

Research Report on Bangla Tagged Lexicon

Kamrul Hayder, Md. Zahurul Islam, and Mumit Khan

Center for Research on Bangla Language Processing, BRAC University

kamrulhayder@hotmail.com, zaharul@bracu.ac.bd, mumit@bracuuniversity.ac.bd

Abstract

This report describes the design and implementation of a Bangla tagged lexicon. The resulting lexicon contains 144,770 entries, out of which 58,145 are verbs. The tags used in the lexicon are reproduced here from a previous report on the Bangla tagset.

1. Introduction

This report describes the design and implementation of a Bangla tagged lexicon. The original lexicon developed at CRBLP was a simple word list primarily for use in a spelling checker. As we started adding context handling to our spelling checker, we needed more information such as the POS tag, along with other lexical features, and needed to develop a tagged lexicon. There are a few major design choices:

1. Tagset: the choice of the primary tagset must be made reasonably end application-neutral. We chose to use the tagset developed at CRBLP in Q3/2006, which is heavily

influenced the Penn Treebank tagset (Santorini, 1990; Marcus, Santorini, and Marcinkiewicz, 1993; Marcus et. al, 1994).

2. Depth vs. breadth: This is of course the classical tradeoff of whether to incrementally tag a large word list, adding tag depth at each step, or to create a deeply tagged lexicon word at a time. We chose breadth over depth, as the end user applications being developed at CRBLP requires a larger word list first, and then tag depth.

We are also experimenting with automated part of speech tagging to extend the current lexicon using a newspaper corpus, but that work is still in its infancy. Eventually, we will use the currently tagged lexicon to bootstrap a new one by tagging a corpus, and then extracting the lexicon from it.

The following sections describe the tagset that we are using (reproduced here from a previous report submitted as part of our 3Q06 deliverables), followed by a discussion on how we selected the words in the list, and process that we followed to tag the words.

2. Bangla tagset

#	Level 1	Level 2	Tag	Examples
1	Noun	<i>Proper</i>	NNP	মতিউর, অক্টোবর
2		<i>Common</i>	NNC	মানুষ, পানি
3		<i>Verbal</i>	NNV	করা, করানো, পরা, পরানো
4		<i>Temporal</i>	NNT	গতকাল, আগামীকাল, আজ, শনিবার, রবিবার
5	Pronoun	<i>First Person</i>	PR1	আমি, আমরা
6		<i>Second Person</i>	PR2	তুমি, তোমরা, ওগো
7		<i>Third Person</i>	PR3	সে, যে, তারা, যারা
8		<i>Non Person</i>	PRN	স্বয়ং, নিজে, সবাই, কে, কেউ

#	Level 1	Level 2	Tag	Examples
9		<i>Creditable</i>	PRC	আপনি, তিনি, যিনি, আপনারা, তাঁরা, যাঁরা
10		<i>Insignificant</i>	PRD	তুই, তোরা, ওরে
11		<i>Possessive</i>	PR\$	আমার, তোমার, তার, আমাদের, তোমাদের, ওর, আপনাদের, কার
12		<i>TO Pronoun</i>	PRTO	আমাকে, তোমাকে, তাকে, তারে, আপনাকে, কাকে
13	Adjective	<i>Simple</i>	AJ	সুন্দর, লাল, গরম, শ্রেষ্ঠ, শ্রেষ্ঠতর, শ্রেষ্ঠতম
14	Verb	<i>First Person</i>	VB1	করি, করছি, করেছি, করলাম, করছিলাম, করেছিলাম, করব, করাই
15		<i>Second Person</i>	VB2	কর, করছ, করেছ, করছিলে, করেছিলে, করাও
16		<i>Third Person</i>	VB3	করে, করছে, করেছে, করল, করছিল, করেছিল, করায়, করুক, হোক
17		<i>Non Person</i>	VBN	করলে, করালে
18		<i>Creditable</i>	VBC	করেন, করছেন, করেছেন, করলেন, করছিলেন, করেছিলেন, করবেন
19		<i>Insignificant</i>	VBD	কর, করছিস, করেছিস, করা
20		<i>Infinite</i>	VBIF	করে, করতে, করাতে
21	Adverb	<i>Adverb</i>	AV	আস্তে, দ্রুত, ধীরে, কেন, কিভাবে
22	Conjunction	<i>Co-ordinating</i>	CC	এবং, ও, কিংবা, অথবা, নতুবা
23		<i>Subordinating</i>	CS	তাই, যে
24	Inflectors	<i>AT</i>	ICAT	এ, য, তে
25		<i>BY</i>	ICBY	এ, তে (ইট-পাটকৈলে/NNC+ICBY অনেক মানুষ হতাহত হয়েছে)
26		<i>Plural</i>	ICS	রা, এরা, গুলি, গণ

