Annotation Guidelines for Bangla

Contents

Table 1.1: Category and their Types

CATEGORY	Attributes
NOUN	Common
	Proper
	Verbal
	Spatio-temporal
VERB	Main
	Auxiliary
PRONOUN	Pronominal
	Reflexive
	Reciprocal
	Relative
	Wh
NOMINAL MODIFIER	Adjective
	Quantifier
DEMONSTRATIVE	Absolute
	Relative
	Wh
ADVERB	Manner
	Location
PARTICIPLE	Verbal (Adverbial)
	Conditional
PARTICLE	Coordinating
	Subordinating
	Classifier
	Interjection
	Others
Punctuation	
RESIDUAL	Foreign word
	Symbol
	Others

Table 1.2: Attributes and their Values

ATTRIBUTE\ SYMBOL	Value\symbol		
NUMBER\NUM	Singular\sg	Plural\pl	
PERSON\PER	First\1	Second\2	Third\3
TENSE\TNS	Present\prs	Past\pst	Future\fut
CASE MARKER\CSM	Accusative\ acc	Genitive\gen	Locative\ loc
ASPECT\ASP	Simple\smp	Progressive\ prg	Perfect\pft
MOOD\MOOD	Declarative\ dcl	Imperative\imp	Habitual\hab
FINITENESS\FIN	Finite\fin	Non-finite\nfn	Infinite\ifn
DISTRIBUTIVE\DSTB	Yes\y	No\n	
DEFINITENESS	Yes\y	No\n	
EMPHATIC\EMPH	Yes\y	No\n	
NEGATIVE\NEG	Yes\y	No\n	
HONORIFICITY\HON	Yes\y	No\n	
NUMERAL\NML	Ordinal\ord	Cardinal\crd	Non-numeral\ nnm
REALIS	Realis \rls	Irrealis\ils	

Annotation Guidelines for Bangla

Introduction

This is a guideline for annotating Bangla (Bengali) text with Parts-of-Speech (POS) tags in accordance with the IL-POSTS (**I**ndian **L**anguage **POS T**ag **S**et); the hierarchical POS tagset framework developed by Microsoft Research India Private Limited in collaboration with linguists and NLP experts. ILPOSTS is the framework which can be used in POS tagging of all Indian languages. We will present the guidelines for Bangla after a short description of ILPOSTS which is the motivation behind this guideline.

The goal of IL-POSTS[1] is to provide a common POS-tagset framework for Indian Languages that offers flexibility, cross-linguistic compatibility and reusability across Indian languages. The framework allows language specific tagsets to be derived from it. An important consideration for its hierarchical structure and decomposable tags is that it should allow users to specify the morphosyntactic information applicable at the desired granularity according to the specific language and task. Thus, IL-POSTS offers broad guidelines for users to define their own tagset for a particular language and/ or a specific application.

Structure of the Framework (ILPOSTS)

The framework has three layers in the hierarchy – the top level having the *universal lexical categories* followed by *types* (of these categories) in the middle layer and *features (attributes)* carrying finer details placed in the last layer. The description of the same is as follows -

Categories:

Categories are the primary grammatical classes to which the words belong. 'Grammatical' means grossly the parts of speech through which each individual word is recognized, e.g., noun, verb, adjective etc. The Category level tags are determined on the basis of the categorization features of the word. It decides on the Parts of Speech the word that it belongs to.

Types:

Types are the subclasses or finer specification of the categories, which are determined on the basis of either form or function. E.g., Common Noun, Proper Noun etc. are the subcategory of the category 'Noun'. Types fine tune the classification of the categories. Each Type group words which are similar in terms of their characteristics, i.e., whether they form a class on the basis of their distribution, references etc. Thus all nemed entities are Proper Nouns under the category of Noun.

---

[1] For detailed understanding of ILPOSTS, we highly recommend reading **Framework for a Common Parts-of-Speech Tagset for Indic Languages.** Draft. Microsoft Research Lab India. Bangalore.

Attributes:

Attributes are the set of basic morpho-syntactic features of a type, like, person, number, gender etc. Each Type has fixed and exhaustive set of Attributes that accumulates the possible morpho-syntactic features to be attached to each word. Some attributes are mandatory and some are optional. Mandatory attributes are those which contributes to the basic meaning of the word in its grammatical specification, e.g., Tense in Verb is mandatory, but Distance in pronouns (e.g., in Bangla, এখানে (proximal), ওখানে (distal), সেখানে (sequal) are optional) since the information is derived from the semantics of the word. These attributes have a fixed set. In this framework we have a set of eighteen attributes distributed over the nominal and verbal morphology.

Description of Bangla Tagset

ILPOSTS is designed in such a way so that it allows language specific tagsets to be derived from it. Any guideline therefore pertains to tagging and the description would be under this framework. It is always possible to derive a slightly different tagset for Bangla (say, a tagset ignoring some attributes or merging some types). Under such cases, the tagging guideline would a slightly different according to the tagset. But, by and large any tagset for Bangla from ILPOSTS should use very similar tagging guidelines.

The objective of the guideline is to provide a complete and disambiguating guideline for tagging Bangla sentences. This tagset consists of **categories, types**, and their **attributes** which are the three different levels of the hierarchy. **Categories** are the top-level part-of-speech classes like noun, adjective etc. and they are obligatory. **Types** are the main sub-classes of categories and may be included depending on whether or not those types exist in a particular Indian language. **Attributes** are morpho-syntactic features of Types. All attributes are optional, and are included in the language specific tagset depending on the morpho-syntactic realizations of the same.

The purpose of this guideline is to facilitate tagging Bangla sentences. Sentences are tagged using an Annotation Tool developed in Microsoft Research India. This Annotation Tool is a GUI for assignment of appropriate tags in to each word in a sentence. The tool can be adopted in any language given the language specific tagset in the proper format. This tool can be adopted for either flat or a hierarchical tagset. Annotation tool guideline[2] is complementary to the tagging guideline. The default values that are mentioned in this guideline are in accordance

---

[2] Dandapat, S. MSRI Part-of-Speech Annotation Interface. Draft. Microsoft Research India Private Limited.

with the values set in the Annotation Tool. Reading the Annotation Tool Guideline is recommended before starting with the tagging process or using the tagset.

Some important points:

- While annotating the words we concentrate on the **forms** for the attributes, i.e., if the attributes are present in the word morpho-syntactically, we mark them accordingly. We focus on **function** while annotating the types. Those special cases would be dealt in Appendix C. However, this is only a general guiding principle and may vary depending on the context.

- The tags are saved or represented just after the words separated by back slash (\). Abbreviations of categories and their types are written without any symbol, but the values of the attributes are written being separated by DOTs (.).

