

Research Report on Bangla Tagset

Altaf Mahmud and Mumit Khan

Center for Research on Bangla Language Processing, BRAC University, Dhaka, Bangladesh.
altaf.mahmud@gmail.com, mumit@bracuniversity.ac.bd

Abstract

This report describes the design of a POS tagset for Bangla, based on the Penn Treebank design. The resulting tagset contains 53 morpho-syntactic tags.

1. Introduction

This report describes the design of a tagset for Bangla, based on the Penn Treebank design. The design is heavily influenced by the work on Penn Treebank tagset, and follows the same methodology [1, 2, 3].

2. Bangla Tagset

Table 1: 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
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, আপনারদের, কার
12		<i>TO Pronoun</i>	PRTO	আমাকে, তোমাকে, তাকে, তারে, আপনাকে, কাকে

#	Level 1	Level 2	Tag	Examples
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>Infinitive</i>	VBIF	করে, করতে, করাতে
21	Adverb	<i>Adverb</i>	AV	আশ্বে, 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, য়, 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
27		<i>TO</i>	ICTO	কে, রে, এরে, দিগকে, দিগেরে
28		<i>Possessive</i>	ICS\$	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

#	Level 1	Level 2	Tag	Examples
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
43	Symbol	<i>Symbol</i>	SYM	বৈজ্ঞানিক বা অংকশাস্ত্রীয় যেকোন চিহ্ন
44	Taka	<i>Taka</i>	/=	৳ (টাকার চিহ্ন)
45	Sentence Punctuation	Final <i>Sentence Final Punctuation</i>		, ?, !
46	Comma	<i>Comma</i>	,	,
47	Colon, colon	Semi- <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>	'	'
52		<i>Opening Double Quote</i>	"	"
53		<i>Closing Double Quote</i>	"	"

3. Results

A sample text tagged with the tagset is shown below.

সব/AJ জল্পনা-কল্পনার/NNC+IC\$অবসান/NNC
ঘটিয়ে/VBIF তত্ত্বাবধায়ক/AJ সরকার/NNC ও/CC নির্বাচন/NNC
কমিশন/NNC সংস্কারের /NNC+IC\$বিষয়ে/NNC+ICAT
প্রধান/AJ দুই /CD দল/NNC বিএনপি/NNP ও/CC
আওয়ামী/NNP লীগের/NNP+IC\$ মহাসচিব-সাধারণ/AJ
সম্পাদক/NNC পর্যায়ে/NNC+ICAT সংলাপ/NNC হচ্ছে/VB3
আজকালের/NNC+IC\$ মধ্যেই/PP\$+ICDF/| আওয়ামী/NNP
লীগের/NNP সাধারণ/AJ সম্পাদক/NNC আব্দুল/NNP
জলিল/NNP গতকাল/NT শনিবার/NNP দুপুরে/NNC+ICAT
বিএনপির/NNP+IC\$ মহাসচিব/NNC ও/CC স্থানীয়/AJ
সরকারমন্ত্রী/NNC আব্দুল/NNP মান্নান/NNP ভূঁইয়াকে

/NNP+ICTO টেলিফোন/NNC করে /VBIFআজকালের
/NNT+IC\$মধ্যেই/PP\$+ICDFসংলাপে/NNC+ICAT
বসতে/VBIF আগ্রহের/NNC+IC\$ কথা /NNC জানান/VBC /|
মান্নান /NNP ভূঁইয়াও /NNP+ICAV জবাবে
/NNC+ICATজানিয়েছেন/VBC,/ সংলাপে/NNC+ICAT
বসতে/VBIF প্রস্তুত/AJ তিনিও/PRC+ICAV/| দুজনে
/NNC+ICAT সুবিধাজনক/AJ সময়ে/NNC+ICAT
বৈঠকের/NNC+IC\$ দিনস্বপ্ন/NNC ও/CC স্থান /NNC
ঠিক/AV করবেন/VBC/| উভয় /DTIনেতা/NNC পৃথক /AJ
সংবাদ /NNC ব্রিফিংয়ে/NNC+ICATবিষয়টি /NNC+ICDT
জানান/VBC/|
সংলাপে/NNC+ICAT বসতে/VBIFদুই/CDদলের/NNC+I
C\$প্রস্তুতি/NNCচূড়ান্ত/AJহওয়ায়/VB3
দেশের/NNC+IC\$বিভিন্ন/AJস্বরের/NNC+IC\$মানুষের/NNC+I
C\$মধ্যে/PP\$স্বস্তির/NNC+IC\$
ভাব/NNCদেখা/VBIFযাচ্ছে/VB3/|

বিভিন্ন/AJরাজনৈতিক/AJদল/NNC
 বিষয়টিকে/NNC+ICDT+ICTO ইতিবাচক/AJ
 বলে/VBIFস্বাগত/NNCজানিয়েছে/VB3||
 অবশ্য/ID এ/DTI অবস্থার/NNC+IC\$ মধ্যেই/PP\$+ICDF
 আজ/NTT রবিবার/NNP ১/CD অক্টোবর/NNP বিএনপি/NNP
 ও/CC তার/PR\$ শরিকেরা/NNC+ICS পালন/NNC করছে/VB3
 ‘/’ ভোট/ AJবিপ্লব/AJ ‘/’ দিবস/NNC ||
 দিনটিকে/NNC+ICDT+ICTO বিরোধী/AJ দল/NNC
 আওয়ামী/NNP লীগ/ NNPপালন/NNC করছে/VB3 ‘/’
 কালো/AJ দিবস/NNC ‘/’ হিসেবে/PP || চলমান/AJ রাজনৈতিক/AJ
 সংকট/NNC নিরসনে/NNC+ICAT দুই/CD দলের/NNC+IC\$
 মধ্যে/PP\$ সমঝোতা/NNC চেস্তার/NNC+IC\$ মধ্যে/PP\$
 অনেক/PRC আজকের/NTT+IC\$ এই/DTI
 ঘটনাকে/NNC+ICTO তাঁৎপর্যপূর্ণ/NNC হিসেবেই/PP+ICDF
 দেখছেন/VBC ||

4. Conclusion

This report presents a Bangla part-of-speech (POS) tagset that is based on the Penn Treebank tagset design. The tagset contains 53 2-level tags. A sample text tagged with this tagset is shown.

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*, Human Language Technology Conference, Association for Computational Linguistics, Morristown, Plainsboro, NJ, March 08 - 11, 1994, pp. 114-119.