Research Report on Bangla Lexicon

Kamrul Hayder
BRAC University, Dhaka, Bangladesh.
kamrulhayder@hotmail.com

Abstract

We report on the compilation of a comprehensive Bangla word list lexicon. The current list contains 80,969 words from the Standard Chalita Bhasha (SCB) vocabulary. The word list is currently being used by the BRAC University Bangla Spelling Checker application.

1. Introduction

The lack of a freely available electronic Bangla lexicon prompted us to compile a list of Bangla words from the Standard Chalita Bhasa (SCB) vocabulary. The words were chosen from a set of commonly used Bangla dictionaries, starting from the standard one produced by Bangla Academy [1-12]. In producing the lexicon for the spelling checker, we were careful in omitting archaic or "dictionary" words, which cause the spelling checkers to flag incorrect words as correct ones if those happen match words from the archaic usage. The word list has been verified by an independent team of native speakers with reasonable level of linguistic knowledge, with a second level of verification is underway during the process of tagging the lexicon. The lexicon is released under the Creative Commons License [13], with full redistribution rights for any purpose.

2. Methods

The words in the lexicon were compiled from the various commonly used dictionaries [1-12], with the Bangla Academy dictionary providing the majority share of the words in the list. The lexicon is currently neither tagged nor annotated with any other information; however, the PAN Localization project is currently tagging the lexicon and annotating it with pronunciation using narrow IPA transcription.

3. Results

The Bangla Lexicon (BLEX) currently contains approximately 80 thousand head words, compared to approximately 70 thousand head words found in the Bangla Academy dictionary. The words in BLEX that are not in Bangla Academy dictionary are those that are commonly used in the literature, but missing from Bangla Academy's list of words. Many of these words happen to be recent imports from foreign languages, and some are technical and scientific terms that have been imported into Bangla.

4. Conclusion

While the 80 thousand word lexicon is certainly the most comprehensive of all freely available electronic lexica available for Bangla, it needs to be tagged with POS tags and annotated with pronunciation and other information before it can be used for advanced applications such as text-to-speech (TTS), automatic speech recognition (ASR) and machine translation (MT). The tagging process has recently begun and we plan to release a fully tagged lexicon by the year end.

5. References

[1] J. Choudhury, *Bangla Banan Abhidhan*, Bangla Academy, Dhaka.

[2] A. Ishaaque, Samakalin Bangla Bhashar Abhidhan, Bangla Academy, Dhaka.

[3] A.K. Mustafa, *Nazrul Shabdakosh*, Bangla Academy.

[4] S. Biswas, *Samsad Bangla Abhidhan*, Sahitya Samsad.

[5] A.T. Deb, *Sabdabodh Abhidhan*, Deb Sahitya Kutir Ptv. Ltd.

[6] R. Bosu, *Chalantika*, M.C. Sarkar and Sons Ptv. Ltd.

[7] H. Bandyopadhyaya, *Bangiya Sabdakosh*, Sahitya Akademi.

[8] G. Das, *Bangala Bhasar Abhidhan*, Sahitya Samsad.

[9] J. Bidyanidhi, B. *Sabdakosh*, Bhurjapattra.

[10] K.A. Odud, S.A. Ghosh, *Baboharic Shabdakosh*, Presidency Library.

[11] M. Datta and A. Mukharji, *Sabdasanchyita*, New Central Book Agency Pvt. Ltd.

[12] S. Mittra, *Saral Bangala Avidhan*, New Bengal Press Pvt. Ltd.

[13] Creative Commons License, http://creativecommons.org/licenses/.