- consider the following example for proper understanding of the hierchical tagset and the way it works -

  o পত্রিকা   অফিসে   কাজ করি ।

The above sentence has four words and a punctuation mark. We find out the Parts of Speech of each word in the sentence. Each word would belong to a Category, thus -

  o পত্রিকা\N অফিসে   \N কাজ   \N করি   \V   ।\PU

Categories are, as specified earlier, the highest level of the hierarchical tagset. After the categories are identified, they are set for finer classification. We identify the Tyoes in the next level, thus –

  o পত্রিকা\NC অফিসে   \NC কাজ   \NC করি   \VM ।\PU

The first three words belong to the type of common nouns (NC). This is the first set of tags that would be represented in the tool. After this level is identified, the next level, i.e., the attributes would be identified on the basis of morpho syntactic markers available on the word.

  o পত্রিকা\NC.0.0.n.n   অফিসে\NC.0.loc.n.n   কাজ\NC.0.0.n.n
    করি\VM.1.prs.sim.dcl.fin.n.n.n ।\PU

This is the second level of tags that are represented in the Annotation tool. This level is depends on the previous level, i.e., the type level and each type would have its unique feature specification. This features or attributes (such as number, person,

8

emphatic etc.) would choose their values according to the morphological   marker present in the particular word.

Thus, the above string has four words and they are tagged in the given convention –

- o The first word {পত্রিকা} is a common noun (NC) and its attribute set is – {number, person, definiteness, emphatic} whose values are - {not applicable (0), not applicable (0), no (n), no (n)} respectively; hence the tag is –

     পত্রিকা\NC.0.0.n.n

     But, consider the sentence - ছেলেটা পত্রিকা অফিসে কাজ করে ।
     ছেলেটা would be annotated as –   ছেলেটা\NC.**sg**.0.**y.**n

- o The last word {করি} is a main verb, containing attribute-value set-

     {person (1), tense (prs), aspect (sim), mood (dcl), finiteness (fin), emphatic (n), negative (n), honorificity (n)} ;

     Hence the tag is –   করি\VM.1.prs.sim.dcl.fin.n.n.n

- The thumb rule for deciding the Category-Type tag of a word is to find out the function; while forms should be taken into consideration while deciding the attributes, which are morphologically present.

     e.g., সুস্থমূল\NC.0.0.n.n **প্রথমে\ALC.0.n.n** সরু\JJ.n.n সরু\JJ.n.n ঝুরি\NC.0.0.n.n হিসবে\PP.0.n বাতাসে\NC.0.loc.n.n ঝোলে\VM.3.prs.sim.dcl.fin.n.n.n ।\PU

     'প্রথমে' is an adverb of location according to the function, though it is a quantifier according to the form.

Common value[3] for all the attributes –

- o *Not-applicable (0)*
     - - When any value is not applicable to the category or the relevant morpho-syntactic feature is not available.
     - - When the category is a binary valued category, i.e., the values of a particular Attribute are 'yes' and 'no' as in the case of Emphatic,

---

[3] These common values are specifically for users of the annotation tool, read the annotation Tool guideline for better understanding of the structure of the tagset and the tool

Negative, Definiteness etc.; annotate/select the value as '**yes**' only when the morphological attribute is present. Otherwise, annotate as '**no**'.

- o *Undecided or doubtful (x)*
  - when the annotator is not sure about a possible attribute, instead of marking on the basis of doubt, tag it as '**x**', e.g., inherently ambiguous cases would be given priority of the contexts; but if they still remain disambiguated, annotate the attributes to be '**x**'.

Description of Tags

- **Categories and their Types: See table 1.1**
- **Attributes and their values: see Table 1.2**

1. **NOUN (N)**

The types and attributes of a NOUN are –

TYPE	ATTRIBUTES
Common (NC)	**Number, Case Marker, Definiteness, Emphatic**
Proper noun (NP)	**Number, Case Marker, Definiteness, Emphatic**
Verbal (NV)	**Case Marker, Definiteness, Emphatic**
Spatio-temporal (NST)	**Case Marker, Definiteness, Emphatic**

1.1. COMMON NOUN (NC)

Common nouns in this tagset are the words that belong to the groups of common noun (person, place or a thing), abstract noun (emotions, ideas etc), collective noun (group of things, animals, or persons), countable and non-countable nouns, nouns in a complex predicate (N+V predicate e.g., অবলুপ্তি ঘটে, দ্বিধা করিনি), nouns in a compound noun (N + N, e.g., জ্ঞান বিজ্ঞান চর্চার, একটি সাইবার কমিউনিটি) etc. Common nouns are different than the proper nouns and the verbal nouns; e.g.,

সভাস্থল\NC.0.0.n.n ত্যাগ\NC.0.0.n.n করার পর গাড়ীতে\NC.0.loc.n.n আরোহনের\NC.0.gen.n.n পর মুহূর্তে\NC.0.loc.n.n আত্মঘাতী হামলায়\NC.0.loc.n.n নিহত হন বেনজির ভুট্টো।

Attributes: Number, Case Marker, Definiteness, Emphatic

Number:

Number in the nouns are marked by the morpho-syntactic markers or plural markers -রা/ -এরা, গুলো/গুলি, -দের etc.. It is annotated as (**0**) if number attribute is not required (as in compound nouns, nouns of the compound verbs, proper nouns etc.) e.g.,

অবলুপ্তি\NC.0.0.n.n ঘটে, দ্বিধা\NC.0.0.n.n করিনি

জ্ঞান\NC.0.0.n.n বিজ্ঞান\NC.0.0.n.n চর্চার, একটি সাইবার\NC.0.0.n.n

কমিউনিটি\NC.0.0.n.n, অনুবাদ\NC.0.0.n.n -\PU স্বীকরণ\NC.0.0.n.n কর্মে,

খিদে\NC.0.gen.n.n পায়, প্রচন্ড\JJ.n.n বিশৃঙ্খলা\NC.0.0.n.n

- Annotate the value of the number as **plural (pl)** if any plural marker is present in the word, e.g., সংগঠকদের\NC.**pl**.gen.n.n, ভোটের\NC.0.gen.n.n বাবুরা\NC.**pl**.0.n.n, বস্তুগুলিও\NC.**pl**.0.n.y

- Select the value of number attribute as '**sg**' (singular) when it is marked with definiteness marker, e.g., বিষয়টিকেও\NC.sg.**acc**.y.y; মেয়েটা\NC.sg.0.y.n etc.

Case marker:

Case markers are also marked if it is present morphologically (-কে, -র/ -এর, -তে/-এতে/-য় etc.); otherwise the default case marker is '0', e.g.,

পুরুষকে\NC.0.**acc**.n.n; মাকে\NC.0.**acc**.n.n; সমাধানের\NC.0.gen.n.n বিষয়টিকেও\NC.sg.**acc**.y.y; রবীন্দ্রনাথের\NP.0.**gen**.n.n; নির্বাচনের\NC.0.**gen**.n.n সময়\NC.0.0.n.n; মানুষের\NC.0.**gen**.n.n জীবন\NC.0.0.n.n; বিপরীত মেরুতে\NC.0.**loc**.n.n; কপাল\NC.0.0.n.n হাঁটুতে\NC.0.**loc**.n.n ঠেকাও; জানলা\NC.0.0.n.n দরজায়\NC.0.**loc**.n.n নতুন রং; পুরানো চেহারায়\NC.0.**loc**.n.n; মনে\NC.0.**loc**.n.n হত

Definiteness:

Definiteness is the feature which makes the noun a definite entity. It is marked with definiteness markers like –টা, -টি, -টে, -জন, -খানা, etc.

