15], we have created *culturomic trajectories* of salient words found in newswire text. For each salient Bengali word or bigram, we have created a vector of normalized term frequency for every month in the 11-year period. Once those vectors are created, we plot them and analyze their temporal trend (Section III-B).[2]

Then we move one step further and model the culturomic trajectories using ARIMA [25]. Here our goal is to predict the usage of a particular term in a future month. As we will show in Section IV, our ARIMA models are able to give predictions that closely follow the actual usage pattern of words. Note, however, that we only have 11 years' worth of data, as opposed to 200 years of English text analyzed in [15]. So we train our ARIMA models on the first ten years (120 months), and test them on the last one year (12 months). Since we only focused on salient words and terms that are of importance in contemporary West Bengal, we do not yet have a comprehensive database of culturomic trajectories of *all* words and n-grams. Building that database is part of our future plan (Section V).

II. Related Work

The Google Culturomics Project[3] and its flagship endeavor - Google N-gram Viewer[4] - constitute the pioneering step in the area of *Culturomics*. With the help of automatic book scanning and advanced Optical Character Recognition (OCR) technology, Google has successfully digitized millions of books across the world [15]. Once digitized and OCR-ed, the next step was to observe *temporal trends* in the usage of words. It was generally observed that topical words tend to follow a particular event, rise sharply just after the event occurs, and then gradually drops [15].

Following Google Culturomics Project, other researchers applied the general idea of Culturomics to a variety of domains. Culturomics has been used in tracking emotion [5, 16], word semantics [24], writing style [10, 17], and popular music [20]. It has also been used in mining marketing history [2], human behavior [12], institutional identities [21, 22], and scientific terms [1]. Moreover, researchers have combined culturomics with random fractal theory [8], and proposed statistical laws governing fluctuation of word usage patterns [19].

[1] http://www.anandabazar.com/

[2] Complete code and data available at http://www.4shared.com/archive/Fa4P5bxe/ialp_2012_code_and_datatar.html

[3] http://www.culturomics.org/

[4] http://books.google.com/ngrams/

Equally varied are the corpora on which culturomic analyses have been performed. Researchers have conducted culturomic studies on scientific articles [9, 21], newswire text [3, 7, 12], literary text [17], commit logs [14], and even fairy tales [16, 23]. Two groups have also suggested conceptual extension to the basic idea of culturomics. The first one, *Hedonometrics* [5], deals with tracking happiness. The second one, *computational history* [3], follows historical events on a large newswire corpus.

While all these studies are extremely important, none have so far looked into *Indic languages*. The main roadblock that continues to stymie culturomic studies in Indic languages like Bengali comes from the shortage of large diachronic corpora, as well as the not-so-mature-yet OCR technology [18]. We circumvent these roadblocks by focusing on an online archive of Bengali newswire text available from Ananda Bazar Patrika (ABP) website.

III. Culturomic Analysis

A. Dataset

We downloaded all HTML files from Ananda Bazar Patrika (ABP) archive[5], and converted them into Unicode Bengali text. Note that ABP started using Unicode Bengali on their website from June 2011. All HTML files before June 2011 used a non-standard Bengali font proprietary to ABP, and we had to convert non-standard HTML files into Unicode HTML first.[6] We removed all kinds of advertisement, feedback pages, English characters, special characters, symbols and numbers from the HTMLs before converting them into Unicode Bengali text. The filtered text only contains news-related content starting from 1st January 2001 till 31st December 2011. It is 3997 days' worth of Bengali text spanning 11 years (132 months).

The Unicode text files were subsequently processed to obtain normalized term frequency of salient Bengali words and bigrams in each of the 132 months. We did not attempt stemming, lemmatization, POS tagging or stop word removal. Stemming, lemmatization and POS tagging models for unrestricted Bengali text are not sufficiently reliable and robust yet, and although there are studies that looked into these issues [4, 6, 11, 13], training data and resources are still scarce. And since Bengali is a highly inflected language, it has no well-defined list of stop words either. Treating most frequent words as stop words does not help because very often the most frequent words are content words. Without stemming, lemmatization and stop word removal, the unigram count of each year was obtained as shown in Table I. Note that the number of unique Bengali words increased over the years (except a small dip in 2005), indicating a possible influx of new terms and foreign words (*Bideshi Shobdo*).