#	Level 1	Level 2	Tag	Examples
27		<i>TO</i>	ICTO	কে, রে, এরে, দিগকে, দিগেরে
28		<i>Possessive</i>	IC\$	এর, দের
29		<i>Determinative</i>	ICDT	টা, টি
30		<i>Adverbial</i>	ICAV	ও
31		<i>Definitive</i>	ICDF	ই
32	Postposition	<i>Common</i>	PP	দ্বারা, কর্তৃক, হতে, হইতে, থেকে
33		<i>Possessive</i>	PP\$	জন্য, চেয়ে, চাইতে
34	Interjection	<i>Interjection</i>	UH	বাহ্!, ওহ্! হায়!
35	Indeclinables	<i>Simple</i>	ID	আর, অবশ্য, তবে, হয়তো, সুতরাং, সর্বাপেক্ষা, সবচেয়ে
36		<i>Infinite</i>	IDIF	যদি
37	Particle	<i>Particle</i>	PT	কি, না, নাকি, যেন, বটে
38	Onomatopes	<i>Onomatopes</i>	ON	টনটন, কনকন, খাঁ খাঁ
39	Cardinal	<i>Cardinal</i>	CD	এক, দুই, ১, ২
40	Determiner	<i>Singular</i>	DT	এটি, ওটি, কি
41		<i>Plural</i>	DTS	সব, ওসব, সকল, তাবৎ, কোন, যেকোন, এই, ঐ, কিছু
42		<i>Predeterminer</i>	DTP	এই/DTP সকল/DTI, যেকোন/DTP কিছু/DTI বৈজ্ঞানিক বা অংকশাস্ত্রীয় যেকোন
43	Symbol	<i>Symbol</i>	SYM	চিহ্ন
44	Taka	<i>Taka</i>	/=	৳ (টাকার চিহ্ন)
45	Sentence Final Punctuation	<i>Sentence Punctuation</i>	<i>Final</i>	, ? , !
46	Comma	<i>Comma</i>	,	,
47	Colon, Semi-colon	<i>Colon, Semi-colon</i>	:	∴, ∵
48	Bracket	<i>Left Bracket</i>	(([
49		<i>Right Bracket</i>))]
50	Quotation	<i>Opening Single Quote</i>	'	`
51		<i>Closing Single Quote</i>	'	'

#	Level 1	Level 2	Tag	Examples
52		Opening Double Quote	"	"
53		Closing Double Quote	"	"

A sample text tagged with the tagset is shown below.

সব/AJ জল্পনা-কল্পনার/NNC+I C\$ অবসান/NNC
ঘটিয়ে/VBIF তত্ত্বাবধায়ক/AJ সরকার/NNC ও/CC
নির্বাচন/NNC কমিশন/NNC সংস্কারের
/NNC+I C\$ বিষয়ে/NNC+I CAT প্রধান/AJ দুই
/CD দল/NNC বিএনপি/NNP ও/CC আওয়ামী/NNP
লীগের/NNP+I C\$ মহাসচিব-সাধারণ/AJ সম্পাদক/NNC
পর্যায়ে/NNC+I CAT সংলাপ/NNC হচ্ছে/VB3
আজকালের/NNC+I C\$ মধ্যেই/PP\$+I CDF ||
আওয়ামী/NNP লীগের/NNP সাধারণ/AJ সম্পাদক/NNC
আব্দুল/NNP জলিল/NNP গতকাল/NTT শনিবার/NNP
দুপুরে/NNC+I CAT বিএনপির/NNP+I C\$
মহাসচিব/NNC ও/CC স্থানীয়/AJ সরকারমন্ত্রী/NNC
আব্দুল/NNP মান্নান/NNP উইয়াকে /NNP+I CTO
টেলিফোন/NNC করে /VBIF আজকালের
/NTT+I C\$ মধ্যেই/PP\$+I CDF সংলাপে/NNC+I
CAT বসতে/VBIF আগ্রহের/NNC+I C\$ কথা /NNC
জানান/VBC || মান্নান /NNP উইয়াও /NNP+I CAV
জবাবে /NNC+I CAT জানিয়েছেন/VBC ,/
সংলাপে/NNC+I CAT বসতে/VBIF প্রস্তুত/AJ
তিনিও/PRC+I CAV || দুজনে /NNC+I CAT
সুবিধাজনক/AJ সময়ে/NNC+I CAT
বৈঠকের/NNC+I C\$ দিনক্ষণ/NNC ও/CC স্থান /NNC
ঠিক/AV করবেন/VBC || উভয় /DTI নেতা/NNC পৃথক
/AJ সংবাদ /NNC ব্রিফিংয়ে/NNC+I CAT বিষয়টি
/NNC+I CDT জানান/VBC ||