Countable nouns would be valued as '*sg*' in number attribute when the definiteness marker –টা, -টি, -টে, are there.

e.g., এই কথাটি\NC.sg.0.y.n; দুধটুকু\NC.0.0.y.n, বইখানা\NC.sg.0.y.n; তালাটা\NC.sg.0.y.n;

Emphatic:

An emphatic is marked when the emphatic marker (-ও, -ই etc.) is present with the noun. An emphasis marker puts stress on the reference of the noun.

E.g., দারোগাবাবুও\NC.0.0.n.**y**; প্রাণকৃষ্ণবাবুকেই\NP.0.acc.n.**y**

Some doubtful cases:

- When nominal modifiers (intensifiers) take nominal inflections, mark them as NCs.
    - Nominal modifiers as nouns-
      বেশির\NC.0.gen.n.n ভাগ ক্ষেত্রে; **ভালোটা\NC.sg.0.y.n** খাব; **একটার\ NC.0.gen.n.n** সময়ে এসো
      But, if বেশির ভাগ is written together, as in – বেশিরভাগ ক্ষেত্রে, it would be annotated as nominal modifier.

    - Quantifiers as nouns – When numeral quantifiers (ordinal and cardinals) take nominal inflections (definiteness marker) are not followed by any noun, mark them as NC, e.g.,
      ৯টির\NC.0.gen.y.n উপরেই, অন্য **দুজন\NC.0.0.y.n** ছিলেন, যে **দুটিকে\NC.0.acc.y.n,** কোন **কিছুরই\NC.0.gen.n.y** নির্দিষ্ট বাঁধাধরা সীমা নেই

- Unit nouns would be annotated as particles though some are annotated as NCs, specifically when they are inflected, e.g., ইত্যাদির\NC.pl.gen.n.n, প্রভৃতিকে\NC.pl.acc.n.n, but consider the next example - বিজ্ঞান\NC.0.0.n.n **বিষয়ক**\CX.n [4] রচনায়, অরিয়নিস\NP.0.0.n.n **নামক**\CX.n তারার চারদিকে

- Borrowed words would be annotated according to their function. This is a highly context depended criterion, therefore the annotation would infer from the context -

  প্রথম\JQ.n.n.ord **বিলিয়নিয়ার**\NC.0.0.n.n, **আমেরিকান**\JJ.n.n ডলারে\NC.0.loc.n.n, গ্র্যান্ড\JJ.n.n **মাস্টার**\NC.0.0.n.n আখ্যায়

---

[4] For detailed discussion on particles, see page 26-27

1.2. PROPER NOUNS (NP)

When the word denotes a specific name of a person, place, shop, institution, date, day, month, species, etc., or whatever is considered to be a name would be marked as proper noun. If the word is of some other category but is used as a proper noun in a context; should be marked as proper noun, e.g.,

কোকিল\NP.0.0.n.n, ময়ূর\NP.0.0.n.n, পেঁচা\NP.0.0.n.n, পাপিয়া\NP.0.0.n.n, কুমের\NP.0.0.n.n,

স্যার\NP.0.0.n.n উইনস্টন\NP.0.0.n.n চার্চিলকে\NP.0.acc.n.n ভোটে হারিয়েছিল ।

তবে ১৯৫৮ সালে ডক্টর\NP.0.0.n.n দীনেশচন্দ্র\NP.0.0.n.n সেন\NP.0.0.n.n এই গানগুলো সম্পাদনা করে কলকাতা\NP.0.0.n.n বিশ্ববিদ্যালয়\NP.0.0.n.n হতে ...

নিউট্রন\NP.0.0.n.n ছাড়া কেন্দ্রীণ\NP.0.0.n.n হয় না,

1938 সালের সেপ্টেম্বর\NP.0.0.n.n ও অক্টোবরেNP.0.loc.n.n নগুচিকে\NP.0.acc.n.n লেখা রবীন্দ্রনাথের\NP.0.gen.n.n পিপলস\NP.0.0.n.n ডেইলিতে\NP.0.loc.n.n তিন কলম জুড়ে...

ভারতীয়\NP.0.0.n.n অলিম্পিক\NP.0.0.n.n য্যাসোসিয়েশন\NP.0.gen.n.n,

জেনেটিক্স\NP.0.0.n.n, বাংলা\NP.0.0.n.n, ইংরেজী\NP.0.0.n.n,

ফাইকাস\NP.0.0.n.n (ডুমুর\NP.0.0.n.n জাতীয়) গোত্রের ইউরোস্টিগ্মা\NP.0.0.n.n উপগোত্রের সদস্য ।

আর্থার\NP.0.0.n.n সি\NP.0.0.n.n ক্লার্ক\NP.0.0.n.n,

রামকৃষ্ণ\NP.0.0.n.n মিশন\NP.0.0.n.n,

তাকে ইয়ুথ\NP.0.0.n.n অফেনডিং\NP.0.0.n.n টীমে\NP.0.loc.n.n পাঠানো...

- <u>বর্তমানের\NP.0.gen.n.n</u>   পুলিশ   প্রতিবেদনে   বলা   হয়েছে ...

The underlined word is a name of a newspaper, so, it should be tagged as an NP, and not JJ or NC.

Attributes: Number, Case Marker, Definiteness, Emphatic

Number:

Number attributes for proper noun would be '**0**' (zero) in case of name of person, place etc., e.g.,

1938 সালের সেপ্টেম্বর\NP.0.0.n.n ও অক্টোবরে\NP.0.loc.n.n নওচিকে\NP.0.acc.n.n লেখা রবীন্দ্রনাথের\NP.0.gen.n.n পিপলস\NP.0.0.n.n ডেইলিতে\NP.0.loc.n.n তিন কলম জুড়ে...

Case Marker:

Case markers are marked only if it is morphologically present, e.g.,

চন্দননগরকে\NP.0.acc.n.n, ধৃতরাষ্ট্রকে\NP.0.acc.n.n, বাংলাদেশের\NP.0.gen.n.n রাজনীতির, পদ্মায়\NP.0.loc.n.n রোজ মাছ ধরতে যাই, বাংলায়\NP.0.loc.n.n লেখো

Definiteness and ***Emphatic:*** Mark if they are morphologically present, e.g.,

রামটা\NP.0.0.y.n আজ আর এলো না

হুমায়ুন\NP.0.0.n.n আজাদেরই\NP.0.gen.n.y বই\NC.0.0.n.n

- ▪ Words like- ডক্টর, মি., শ্রী, etc. are tagged as NP. E.g.,

ডক্টর\NP.0.0.n.n বিধান\NP.0.0.n.n চন্দ্র\NP.0.0.n.n রায়\NP.0.0.n.n, মি.\NP.0.0.n.n রায়\NP.0.0.n.n, শ্রী\NP.0.0.n.n রায়\NP.0.0.n.n

1.3. VERBAL NOUNS (NV)

A verbal noun is a derived from of a verb which allows nominal inflections. Verbal nouns are gerunds and gerundival constructions in Bangla.