TABLE I: Unigram Counts for Each Year

Year	Total Unigrams	Unique Unigrams
2001	33,829,382	200,995
2002	15,853,776	221,419
2003	18,069,992	262,542
2004	35,531,251	302,969
2005	19,354,115	272,571
2006	23,121,779	306,113
2007	22,747,540	307,130
2008	21,996,939	309,994
2009	19,386,663	311,938
2010	20,620,798	315,725
2011	19,772,468	318,239

B. Culturomic Trajectories

We studied Bengali words and bigrams that are important in contemporary West Bengal. For each salient term, we plot its normalized monthly frequency across the 132-month period from January 2001 to December 2011, and analyze the trend. To compensate for the effects of not doing stemming and lemmatization, we created vectors of normalized term frequency by treating each term as a "prefix". For example, the term frequency of *Kaaj* ("work") also includes terms like *Kaajer* ("of work") and *Kaaj-kormo* ("chores"). With this slight modification, all words that begin with *Kaaj* are treated as a single term. In the following discussion we will only use these "prefix" terms.

The first set of figures (Table II first row) explores orthographic variation of Bengali words. Note that Ananda Bazar Patrika adopted a spelling reform in 1994, and old spellings were gradually replaced with new ones. However, old variants still surface occasionally. The spelling reform affected both native and foreign words. By 2001, however, old spellings were mostly suppressed and new variants firmly established. For example, Figure 1_ta (Table II) shows two spellings of the word *SongGhaat* ("conflict") - an older spelling (*Songhaat*) and a newer spelling (*SongGhaat*). Note the extremely low frequency of the old spelling as compared to the new spelling. The same observation is repeated in Figure 1_tb, where we show two spellings of the word *Cheen* ("China") - the old spelling (*Cheen*) and the new spelling (*Chin*).

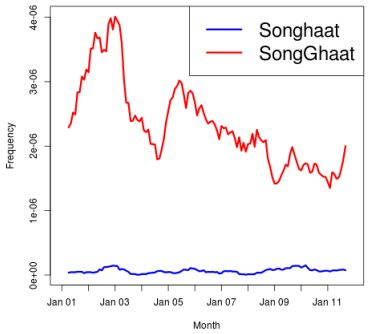
A different trend is observed in Figure 1_tc, where we have shown three orthographic variants of the word "Gandhi" - the original spelling (*Gandhee*), and two new spellings (*Gnadhi* and *Gandhi*). Note that the original spelling (*Gandhee*) dominates the other two variants. We conjecture that this discrepancy comes from our "prefix" method. The old spelling *Gandhee* still enjoys wide usage in words like *Gandheemurti* ("Gandhi statue"), *Gandheejoyonti* ("Gandhi's birthday"), etc. Since these words are coalesced with *Gandhee* into a single term, the overall frequency remains high.

The second row of Table II talks about politics in West Bengal. There are three most prominent political parties - CPM, Congress and Trinamool (Figure 2_ta in Table II). After some initial ups and downs, frequency of all three parties increased from January 2007 to January 2010, then stabilized with Trinamool having the highest frequency. As if on a cue, Trinamool-chief *Mamata Bandyopadhyay* superseded the erstwhile chief minister *Buddhadeb Bhattacharya* in popularity (Figure 2_tb) and was elected chief minister of West Bengal

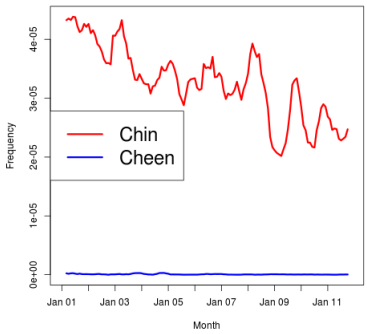
[5]Available at http://anandabazar.com/archive/, as of December 1, 2011.

[6]We used the Python proxy from http://anandabazar-unicode.appspot.com/.

