A Phrase Set for Bengali Text Entry Evaluations Based on Actual Text Messages

Ahmed Sabbir Arif
Ryerson University
Toronto, Ontario, Canada
a.s.arif@gmail.com

Sarah Fardeen
North South University
Dhaka, Bangladesh
sarah.fardeen@northsouth.edu

CHI'16 Extended Abstracts, May 07-12, 2016, San Jose, CA, USA
ACM 978-1-4503-4082-3/16/05.
http://dx.doi.org/10.1145/2851581.2892394

Abstract
User studies evaluating text entry techniques usually require participants to transcribe phrases. Yet at present, there is no dataset available for Bengali text entry research that includes phrases entered on mobile devices. This forces researchers to collect phrases from various sources, compromising the external validity of the work. In this paper, we present a set of Bengali phrases composed by real users on actual mobile devices. Through an analysis of the dataset, we show that it contains phrases with varying lengths, symbols, and numbers.

Author Keywords
Bengali/Bangla text entry; evaluation; mobile phone; smartphone; text messages; texts; phrase set.

ACM Classification Keywords
H.5.2 User Interfaces: Evaluation/methodology.

Introduction
Text entry techniques are usually evaluated by measuring entry speed and error rates in transcription tasks. Participants are presented with phrases of text from a corpus that they have to transcribe as quickly and accurately as possible. At first glance, it may seem

Figure 1. An interviewer interviewing a participant in private.

more appropriate to permit participants to freely input whatever they desire, because this replicates natural usage and improves the external validity of the study. The drawback of this approach is the absence of a source text to compare the transcribed text with to determine errors. In addition, such data may be contaminated with spurious behavior, such as taking a break or performing a secondary task [7]. Therefore, the most common procedure is to present participants with predetermined phrases from a set to transcribe. Two such phrase sets for English text entry research are the MacKenzie and Soukoreff set [7] and the Enron dataset [12]. Bengali text entry on mobile devices is becoming increasingly popular among native speakers, mainly because it allows them to communicate with each other without the assistance of a second language [8]. As a result, there has been a growing interest in developing novel and improved mobile text entry techniques for the Bengali language. However at present, there is no phrase set available to accommodate the evaluation of these techniques [4].

The most popular phrase set for Bengali text entry research is arguably the Central Institute of Indian Languages (CIIL) corpus [3], which collected the phrases from various sources, such as novels, textbooks, newspapers, and scientific journals. Some have also collected phrases from books, textbooks, songs, poems, quotes, and day to day conversations [4, 5, 9]. However, most of these sets were created for particular user studies, thus are not publicly available for others to use. Besides, none of these sets include messages composed by real mobile device users, thus not representative of actual mobile messages. These sets also do not account for different writing styles of Bengali (i.e., Shadhubhasha vs. Cholitobhasha [6]) and include phrases that are too long to be transcribed on mobile devices. In this paper, we present a set of Bengali phrases composed by actual users on actual mobile devices. We also analyze the set and show that it contains phrases with varying lengths, symbols, and numbers.

Data Collection

As part of a larger project, we conducted a survey of mobile text entry behaviors in Dhaka, Bangladesh. For the survey, we recruited 155 undergraduate students as interviewers from an introductory level psychology course (average age 22.5, 57% male and 43% female). We instructed them to spread out in the densest areas of the city and approach potential volunteers in public places, such as park, tea stall, university campus, etc. First, they explained our research to potential volunteers, collected their consents, and screened them for age, education, and mobile ownership. All volunteers had to be adults (19 years or older), own at least one mobile device, and have primary-level education (from grade 1 to 4) as a minimum requirement to participate in the survey. They were screened for educational background to make sure that they possess the language skills necessary to compose texts. Anyone who did not meet these criteria was excluded from the study. Then, the interviewers separated the volunteers from the crowd for a private interview (Figure 1) involving a predetermined set of questions. The interview included an *optional* question that asked participants to share their last five sent/received texts with us.

Participants

Over the period of three months the interviewers interviewed 643 mobile users. About 42% of them decided to respond to the optional question. Although

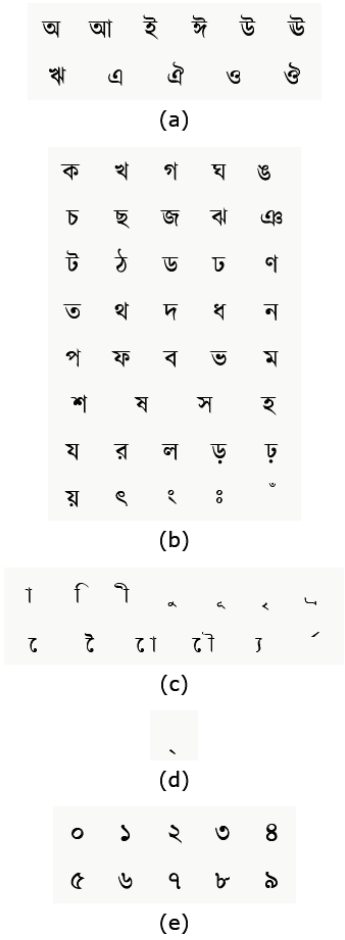

Figure 2. Bengali (a) vowels, (b) consonants, (c) matras, (d) halant, and (e) digits.