Attributes: Number, Case Marker, Definiteness, and ***Emphatic.***

The rule for all attributes works in the same way as Common Noun and Proper Noun; e.g.,

ঢাকা\NC.0.0.n.n দেওয়া**NV.**0.n.n ভাত\NC.0.0.n.n; বই পড়া**NV.**0.n.n;

দেখা**NV.**0.n.n যায়; বর্ণনা করার**NV.**gen.n.n ক্ষমতা; ছবি তোলারই**NV.**gen.n.y অপচেষ্টা

চাপাই**NV.**0.n.y থাক; গার্ড অভ অনার পাওয়াটা**NV.**0.y.n বেশ সম্মানজনক

- Some words would be ambiguous between NV and NC, as in –
কথাটা চাপাই**NV**.0.n.y থাক; ঢাকা\NC.0.0.n.n দেওয়া**NV.**0.n.n ভাত\NC.0.0.n.n
These specific cases would be dealt according to the context.

- Borrowed words that are used as gerunds originally would not retain their tags from the source language, and would be annotated according to the rules and specifications of the target language, e.g.,
টাইম\NC.0.0.n.n ব্লগিং\NC.0.n.n, সকালে উঠে জগিং\NC.0.n.n করা\NV.0.n.n ভালো

1.4. SPATIO-TEMPORAL NOUNS (NST)

These are the nouns that denote space, time, direction etc. e.g., আগে, নীচে, ওপর. These words will be marked as NSTs irrespective of their occurrence as a pre/post position or spatio-temporal noun. Spatio-temporal nouns are those which are easily interchangeable with post positions.

Attributes: Case Marker, Definiteness, Emphatic

They are marked only if they are morphologically present.

- When the spatio-temporal nouns are in nominal distribution, i.e., when they are functioning as nominals and not as post positions (e.g., উপরটাকে, নীচেটার সামনের দিকটা) case markers would be mandatory. Case markers are not semantically relevant when they are used as post positions, but for the sake of symmetry in our usage, we mark case marker in both instances.

- In some of the spatio-temporal nouns locative case marker would be inherently present in the nouns (সামনে, পিছনে). The case markers are not annotated in these cases since spatial features inherently have locative case markers attached to them. Such nouns which have non-case marked forms available, annotate the case marker in those cases. e.g.,
ভিতির ওপর\NST.0.n.n, ওপরে\NST.loc.n.n দাড় করিয়ে দেয়; বারান্দায় দাঁড়ালেই সামনে\NST.0.n.n, যে বিশাল অন্ধকার..., সমাবেশ শেষে\NST.loc.n.n সভাস্থল ত্যাগ করার পর\PP.0.n; দুই মাস আগেও\NST.loc.n.y একবার বেনজীর হত্যার ব্যর্থ চেষ্টা...; আলেকজান্ডার পশ্চিমে\NST.loc.n.n অভিযান...

- **Some doubtful cases:**

পুলিশের ওপর\NST.0.n.n, নির্বাচনের আগে\NST.loc.n.n, কাজের মাঝে\NST.loc.n.n, কাজটি করার আগেই\NST.loc.n.y, আমাদেরকে পিছনের\NST.gen.n.n দিকে\NST.loc.n.n টানছে, ছাত্র পড়ানো শুরু\NST.0.0.n.n করলো, ২ দিন পরেই\NST.loc.n.y ঈদ, বারান্দায় দাঁড়ালেই সামনে\NST.0.n.n যে বিশাল অন্ধকার...

2. VERBS (V)

The types and attributes of verbs are –

TYPE	ATTRIBUTES
Main verb (VM)	*Person, Tense, Aspect, Mood, Finiteness, Emphatic, Negative, Honorificity.*
Auxiliary verb (VA)	*Person, Tense, Aspect, Mood, Finiteness, Emphatic, Negative, Honorificity.*

2.1. MAIN VERBS (VM)

A Verb in a sentence denotes the eventual information expressed by the sentence. This information can be carried out by one or many verbs, but there should be only one main verb in a simple declarative sentence. This main verb can be accompanied by one or many helping verbs (auxiliary verbs or modal verbs) for proper manifestation of the eventual information. A main verb in its finite form is usually inflected in agreement with the subject (object[5] in the ergative cases in Hindi). If it is followed by one or several auxiliary verbs, usually the last verb in the verb group picks up the finiteness agreement. If the verb expresses infiniteness, no morphological attributes, other than the (in)finiteness information, is attached to the verb.

Attributes are same for both *Main verb (VM)* and *Auxiliary verbs (VA).* They are –

Person, Tense, Aspect, Mood, Finiteness, Emphatic, Negative, Honorificity

Person:

Among the nominal inflections (Person, Number, and gender) verbs in Bangla are inflected only for Person. Since Bangla verbs show agreement with the subject, the verb agrees with the person information of the subject. Bangla lacks gender and number agreement on the verb.

---

[5] Object agreement is found in ergative case agreement, e.g., raam ne **roTi** khaay**ii**

Default agreement for PERSON in passive sentences is marked as third person {3}, e.g.,

ঘোষণা করা হয়\VM.3.prs.sim.dcl.fin.n.n.n

Tense:

Tense is the basic temporal information on the verb which expresses the temporal location of the event. If the event occurs at the speech time, the information is expressed as the Present tense. If the event is before the speech time, it is expressed as Past tense and if the event is after the speech time, or the event is yet to take place, the information is expressed as Future tense. Thus the temporal information of a sentence is manifested by three tenses – Present, Past and Future. The present tense is morphologically not marked, Past tense is marked with the morpheme /-ল/*(-la)* and the Future tense is marked with /-ব/ (**-ba**) in the verbal inflections. Default value for tense in the tool is – Present.

e.g.,

ঘুমে দু চোখ জড়িয়ে\VM.0.0.0.0.nfn.n.n.n আসে\VA.3.prs.sim.dcl.fin.n.n.n

মাও এই বাউন্ডুলেপনা মেনে\VM.0.0.0.0.nfn.n.n.n নিয়েছেন\VA.3.prs.pft.dcl.fin.n.n.n

রাত করে ফিরি\VM.1.prs.sim.dcl.fin.n.n.n

আগেরদিন রাতেই বুঝে\VM.0.0.0.0.nfn.n.n.n গিয়েছিলাম\VA.1.pst.pft.dcl.fin.n.n.n

বারান্দায় দাঁড়াব\VM.1.fut.sim.dcl.fin.n.n.n

এই এন্ট্রিটি লিখব\VM.1.fut.sim.dcl.fin.n.n.n,

যখন শুতে গেলাম\VM.1.pst.sim.dcl.fin.n.n.n

Aspect:

Aspect is the internal temporal constituency of an event expressed by the verb. This indicates whether the event is completed or ongoing etc. Morphologically it is expressed by three aspects – simple, perfective and progressive. Simple aspect is expressed by present simple tenses, which also expresses the habitual present information. The perfective aspect denotes completion of an event. The progressive aspect denotes that the event is still ongoing.