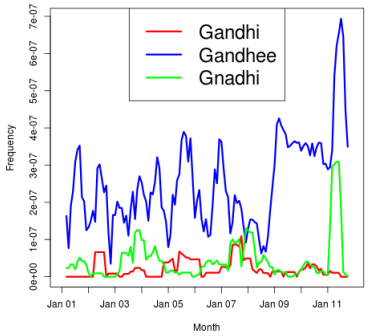
TABLE II: Culturomic trajectories



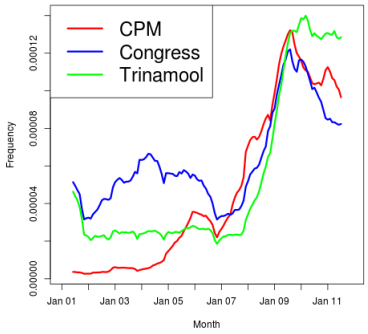
1_ta. *Songhaat* ("conflict") - smoothing 7



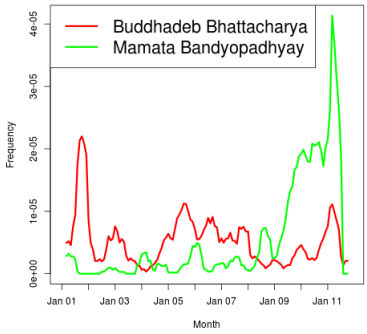
1_tb. *Cheen* ("China") - smoothing 5



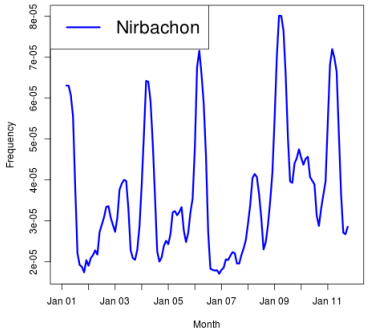
1_tc. *Gandhi* - smoothing 5



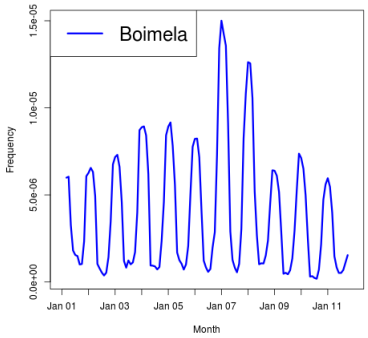
2_ta. Political parties - smoothing 11



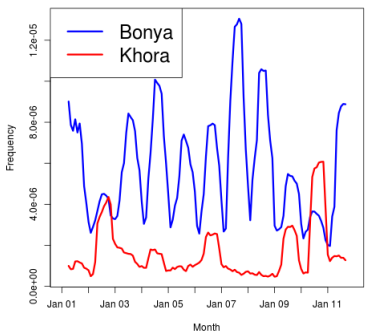
2_tb. Chief ministers - smoothing 5



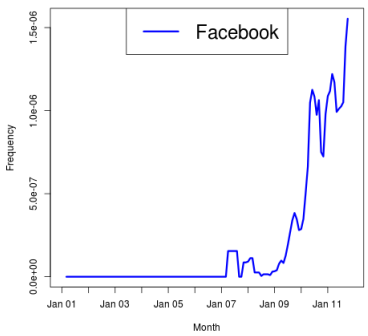
2_tc. *Nirbachon* ("election") - smoothing 5



3_ta. *Boimela* ("book fair") - smoothing 5



3_tb. Natural disaster - smoothing 7



3_tc. *Facebook* - smoothing 5

in 2011 assembly election. The word "election" (*Nirbachon*) itself shows a periodic trend (Figure 2_tc). Its stronger peaks align with Lok Sabha and West Bengal assembly elections that happen every five years, and weaker peaks correspond to other elections (like municipality or Panchayat elections).

West Bengal is also known for its *Boimela* ("book fair"). The most celebrated *Boimela* - Kolkata Book Fair, is organized at the end of January. As Figure 3_ta shows, there is a strong peak every January and the trajectory is clearly periodic. Periodicity is also observed in natural disasters. Every year, some parts of West Bengal suffer from severe flood. As we note from Figure 3_tb, *Bonya* ("flood") repeats every year with strong peaks in July-August. Droughts (*Khora*), however, are much less prevalent as well as somewhat irregular. *Khora* does not have strong peaks, and we found that its frequency increased every four years, as opposed to *Bonya* ("flood").