they were asked to share at least five text messages, many shared fewer than five, resulting in 987 texts in total. Participants' age ranged from 19 to 57 years, on average 26.5 years (SD = 7.6). About 74.1% of them were male and 25.9% were female. Based on the World Bank classification [13], 1.7%, 28.8%, 57.9%, 6.9%, and 4.7% of them were from the poor, low, middle, upper-middle, and high income groups, respectively. Roughly 64.5% participants owned a smartphone and the remaining 35.5% owned a feature phone. They were all frequent mobile users – that is, on average used their mobile devices for about 6.2 hours (SD = 4.9) daily.

Review Process

We wrote a simple console application with C# to automatically remove all repeated messages and messages composed of only one word and/or emoticons. We also manually reviewed the data to discard all messages that are:

- Written in a different language, such as English.
- Inappropriate for a user study, such as use coarse or offensive language.
- Contain sensitive information, such as bank account information.
- Contain information that can compromise anonymity, such as a participant's email address or phone number.
- Short codes for value-added services, such as ordering ringtones, transferring balance, and various mobile services.

This process eliminated 721 phrases from the set (about 73% of the data), resulting in 266 phrases in total. We corrected all misspellings and converted all Roman scripts to Bengali[1] in these phrases. However, some minor modifications may be necessary to convert spellings to a local dialect, such as "আব্বু" and "আব্বা", both mean "dad". Table 1 displays a selection of phrases from the final set. We have made the complete set publicly available at http://www.asarif.com/resources/bncorpus.

সাড়ে ছয়টার আগে বেরুতে পারবো না
আসলে আমি তোর কাছে আসতে চেয়েছিলাম কিন্তু পরে সময় পাই নাই, দেখি পরে আসবো
খাবার টেবিলে দেয়া আছে
বাবা, স্কুল থেকে আসার সময় অষুধগুলো কিনে এনো
শুভ নববর্ষ
সিরিয়াল নাম্বার লিখে জানিও
পরে কথা বলবো, ক্লাসে আছি
আমি আজ রাত ১১টায় বাসায় আসবো

Table 1: Eight example phrases from the final set.

Analysis of the Phrase Set

There are over sixty basic characters in Bengali [2, 11] that can be divided into the following categories:

- *Vowels:* There are eleven commonly used vowels in the Bengali language. Illustrated in Figure 1 (a).
- *Consonants:* There are thirty-nine consonants. Illustrated in Figure 1 (b).

[1] Due to the unavailability of an effective, user friendly, and easy to learn Bengali mobile text entry technique, many users write Bengali messages using Roman characters. For example, writes "Ami bhalo achi" instead of "আমি ভাল আছি", which means "I am well".

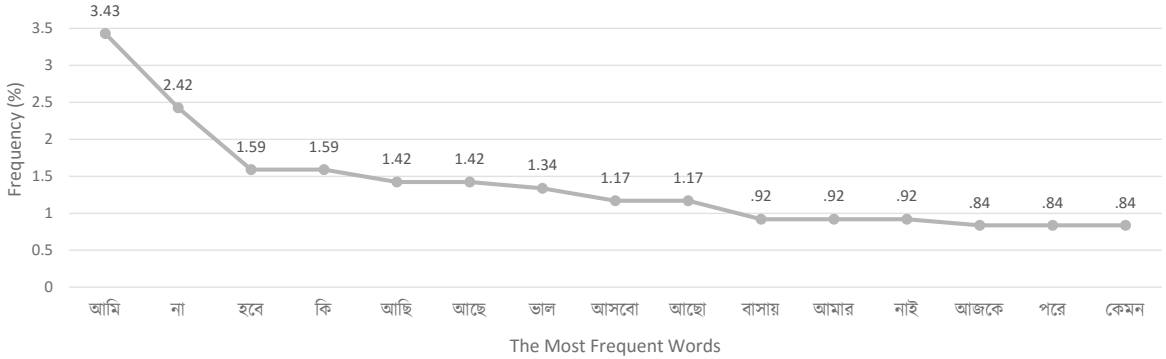

Figure 3. The most frequent words in the phrase set.