Default aspect in the tool is **simple**, mark progressive and perfective if they are morphologically present. Usually the perfective is marked by V-root+–এছ+agreement marker, e.g.,

কর+এছ+এ = করেছে\VM.3.prs.pft.dcl.fin.n.n.n

কর+ছ+এ = করছে\VM.3.prs.prg.dcl.fin.n.n.n

e.g.,

বইটা রামবাবু নিয়েছেন\VM.3.prs.pft.dcl.fin.n.n.n

আমি দোকানে গিয়েছিলাম\VM.1.pst.pft.dcl.fin.n.n.n

আমি দিল্লী গিয়েছি\VM.1.prs.pft.dcl.fin.n.n.n

এক্ষুনি ওদের বাড়িতে খবর দিচ্ছি\VM.1.prs.prg.dcl.fin.n.n.n

গল্প লেখা হচ্ছে\VM.3.prs.prg.dcl.fin.n.n.n

আলো ঠিকরে\VM.0.0.0.0.nfn.n.n.n পড়ছে\VM.3.prs.prg.dcl.fin.n.n.n

এই রাস্তা দিয়ে অনেক ট্রেন যায় \VM.3.prs.sim.dcl.fin.n.n.n

তারপরে সে কোনোরকমে হোটেলে পৌঁছায়\VM.3.prs.sim.dcl.fin.n.n.n

তারপরে তিনি লম্বা এক বক্তৃতা শুরু করেন\VM.3.prs.sim.dcl.fin.n.n.y

Mood:

Mood or modality is the semantic information inflected on the verb which essentially expresses speaker's attitude towards the listener or the event. Modality information is conveyed by the morphological markers as well the position of the modal auxiliaries in the sentence. Modal information contains the information about the basic purpose of the speaker for making an utterance (illocutionary point of utterance), strength of the illocutionary points, presuppositional knowledge of utterance and the situation etc. in order to derive various types of moods such as declarative, imperative, conditional, subjunctive, interrogative, presumptive, optative etc.

Default mood in the tool is declarative. Mark Imperative, Habitual moods and others if morphologically present in the data; e.g.,

নেতা বক্তৃতা শুরু করেন\VM.3.prs.sim.dcl.fin.n.n.y

বইটা ধরুন\VM.3.prs.sim.imp.fin.n.n.

রোজ বিকেলে বাবা আমাদের পড়াতেন\VM.3.pst.sim.hab.fin.n.n.y

Finiteness:

Finiteness conveys the information whether the event the verb denotes is finite or not. Any verb can take finite or non finite forms according to the event information. A finite verb is marked for nominal and verbal inflections, whether non-finite verbs are not marked.

Finiteness is expressed as –

- **Finite(fin):** verb forms that inflect for agreement, e.g., নিয়েছেন\VM.3.prs.pft.dcl.fin.n.n.n, দিচ্ছি\VM.1.prs.prg.dcl.fin.n.n.n

- **Infinite (ifn)** –তে, as in - বাড়তে\VM.0.0.0.0.ifn.n.n.n থাকলো; বলতে\VM.0.0.0.0.ifn.n.n.n পারি

- **Non-finite(nfn)** –এ/-য়ে, as in – একটা দুর্ঘটনা ঘটিয়ে\VM.0.0.0.0.nfn.n.n.n ফেলেছে; মাটি ভেদ করে\VM.0.0.0.0.nfn.n.n.n বেরিয়ে\VM.0.0.0.0.nfn.n.n.n এলো

- In case of non-finite or infinite, all the values of Tense, Aspect, Mood, and Person would be '0'.

- In case of root verbs, as in বের করে ফেলেছে, the value for finiteness would be '0' as well.

Emphatic:

Emphatic particles (–ই, –ও) can be attached to any categories, annotate them if they are morphologically present. The default value is '**n**' for emphatic particles in the tool.

Negative:

 Negative in Bangla also carries particle-like properties and can float anywhere in the sentence. It can be attached to the verbs too. Usually the negative particles are written separately, but it serves to be an attribute of the verb if it is attached to it. Bangla stative auxiliaries - হওয়া, থাকা take special forms in negatives, e.g., নয়, নেই etc. Negative particles attached to the verb in perfect aspect is –নি, as in করেনি, করিনি etc. e.g.,

অবনী বাড়ি নেই\VM.3.prs.sim.dcl.fin.n.y.n

আমি রামায়ণ পড়িনি\VM.1.prs.pft.dcl.fin.n.y.n

19

আজ আর কোন কথা নয়\VM.3.prs.sim.dcl.fin.n.y.n

আমার আর তর সইছিলনা\VM.3.pst.prg.dcl.fin.n.y.n

- ▪ This can be written separately also, if this is the case, annotate them as particles.

- ▪ The default value of the negative is also set in '**n**', mark it '**y**' if it is morphologically present.

Honorificity:

Honorificity is an agreement feature on the verb which agrees with the subject. Honorificity in Bangla is marked in three forms, Formal, informal and intimate. The same feature is applicable on the pronouns too. Honorificity is marked only on the second person and third person, and never on the first person. However, honorificity in this framework is not finely classified. It has binary value in the framework: **Yes** and **No**. Only the formal is considered to have morphological marker in this framework. Therefore, **আপনি** আমাদের বাড়ি আসবেন   would be marked as honorific, but **তুমি** আমাদের বাড়ি আসবে/ **তুই** আমাদের বাড়ি আসবি would be marked as non-honorific. Default value for HONORIFICITY is '**n**', annotate only if it is morphologically present in the verb, e.g.,

ছাত্রদের বিনামূল্যে পড়াতেন\VM.3.pst.sim.hab.fin.n.n.y

ছেলেটা রোজ সকালে কাগজ দিয়ে যেত\VM.3.pst.sim.hab.fin.n.n.n

পড়তে\VM.0.0.0.0.ifn.n.n.n পারতেন\VA.3.pst.sim.hab.fin.n.n.y না

কতটা কষ্ট করেছেন\VM.3.prs.pft.dcl.fin.n.n.y

কে কি বলবেন\VM.3.fut.sim.dcl.fin.n.n.y

2.2. AUXILIARY VERBS (VAUX)

Auxiliary verbs are the helping verbs which along with the main verb completes the temporal or eventual information expressed by the predicate. There should be ideally one main verb in a simple sentence but a verb group forming a complex predicate (V +V) is also found. Auxiliary verbs take the agreement relations in complex predicates. We do not distinguish between auxiliary verbs and light verbs or vector verbs in the framework. We provide a list of light verbs (adopted and extended from Paul, S. (2004)) if the second verb in the verb complex belongs to

the list of verbs, annotate it as auxiliary verb. Usually the first verb will be tagged as a VM and others will be tagged as VA. If the second verb does not belong to the Auxiliary list given below, consider the meaning to decide whether it is an auxiliary verb or a main verb. Auxiliary verbs comprise a closed class of verbs universally which comes after a main verb. Usually the last auxiliary verb carries the verbal inflection.