সংলাপে/NNC+I CAT বসতে/VBIF দুই/CD দলের/NN
C+I C\$ প্রস্তুতি/NNC চূড়ান্ত/AJ হওয়ায়/VB3
দেশের/NNC+I C\$ বিভিন্ন/AJ স্তরের/NNC+I C\$ মানুষের
/NNC+I C\$ মধ্যে/PP\$ স্তির/NNC+I C\$
ভাব/NNC দেখা/VBIF যাচ্ছে/VB3 ||
বিভিন্ন/AJ রাজনৈতিক/AJ দল/NNC
বিষয়টিকে/NNC+I CDT+I CTO ইতিবাচক/AJ
বলে/VBIF স্বাগত/NNC জানিয়েছে/VB3 ||

অবশ্য/ID এ/DTI অবস্থার/NNC+I C\$
মধ্যেই/PP\$+I CDF আজ/NTT রবিবার/NNP ১/CD
অক্টোবর/NNP বিএনপি/NNP ও/CC তার/PR\$
শরিকের/NNC+I CS পালন/NNC করছে/VB3 ‘/ ভোট/
AJ বিপ্লব/AJ ‘/ দিবস/NNC ||
দিনটিকে/NNC+I CDT+I CTO বিরোধী/AJ
দল/NNC আওয়ামী/NNP লীগ/ NNP পালন/NNC
করছে/VB3 ‘/ কালো/AJ দিবস/NNC ‘/ হিসেবে/PP ||
চলমান/AJ রাজনৈতিক/AJ সংকট/NNC
নিরসনে/NNC+I CAT দুই/CD দলের/NNC+I C\$
মধ্যে/PP\$ সমঝোতা/NNC চেষ্টার/NNC+I C\$

মধ্যে/PP\$ অনেকে/PRC আজকের/NTT+I C\$
এই/DTI ঘটনাকে/NNC+I CTO তাৎপর্যপূর্ণ/NNC
হিসেবেই/PP+I CDF দেখছেন/VBC ||

3. Word-list selection

The first decision that we had to wrestle with was how to select the proper word list to start with. The lack of balanced corpus made it difficult to use the frequency metric to select the set, so we used a combination of “native-speaker instinct”, head-words from official dictionary, and frequency distribution from a newspaper corpus to select 86,625 non-verb entries. The verbs were generated from 900 verb roots, creating a total of 58,145 inflected forms.

4. Conclusion

This report presents a Bangla tagged lexicon with 144,770 entries, out of which 58,145 are verbs. Each verb is tagged to level 3, while 2nd level tagging is underway for the other categories. The lexicon is now used to provide the contextual information in addition to the word entries to the CRBLP spelling checker, as well as being used to automatically tagging a corpus.

5. References

- [1] B. Santorini, “Part-of-speech tagging guidelines for the Penn Treebank Project”, *Technical report MS-CIS-90--47*, Department of Computer and Information Science, University of Pennsylvania, 1990.
- [2] M.P. Marcus, M.A. Marcinkiewicz and B. Santorini, “Building a large annotated corpus of English: the penn Treebank”, *Comput. Linguist.* 19, 2, June 1993, pp. 313-330.
- [3] M. Marcus, G. Kim, M.A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger, “The Penn Treebank: annotating predicate argument structure”, In *Proceedings of the Workshop on Human Language Technology*, Plainsboro, NJ, Human Language Technology Conference, Association for Computational Linguistics, Morristown, NJ, March 08 - 11, 1994, pp. 114-119.