Now we look into the influx of "new culture" as a result of globalization. Note from Figure 3_tc that the term *Facebook* has been gaining popularity since mid-2007, and the popularity has been increasing ever since, at a steady rate. This validates our earlier proposition (cf. Section III-A) that Bengali language is experiencing an influx of new words. The rate of assimilation (Figure 3_tc has a sharp increasing trend) also hints at the possibility of an overhaul of the Bengali lexicon in near future.

We have been able to extract many more trajectories like the ones shown in Table II. Here we only present the most

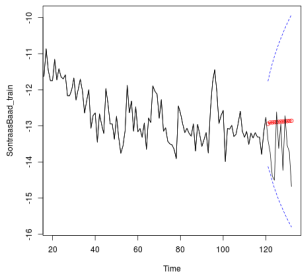


Fig. 1: Forecast of last one-year usage of *SontraasBaad*.

salient ones due to space constraints.[7] One important point to note, however, is that the trajectories give us a clear idea about the *temporal trend* of natural, socio-political, cultural and economical events in contemporary West Bengal.

IV. Time Series Modeling

Modeling the temporal trend of these culturomic trajectories is important because the models allow us to *forecast* future events. With this view in mind, we move one step further and model time series trajectories with ARIMA [25]. We found two studies that looked into time series modeling of culturomic trajectories - [8] and [19]. But they did not *forecast* future usage of words - a feature we introduce.

Recall that our data comes from 132 months of Bengali newswire text (cf. Section III-A). In other words, for each salient Bengali term, we have a time series vector of 132 consecutive values. We use the first ten years (120 months) for estimating our ARIMA model parameters, and last one year (12 months) for testing the model (i.e. forecasting the word usage in last one year). If the forecast is close to actual word usage in the last one year, then we can conclude that our models are good.

We used the Box-Jenkins Method [27] for estimating ARIMA model parameters. This involves plotting the data, taking log transforms if necessary, differencing and seasonal differencing, and plotting auto-correlation (ACF) and partial auto-correlation (PACF) functions of the transformed data[8]. At the end of Box-Jenkins procedure, we determine whether a particular time series is fit for ARIMA modeling, and if it is fit, then what the model parameters are. Note that there are seven parameters (p, d, q, P, D, Q, S). Once the parameters have been estimated, we train our ARIMA model on the first 120 months and test it on the last 12 months.

Here we present our analysis of the Bengali word *Sontraas-Baad* ("terrorism"). For *SontraasBaad*, the estimated ARIMA parameters are: $p = 1$, $d = 1$, $q = 1$, $P = 0$, $D = 0$, $Q = 0$, $S = 1$. After ARIMA modeling, we get forecast of last one year as shown in Figure 1. In Figure 1, the solid line represents original time series trajectory, red circles represent forecasts of the last 12 months, and the blue dashed lines represent 95% confidence interval around the predicted values in last 12 months. Note that the actual usage of *SontraasBaad* in last 12 months lies well within the 95% confidence interval, and is actually quite close to the predicted values (red circles of Figure 1). We obtained similar results for many other trajectories, but include only one for space constraints.[9]

[7] All trajectories are available at http://www.4shared.com/archive/Fa4P5bxe/ialp_2012_code_and_datatar.html

[8] We used R for parameter estimation, modeling and forecasting.

V. Conclusion

In this paper we introduced, for the first time, a culturomic study on an Indic language (Bengali). Based on 11 years' worth of Bengali newswire text, we were able to plot and analyze culturomic trajectories of salient words and bigrams. We also modeled the trajectories using ARIMA and obtained word usage predictions that closely followed actual usage patterns.

We plan to implement a Bengali n-gram viewer in the same spirit as Google n-gram viewer. To accomplish this idea, we need a comprehensive database of all Bengali n-grams used in the 132 months. We are currently working on building this database. Another extension to our present work would be to include other Bengali newspapers from West Bengal, Bangladesh, Tripura and Assam, and observe if the temporal trends of word usage differ significantly among newspapers. As a future work we also plan to incorporate other types of genre, e.g. Bengali literary text, into our study.