List of auxiliary verbs in Bangla[6]

আসা, যাওয়া, আনা, পারা, ওঠা, ফেলা, তোলা, বসা, মরা, রাখা, দেওয়া, নেওয়া, বেড়ানো, চলা, পাঠানো, দাঁড়াAdded to this- থাকা, চাওয়া, হওয়া, লাগা

- Other kind of complex predicates might be VM–তে + VM

  e.g., খেতে ভালোবাসে, খেলতে চায়, যেতে দেখলো, বলতে শুনলো

- A sequential verbal complex, which might be ambiguous, as evident from the given examples e.g.,করে ফেলেছে, বলে দাও, খেয়ে নাও

  o করে ফেলেছে –

    a. 'have done it completely'

    b. 'have done it and then have thrown it'

In the above case, when the meaning is aspectual (case a.), annotate the second verb as VA, otherwise VM.

a. করে\VM.0.0.0.0.nfn.n.n.n   ফেলেছে\ VA.3.prs.pft.dcl.fin.n.n.n

b. করে\VM.0.0.0.0.nfn.n.n.n   ফেলেছে\ VM.3.prs.pft.dcl.fin.n.n.n

Examples of verb types –

> যাত্রা\NC.0.0.n.n করবেন\VM.3.fut.sim.dcl.fin.n.n.y, বলা\NV.0.n.n হচ্ছে\VM.3.prs.prg.dcl.fin.n.n.n,
>
> মৃতের সংখ্যা বাড়তে\VM.0.0.0.0.ifn.n.n.n পারে\VA.3.prs.sim.dcl.fin.n.n.n বলে উল্লেখ করা\NV.0.n.n হয়েছে\VM.3.prs.pft.dcl.fin.n.n.n
>
> মৃত্যু ঘটিয়ে\VM.0.0.0.0.nfn.n.n.n থাকে\VA.3.prs.sim.dcl.fin.n.n.n যেতে\VM.0.0.0.0.ifn.n.n.n হল\VA.3.pst.sim.dcl.fin.n.n.n

---

[6] Paul, S. 2004. An HPSG Account of Bangla Compound Verbs with LKB Implementation. Ph.D. Dissertation. CALTS: University of Hyderabad.

শুতে\VM.0.0.0.0.ifn.n.n.n গেলাম\VA.1.pst.sim.dcl.fin.n.n.n
পড়তে\VM.0.0.0.0.ifn.n.n.n পারিনি\VA.1.prs.pft.dcl.fin.n.y.n
দিতে\VM.0.0.0.0.ifn.n.n.n চাচ্ছিলাম\VA.1.pst.prg.dcl.fin.n.n.n
বেরোতে\VM.0.0.0.0.ifn.n.n.n হবে\VM.3.fut.sim.dcl.fin.n.n.n
হতে\VM.0.0.0.0.ifn.n.n.n হবে\VA.3.fut.sim.dcl.fin.n.n.n
দেখতে\VM.0.0.0.0.ifn.n.n.n গিয়েছিলাম\VA.1.pst.pft.dcl.fin.n.n.n

- All the verbs in their default agreement or in passive sentence patterns will show third person agreement.

3. PRONOUNS (P)

A pronoun is a grammatical category which can substitute a noun or a noun phrase. Pronouns are referential and anaphoric in nature, i.e., they always refer to an antecedent either in the sentence or in the discourse.

Types and attributes for Pronouns are -

TYPES	ATTRIBUTES
Pronominal (PPR)	Number, Person, Case Marker, (Distributive), Definiteness, Emphatic, Honoroficity
Reflexive (PRF)	Number, Case marker, Definiteness, Emphatic
Reciprocal (PRC)	Case Marker, Emphatic
Relative (PRL)	Number, Case Marker, (Distributive), Definiteness, Emphatic, Honorificity
Wh-pronoun (PWH)	Number, Case Marker, (Distributive), Definiteness, Emphatic, Honorificity

3.1. PRONOMINALS (PPR)

Pronominals include all personal pronouns, demonstrative pronouns, inclusive pronouns and indefinite pronouns. In the case of indefinite pronouns, person attribute should be marked as [0] i.e., not applicable, e.g.,

- Personal pronouns:
  আমি, তুমি, আমার, সে, তাঁরা, তাদের, তোমাকে, তোর, আমাদেরকে, ইনি, উনি, এরা, এদের, এটা, এখানা, এগুলো, ওরা, ওগুলো, ওকে, ওদের etc.
- Demonstrative pronouns: এ, এই, ও, ওই etc.
- Inclusive Pronouns: সবাই, অনেকে, উভয় etc.

- Indefinite pronouns: কেউ, কারা, কারো, কিছু, কোনো etc.
- Others: অন্য, প্রত্যেক etc. which are used as quantifiers if followed by a noun.

Demonstrative pronouns can be used as demonstrative[7] also, as in – **ওই বইটা** দাও, **এ বইটার** মলাটটা ছিঁড়ে গেছে. The main difference in the distribution between a pronoun and a demonstrative is that a demonstrative is necessarily followed by a noun, pronoun or adjective, but a pronoun has the distribution of a noun, can take nominal inflections. A pronoun need not be followed by another noun or other parts of speech. As seen in the above examples that the demonstratives are followed by the nouns in the sentences.

Attributes: Number, Person, Case Marker, Distributive, Definiteness, Emphatic, Honoroficity

Number:

Number is morphologically as well as lexically marked in most of the cases in pronouns. The default value is singular 'sg'. Mark it as plural 'pl' if number is morphologically present in the word. In case of inclusive pronouns, the number attribute should be marked as 'pl', as they are inherently plural.

Person:

Person attribute would be marked in personal pronouns. All the demonstrative pronouns would be marked as '3'. Mark the person as '0' (not applicable) in inclusive, indefinite, and other pronouns.

Case Marker:

The default value of the case-marker is '0' when it is not marked or not applicable. Mark the Case-Markers only if the morphological attribute is present there.

Distributive:

This is an optional attributes, i.e., while annotating the corpora one has option to drop it. Distributive is the feature for the pronoun when it occurs twice consecutively in the sentence. This is the feature that refers to the iterativity of subjects. E.g., কেউ\PPR.sg.0.0.y.n.n.n কেউ\PPR.sg.3.0.y.n.n.n, কোনো\ PPR.0.0.0.y.n.n.n কোনো\ PPR.0.0.0.y.n.n.n etc. this feature would be marked on both the pronouns.

---

[7] See page 25-26 for detailed discussion on demonstratives.

Definiteness:

Definiteness is the feature which makes the pronoun a definite entity. It is marked with definiteness markers like –টা, -টি, -টে, জন, খানা as in এটা, সেখানা, কোনজন etc.

