



Research Report on Bangla Corpus Tagset for Tagging

Country Component	Bangladesh	Report no.	QPR - 01
Phase no.	1.1	Report Ref no.	PANL10n/Admn/QPR/001

Prepared By: Altaf Mahmud_____
(Name)

Designation: Research Associate

Project Leader: Mumit Khan_____
(Name)

Signature: _____

Date: October 8, 2007_____

TABLE OF CONTENTS

Title	- 3 -
Abstract	- 3 -
Introduction	- 3 -
Bangla Tagset	- 3 -
Results	- 5 -
Conclusion	- 6 -
References	- 6 -

Title

Bangla Tagset

Altaf Mahmud and Mumit Khan
Center for Research on Bangla Language Processing
BRAC University

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.

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 (Santorini, 1990; Marcus, Santorini, and Marcinkiewicz, 1993; Marcus et. al, 1994).

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	Pronoun	<i>Temporal</i>	NNT	গতকাল, আগামীকাল, আজ, শনিবার, রবিবার
5		<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, তঁারা, যাঁরা
10		<i>Insignificant</i>	PRD	তুই, তোরা, ওরে

#	Level 1	Level 2	Tag	Examples
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 করি, করছি, করেছি, করলাম, করছিলাম, করেছিলাম, করব, করাই
14	Verb	<i>First Person</i>	VB1	কর, করছ, করেছ, করছিলে, করেছিলে, করাও
15		<i>Second Person</i>	VB2	করে, করছে, করেছে, করল, করছিল, করেছিল, করায়, করুক, হোক
16		<i>Third Person</i>	VB3	করলে, করালে
17		<i>Non Person</i>	VBN	করেন, করছেন, করেছেন, করলেন, করছিলেন, করেছিলেন, করবেন
18		<i>Creditable</i>	VBC	কর, করছিস, করেছিস, করা
19		<i>Insignificant</i>	VBD	করে, করতে, করতে
20		<i>Infinite</i>	VBIF	আস্তে, দ্রুত, ধীরে, কেন, কিভাবে
21	Adverb	<i>Adverb</i>	AV	এবং, ও, কিংবা, অথবা, নতুবা
22	Conjunction	<i>Co-ordinating</i>	CC	তাই, যে
23		<i>Subordinating</i>	CS	এ, য়, তে
24	Inflectors	<i>AT</i>	ICAT	এ, তে (ইট-পাটকেলে/NNC+ICBY অনেক মানুষ হতাহত হয়েছে)
25		<i>BY</i>	ICBY	রা, এরা, গুলি, গণ
26		<i>Plural</i>	ICS	কে, রে, এরে, দিগকে, দিগেরে
27		<i>TO</i>	ICTO	এর, দের
28		<i>Possessive</i>	IC\$	টা, টি
29		<i>Determinative</i>	ICDT	ও
30		<i>Adverbial</i>	ICAV	

#	Level 1	Level 2	Tag	Examples
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, তাবঁ , 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 Sentence Final	<i>Taka</i>	/=	৳ (টাকার চিহ্ন)
45	Punctuation	<i>Sentence Final Punctuation</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>	'	'
52		<i>Opening Double Quote</i>	"	"
53		<i>Closing Double Quote</i>	"	"

Results

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 নির্যাসন/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 বিএনপি/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/NNC+I C\$ প্রস্তুতি/NNC চূড়ান্ত/AJ হওয়ায়/VB3 দেশের/NNC+I C\$ বিভিন্ন/AJ স্তরের/NNC+I C\$ মানুষের/NNC+I C\$ মধ্যে/PP\$ স্বস্তির/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/I D এ/DTI অবস্থার/NNC+I C\$ মধ্যেই/PP\$+I CDF আজ/NTT রবিবার/NNP ১/CD অক্টোবর/NNP বিএনপি/NNP ও/CC তার/PR\$ শরিকেরা/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 তাপর্যপূর্ণ/NNC হিসেবেই/PP+I CDF দেখছেন/VBC ||

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.

References

Santorini, B. 1990. 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.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. 1993. Building a large annotated corpus of English: the penn treebank. *Comput. Linguist.* 19, 2 (Jun. 1993), 313-330.

Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., and Schasberger, B. 1994. The Penn Treebank: annotating predicate argument structure. In *Proceedings of the Workshop on Human Language Technology (Plainsboro, NJ, March 08 - 11, 1994)*. Human Language Technology Conference. Association for Computational Linguistics, Morristown, NJ, 114-119.