- *Matras:* There are thirteen matras. See Figure 1 (c). Matras are diacritics of the elements in set of vowels and unique to Indic languages. They appear in most Bengali words before, after, above, or below consonants. The composition of a consonant and matra is often referred to as a glyph. For example, the word "পরে", which means "later", is consisted of the consonant "প" and the glyph of the consonant "র" and the matra "ে".
- *Halant*: Apart from the above mentioned matras, there is a special matra in Bengali, called halant. See Figure 1 (d). It is used mainly to form yuktakshars or conjugate symbols. A yuktakshar is a composition of multiple basic consonants that is represented with a special composite character. Similar to matras, many Bengali words cannot be written without yuktakshars. For example, the word "রক্ত", which means "blood", is consisted of the consonant "র" and the yuktakshar of the consonant "ক", halant "্", and the consonant "ত".
- *Numeric Characters*: There are nine base numbers in Bengali, as displayed in Figure 1 (e).

Due to the complex nature of the Bengali writing system [6] and the absence of standard performance metrics for evaluating Bengali text entry techniques [4], researchers use different conventions to count characters in a Bengali phrase. Most researchers consider glyphs and/or conjuncts as individual characters, while some consider only the constituent symbols as distinct characters [9]. There are also disagreements about whether to consider halant as an individual character or not. Therefore, it is sometimes difficult to compare studies or to extract meaningful average text entry speeds and accuracy rates from this body of work. This makes it hard for researchers to use and apply these results and prevents the synthesis of a larger picture. Recently, Sarcar et al. [9] compared these conventions and showed that conventional text entry metrics (i.e.,

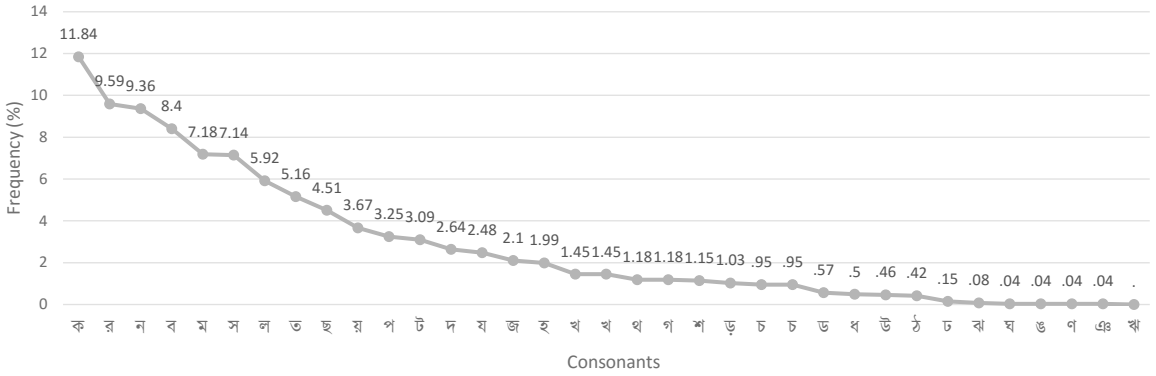

Figure 4. Frequencies of consonants in the dataset.

words per minute and error rate) are more representative of the actual performances observed with a text entry technique in a user study when only constituent symbols are considered as distinct characters. Therefore, we use this approach to analyze our phrase set. However, to provide a better comparison we also report results from a different analysis that counted glyphs and conjuncts as individual characters. Table 2 summarizes the results.

Phrases

On average, there are 23.2 constituent characters per phrase (SD = 10.6). About 54% of the phrases contain punctuation marks (63% of which are question marks) and about 5% contain numeric characters. There are on average 4.48 words per phrase (SD = 2.02), which is substantially lower than the average 10.81 words per phrase found in the CIIL corpus [1, 3], suggesting that mobile users tend to input relatively shorter phrases on mobile devices.

Total phrases	266
Phrases containing numbers	14
Phrases containing punctuations	144
Average phrase length (constituent)	23.2 (*SD = 10.6*)
Average phrase length[2]	15.7 (*SD = 7.41*)
Total words	1196
Total unique words	576
Average words per phrase	4.48 (*SD = 2.02*)
Average word length (constituent)	4.22 (*SD = 0.88*)
Average word length[2]	2.76 (*SD = 0.54*)

Table 2. Statistics about the phrase set.

[2] Results from a different analysis where we counted both glyphs and conjuncts as distinct characters (i.e., "ক্ত" was counted as one character, which is actually the composite of "ক", "্", and "ত").

Words

There are in total 1,196 words in the set, roughly 47% of which are unique. Figure 2 displays the most frequent words in the set. The average word length is 5.48 constituent characters (SD = 2.02) and 2.76 (SD = 0.54) characters[2], which is substantially lower than the average 5.11 characters per word found in the CIIL corpus [1, 3]. This indicate towards the possibility that mobile users avoid using longer words when composing texts on their mobile devices. The longest and the shortest words in the set are of length 16 and 2 constituent characters, respectively.