Emphatic and *Honorificity* are tagged as 'y' only if they are morphologically present.

e.g.,

আমার\PPR.sg.1.gen.n.n.n.n;   এর\PPR.sg.3.gen.n.n.n.n;   সেটা\PPR.sg.3.0.n.y.n.n;
এটুকু\PPR.sg.3.0.n.y.n.n   তা\PPR.sg.3.0.n.n.n.n;   তাঁর\PPR.sg.3.gen.n.n.n.y;
সেটাই\PPR.sg.3.0.n.y.y.n;   তাঁদের\PPR.pl.3.gen.n.n.n.y;   আপনারা\PPR.pl.3.0.n.n.n.y;
সেইটে\PPR.sg.3.0.n.y.n.n;   সেগুলোও\PPR.pl.3.0.n.n.y.n;   তাতে\PPR.sg.3.loc.n.n.n.n;
কেউ\PPR.sg.0.0.n.n.n.n;   কারোই\PPR.sg.0.gen.n.n.y.n;   কাউকে\PPR.sg.0.acc.n.n.n.n;
সবার\PPR.pl.3.gen.n.n.n.n জন্য;   সবাই\PPR.0.0.0.n.n.n.n;   অনেকেই\PPR.pl.0.0.n.n.y.n;
সেইটে\PPR.sg.3.0.n.y.n.n;   আমারটা\PPR.sg.1.gen.n.y.n.n,   অনেকটা\PPR.pl.0.0.n.y.n.n

- এমনি, তেমন, যেমন etc. are tagged as PPRs if not followed by any noun or some other categories, e.g. যেমন\DAB[8].sg.n কুকুর তেমন মুগুর; but,
- এমনিতেই\PPR.0.0.loc.n.n.y.n; যেমন\PPR.0.0.0.n.n.n.n –\PU খরগোশের চোখ... are annotated as pronouns.

3.2. REFLEXIVE (PRF)

A reflexive pronoun is an anaphoric pronoun that refers to the antecedent in the same sentence. it can co-refer to the subject or the object.

Attributes – Number, Case-Marker, Definiteness, Emphatic

In Bangla, the reflexive pronouns – নিজে is found in its various nominal inflection patterns, as in – নিজেরা, নিজেদের, নিজের, নিজেদের জন্য . আপন is also used in some archaic cases. All the attributes are applicable only if they are morphologically present, e.g.,

আমি নিজেই\PRF.pl.0.y ফোন করেছিলাম

আমরা নিজেরা\PRF.pl.0.n চলে গেছিলাম,

---

[8] See discussions on Demonstratives in page 24-25

ছেলেগুলো নিজেদের`\PRF.pl.gen.n` মধ্যে কি সব আলোচনা করলো

আমরা নিজেরা`\PRF.pl.0.n` নিজেদেরকে`\PRF.pl.acc.n` অনেক বোঝালাম `etc.`

3.3. RECIPROCAL (PRC)

A reciprocal pronoun is the one which refers to the mutual relation between the plural subjects of a sentence. Reciprocal pronouns in Bangla are - একে অপর; একে অন্য, পরস্পর etc. The second word of the set is inflected

Attributes: Case-Marker, Emphatic. e.g.,

রাম আর শ্যাম একে\PRC.0.n অপরের\PRC.gen.n প্রতি ভিষণ বিরক্ত।

রাম আর শ্যাম একে\PRC.0.n অপরকে\PRC.acc.n দোষারোপ করলো।

3.4. RELATIVE (PRL)

A relative pronoun is a <u>pronoun</u> that links two clauses into a single complex clause. It attaches the relative clause with the referent in the matrix clause. It is anaphoric in nature.

Attributes: Number, Case Marker, Distributive, Definiteness, Emphatic, Honorificity

Number and ***Case Marker*** would be annotated according to presence of the morphological markers.

Distributive: distributive is the feature which denotes iterativity of events or subjects. This is an optional attribute. Thus, in the following sentence -

যারা যারা নাটকে অংশ নিতে ইচ্ছুক হাত তোলো, যারা

would be annotated as -   যারা\PRL.pl.0.**y**.n.n.n   যারা\PRL.pl.0.**y**.n.n.n

e.g.,

যেটাকে\PRL.sg.acc.n.y.n.n; যা\PRL.sg.0.n.n.n.n এককালে ফ্রাসিদের; যার\PRL.sg.gen.n.n.n.n কোন প্রমাণ; যাঁরা\PRL.pl.0.n.n.n.y গণহত্যা বলেন; যেমন\PRL.0.0.n.n.n.n ধরুন; যাদের\PRL.pl.gen.n.n.n.n কেন্দ্রীণে; যিনি\PRL.sg.0.n.n.n.y etc.

3.5. WH PRONOUN (PWH)

Wh-pronouns are question pronouns that are used in interrogation or in asking questions.

Attributes: Number, Case-Marker, Distributive, Definiteness, Emphatic, Honorificity

Number and ***Case Marker*** would be marked only if it is morphosyntactically.

Rules for other attributes are the same as the pronominals, i.e., they are marked only if they are morphologically present.

Wh-pronouns in Bangla are - কে, কি, কেন, কোথায়, কবে, কারা, কাদের

e.g., কে\PWH.sg.0.n.n.n.n ?\PU বকেছে   তোমাকে? ওরা কোথায়\PWH.0.loc.n.n.n.n গেলো?

4. NOMINAL MODIFIER (J)

Nominal modifier is the category which usually modifies nominal constructions (nouns and pronouns) in the sentence. Adjectives and quantifiers have been put into the same group of nominal modifiers on the basis of their function. Intensifiers are also put into this category though intensifiers can modify verbal domains too. Intensifiers basically form the type of 'non-numeral quantifier'

The types and attributes of nominal modifiers are –

TYPES	ATTRIBUTES
Adjectives (JJ)	**Definiteness, Emphatic**
Quantifier (JQ)	**Definiteness Emphatic, Numeral**

4.1. ADJECTIVE (JJ)

An adjective modifies a noun; hence it is kept as a type of nominal modifier in the framework. Though adjectives are not always followed by nouns, it can be used as a predicate too. The first kind is called an attributive adjective and the second type is called a Predicative adjective. An adjective can function as a noun if not followed by a modified noun; in that case it is called an absolute adjective. However, these do not make any difference in the attribute set of the tagset. Nor do the comparative and superlative adjectives.

Attributes: Definiteness and Emphatic

The attributes are marked only if they are morphologically present in the words.

An adjective modifies a noun, e.g.,

গত\JJ.n.n আড়াই মাস, বিভিন্ন\JJ.n.n রাস্তা, সুবিধাজনক\JJ.n.n.n ঘরটি, রং থানিকটা ম্যাড়ম্যাড়ে\JJ.n.n, হালকা\JJ.n.n সবুজ\JJ.n.n, নাকি ফিরোজা\JJ.n.n কালার, দৃশ্যটা আরো সুন্দর\JJ.n.n, চমকপ্রদ\JJ.n.n সাফল্য, ঐতিহ্যের বাঁধা\JJ.n.n পথ থেকে, সেটাই স্বাভাবিক\JJ.n.n, আপাদমস্তক\JJ.n.n সিপিএম সমর্থক, বুশোচিত\JJ.n.n রসিকতা; মিড\JJ.n.n – টার্মের ছুটি, শান্তিপূর্ণ\JJ.n.n ভাবে গ্রামে ঢুকে; ক্ষিপ্ত\JJ.n.n মানুষ; কাঁদানে\JJ.n.n গ্যাস; ভারত পরিভ্রমণরত\JJ.n.n স্বামিজীকে ক্ষেত্রীর মহারাজ বিবেকানন্দ নাম অর্পণ করেন ।

If **emphatic** markers are present, mark the value to be (**y**) E.g., সহজেই\JJ.n.y, কমই\JJ.n.y, অবিবাহিতাও\JJ.n.