References

[1] Relative Trends in Scientific Terms on Twitter [v0]. http://altmetrics.org/workshop2011/uren-v0/.
[2] The Evolution of Marketing History: a peek through Google Ngram Viewer. http://lightofyouth.wordpress.com/2011/09/28/the-evolution-of-marketing-history-a-peek-through-google-ngram-viewer/.
[3] C.-m. Au Yeung and A. Jatowt. Studying how the past is remembered: towards computational history through large scale text mining. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 1231–1240, New York, NY, USA, 2011. ACM.
[4] S. Dandapat, S. Sarkar, and A. Basu. Automatic part-of-speech tagging for Bengali: an approach for morphologically rich languages in a poor resource scenario. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 221–224, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.
[5] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE*, 6(12):26, 2011.
[6] A. Ekbal and S. Bandyopadhyay. Part of Speech Tagging in Bengali Using Support Vector Machine. *2008 International Conference on Information Technology*, pages 106–111, 2008.
[7] I. Flaounas. Pattern Analysis of News Media Content. Master's thesis, Univ. of Bristol, 2011.
[8] J. Gao, J. Hu, X. Mao, and M. Perc. Culturomics meets random fractal theory: insights into long-range correlations of social and natural phenomena over the past two centuries. *Journal of The Royal Society Interface*, 2012.
[9] J. Gasiorek, H. Giles, T. Holtgraves, and S. Robbins. Celebrating Thirty Years of the JLSP : Analyses and Prospects. *Journal of Language and Social Psychology*, 2012.
[10] J. M. Hughes, N. J. Foti, D. C. Krakauer, and D. N. Rockmore. Quantitative patterns of stylistic influence in the evolution of literature. *Proceedings of the National Academy of Sciences*, 109(20):7682–7686, 2012.
[11] M. Z. Islam, M. N. Uddin, and M. Khan. A Light Weight Stemmer for Bengali and Its Use in Spelling Checker.
[12] Kalev. Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, 15(9), Sept. 2011.
[13] P. Majumder, M. Mitra, S. Parui, G. Kole, P. Mitra, and K. Datta. YASS: Yet another suffix stripper. *ACM Trans Inf Syst*, 25(4), 2007.
[14] B. Michalski, M. Krishnamoorthy, and T.-Y. Lau. Temporal Analysis of Literary and Programming Prose. *CoRR*, abs/1202.2131, 2012.
[15] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, T. G. B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014):176–182, January 14 2011.
[16] S. M. Mohammad. From Once Upon a Time to Happily Ever After: Tracking Emotions in Mail and Books. *Decision Support Systems*, May 2012.
[17] A. Muralidharan. WordSeer: Exploring Language Use in Literary Text. http://hci.berkeley.edu/cs260-fall10/images/archive/c/c4/20101213064903!FinalPaper-Muralidharan.pdf.
[18] F. Y. Omee, S. S. Himel, and M. A. N. Bikas. Article: A Complete Workflow for Development of Bangla OCR. *International Journal of Computer Applications*, 21(9):1–6, May 2011. Published by Foundation of Computer Science.
[19] A. M. Petersen, J. Tenenbaum, S. Havlin, and H. E. Stanley. Statistical Laws Governing Fluctuations in Word Use from Word Birth to Word Death. *eprint arXiv:1107.3707*, July 2011.
[20] J. Serrà, A. Corral, M. Boguñá, M. Haro, and J. L. Arcos. Measuring the evolution of contemporary western popular music. *CoRR*, abs/1205.5651, 2012.
[21] D. S. Soper and O. Turel. An n-gram analysis of Communications 2000–2010. *Commun. ACM*, 55(5):81–87, May 2012.
[22] D. S. Soper and O. Turel. Who Are We? Mining Institutional Identities Using n-grams. *Hawaii International Conference on System Sciences*, 0:1107–1116, 2012.
[23] S. Weingart and J. Jorgensen. Computational analysis of the body in European fairy tales. *Literary and Linguistic Computing*, 2012.
[24] D. T. Wijaya and R. Yeniterzi. Understanding semantic change of words over centuries. In *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web*, DETECT '11, pages 35–40, New York, NY, USA, 2011. ACM.
[25] Wikipedia. Autoregressive integrated moving average, 2012. [Online; accessed 30-May-2012].
[26] Wikipedia. Bengali language, 2012. [Online; accessed 30-May-2012].
[27] Wikipedia. Box-Jenkins, 2012. [Online; accessed 9-June-2012].
[28] Wikipedia. Shashanka, 2012. [Online; accessed 30-May-2012].

[9] All prediction results are available at http://www.4shared.com/archive/Fa4P5bxe/ialp_2012_code_and_datatar.html