Characters

We compared the letter frequencies of our corpus with the letter frequencies from a study that used a corpus of the 950,000 most frequent words in the Bengali language [10]. However, we only considered the consonants for this comparison, as the study did not report frequencies for the other characters and symbols. Results revealed that the frequencies correlate reasonably well ($R^2 = 0.75$). Figure 4 illustrates the frequencies of the consonants in our phrase set.

Discussion

Results revealed that the average phrase and word lengths are substantially lower in our phrase set than the commonly used CIIL corpus [3] that collected phrases from novels, textbooks, newspapers, poems, scientific journals, etc. This indirectly validates our claim that the CIIL and similar corpora are not representative of the actual messages composed on actual mobile devices. This further highlights the necessity of a phrase set for evaluating Bengali mobile text entry techniques.

Limitations

Results revealed that about 85% users used English characters to compose Bengali text messages – that is, they wrote Bengali texts using the Roman alphabets[1]. This is mainly due to the unavailability of an effective Bengali text entry technique for mobile devices. Thus, it is possible that they would have written the texts differently if such a technique was available to them. Besides, we recruited participants from public places, such as parks, tea stalls, etc., thus may have excluded a subset of the users who do not visit these places.

Conclusion

We presented a phrase set of Bengali texts composed by real users on actual mobile devices. The set contains phrases with varying lengths, symbols, and numbers. We also reported results of an analysis of the set. Considering that currently there is no realistic Bengali phrase set available, we hope that this set will assist Bengali text entry researchers to increase the external validity of their work.

Acknowledgements

We thank the students of North South University, Dhaka, Bangladesh for helping us with the project.

References

1. Akshar Bharati, K. Prakash Rao, Rajeev Sangal, and S. M. Bendre. 2000. Basic statistical analysis of corpus and cross comparison among corpora. Technical Report of Indian Institute of Information Technology.
2. Samit Bhattacharya and Subrata Laha. 2013. Bengali text input interface design for mobile devices. *Universal Access in the Information Society* 12, 4: 441-451.

3. The Central Institute of Indian Language. 2005. Bengali Sample Pages. Retrieved January 7, 2016 from http://www.ciilcorpora.net/bensam.htm
4. Girish Dalvi, Shashank Ahire, Nagraj Emmadi, Manjiri Joshi, Nirav Malsettar, Debasis Samanta, Devendra Jalihal, and Anirudha Joshi. 2015. A Protocol to Evaluate Virtual Keyboards for Indian Languages. In *Proceedings of the 7th International Conference on HCI, IndiaHCI 2015* (IndiaHCI'15). ACM, New York, NY, USA, 27-38.
5. Anirudha Joshi, Girish Dalvi, Manjiri Joshi, Prasad Rashinkar, and Aniket Sarangdhar. 2011. Design and evaluation of Devanagari virtual keyboards for touch screen mobile phones. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services* (MobileHCI '11). ACM, New York, NY, USA, 323-332.
6. Ray S. Kumar. The Bengali Language and Translation. Translation Articles. Kwintessential. Retrieved January 7, 2016 from http://www.kwintessential.co.uk/translation/articles/bengali-language.html
7. I. Scott MacKenzie and R. William Soukoreff. 2003. Phrase sets for evaluating text entry techniques. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '03). ACM, New York, NY, USA, 754-755.
8. Md Enamul Hoque Prince, Gahangir Hossain, Ali Akbar Dewan, and Pijush Debnath. 2009. An audible Bangla text-entry method in mobile phones with intelligent keypad. In *Proceedings of the 12th International Conference on Computer and Information Technology* (ICCIT '09). IEEE, Washington, DC, USA, 279-284.
9. Sayan Sarcar, Ahmed Sabbir Arif, and Ali Mazalek. 2015. Metrics for Bengali text entry research. In CHI 2015 Workshop on Text Entry on the Edge (April 18, 2015). Seoul, South Korea, 4 pages. Retrieved January 7, 2016 from http://www.asarif.com/pub/workshop/CHI15Workshop_BengaliTextEntryMetrics.pdf
10. Rezwana Sharmeen, Md Abul Kalam Azad, Shabbir Ahmad, and S. M. Kamruzzaman. 2005. Smart Bengali cell phone keypad layout. In *Proceedings of the 8th International Conference on Computer and Information Technology* (ICCIT '05). Islamic University of Technology, Gazipur, Dhaka, Bangladesh.
11. The Unicode Consortium. 2015. Unicode 8.0 Bengali Character Code Charts. Retrieved January 7, 2016 from http://unicode.org/charts/PDF/U0980.pdf
12. Keith Vertanen and Per Ola Kristensson. 2011. A versatile dataset for text entry evaluations based on genuine mobile emails. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services* (MobileHCI '11). ACM, New York, NY, USA, 295-298.
13. World Bank Group. 2015 A measured approach to ending poverty and boosting shared prosperity. The World Bank, USA.