- It can be the first word of a complex predicate, e.g.,

প্রতি কালীপুজোয় ওথানে ধুমধাম সহকারে বলি **অনুষ্ঠিত\JJ.n.n** হয়,

আবার সেই কাজে **জড়িত\JJ.n.n** হতে হবে,

এইসব আমোদ আহ্লাদ **বন্ধ\JJ.n.n** করতে চাই,

এভাবে বললে যে কেউ একাজ করতে **বাধ্য\JJ.n.n** হবে;

৫০জন **আহত\JJ.n.n** হন;

চাটুটা **গরম\JJ.n.n** হয়ে উঠেছে;

কারখানাটা আর **চালু\JJ.n.n** হল না;

- When an adjective has a definiteness marker attached to it, annotate it as a common noun. e.g., ভালোটা থাবো, এবারের ঠান্ডাটা যেন বড় বেশী

- Attributve adjectives are different than predicative adjectives. Attributive adjectives are those which can modify nouns and usually appears infront of a noun, where predicative adjectives are those which can be used as a predicate i.e., it can take a subject. E.g.,এই বইটা ভালো**JJ.n.n**
কাল যে ছেলেটা এসেছিল সে বেশ লম্বা/ **JJ.n.n**

4.2. QUANTIFIERS (JQ)

A quantifier is a word which quantifies the noun, i.e., it expresses the definite or indefinite number or amount of the noun, e.g.,

দুটি\JQ.y.n.crd বিভিন্ন রাস্তা দিয়ে; গত\JJ.n.n আড়াই\JQ.n.n.crd মাস

আর কিছু\JQ.n.n.nnm না; রং খানিকটা\JQ.y.n.nnm ম্যাড়ম্যাড়ে; এত\JQ.n.n.nnm জীবন্ত\JJ.n.n নীলরং; কিঞ্চিৎ\JQ.n.n.nnm সন্দেহ ; অত\JQ.n.n.nnm ভালো\JJ.n.n আরো\JQ.n.n.nnm সুন্দর\JJ.n.n; একশ\JQ.n.n.crd ভাগ একমত, ২রা\JQ.n.n.crd অক্টোবর; খুবই\JQ.n.y.nnm বিখ্যাত, অত্যন্ত\JQ.n.n.nnm জাকজমকের সাথে, একটা\JQ.y.n.crd সময়ের পর, দশজন\JQ.n.n.crd উপদেষ্টা দিয়েই, প্রচুর\JQ.n.n.nnm লেখা, প্রতি\JQ.n.n.nnm সন্ধ্যায়, অপেক্ষাকৃত\JQ.n.y.nnm কম\JJ.n.n নির্যাতিত\JJ.n.n হয়ে

Attributes – Numeral, Definiteness, Emphatic

Numeral: Numerals have three values – ***cardinal* (crd)*, ordinal* (ord) *and non-numeral* (nnm)**.

- **Cardinal** numbers are numerals that are numbers, e.g., শত\JQ.n.n.crd শত\JQ.n.n.crd তারা, এক\JQ.n.n.crd নির্বাচনী\JJ.n.n সমাবেশ, তিন\JQ.n.n.crd ধরনের; একশ\JQ.n.n.crd ভাগ একমত, ২রা\JQ.n.n.crd অক্টোবর; একটা\JQ.y.n.crd সময়ের পর, দশজন\JQ.n.n.crd উপদেষ্টা দিয়েই,

- **Ordinals** are the numerals which are ordered numbers, e.g., সর্বপ্রথম\JQ.n.n.ord বহির্জাগতিক\JJ.n.n সৌরমণ্ডল, একবিংশ\JQ.n.n.ord শতাব্দীর, প্রথম\JQ.n.n.ord পরিচয়, ৮ম\JQ.n.n.ord শতাব্দী

- **Non-numerals** are usually the modifiers and intensifiers; the framework does not provide separate tag for nominal or verbal intensifiers. All the modifiers and intensifiers are categorized under the non-numeral quantifier.
  e.g., সমস্ত\JQ.n.n.nnm ভাষাতাত্ত্বিক\JJ.n.n চিন্তাধারা, বেশ\JQ.n.n.nnm কয়েক\JQ.n.n.crd শতাব্দী, সবচেয়ে\JQ.n.n.nnm বিখ্যাত\JJ.n.n, একাধিক\JQ.n.n.nnm পুরস্কারের মাধ্যমে, খুব\JQ.n.n.nnm বড়\JJ.n.n জায়গা; আর\JQ.n.n.nnm কিছু\JQ.n.n.nnm না; রং খানিকটা\JQ.y.n.nnm ম্যাড়ম্যাড়ে\JJ.n.n; এত\JQ.n.n.nnm জীবন্ত\JJ.n.n নীলরং; কিঞ্চিৎ\JQ.n.n.nnm সন্দেহ; অত\JQ.n.n.nnm ভালো\JJ.n.n; আরো\JQ.n.n.nnm সুন্দর\JJ.n.n; খুবই\JQ.n.y.nnm বিখ্যাত, অত্যন্ত\JQ.n.n.nnm জাকজমকের সাথে, প্রচুর\JQ.n.n.nnm লেখা,

প্রতি\JQ.n.n.nnm সন্ধ্যায়, অপেক্ষাকৃত\JQ.n.y.nnm কম\JJ.n.n নির্যাতিত\JJ.n.n হয়ে; বিশ্বাস যত\JQ.n.n.nnm বেশী\JJ.n.n সন্দেহ তত\JQ.n.n.nnm কম\JJ.n.n

- When a non-numeral quantifier is not followed by a noun, and the definiteness marker is attached to it, mark it as pronoun, e.g., অন্যটির\PPR.0.0.gen.n.y.n.n ভর, অন্যদের\PPR.pl.0.gen.n.n.n.n সঙ্গে\PP.0.n

  But, দশজন\JQ.n.n.crd উপদেষ্টা দিয়েই is a quantifier.

- In case of time specifications, as in Bangla, e.g., একটার\NC.sg.gen.y.n সময়ে 'one-definiteness marker-genitive time', mark the time word as noun; but not in e.g., রাত ১টা\RDX;

Definiteness and Emphatic are marked / annotated only if they are morphologically marked.

Some problematic cases:

আমি আর খাবো না, আর আমি খাবো না, আমি খাবো না আর

আর ভাত খেতে পারছি না, ভাত আর খেতে পারছি না, ভাত খেতে পারছি না আর

Here the modifier or the quantifier is not modifying the nominal or the nominal compound and behaves more like an adverb.

5. DEMONSTRATIVES (D)

Demonstratives form a class of words which deictically refers to the entity or the object. It refers to a spatial deixis for the location of the referent (i.e., object or the entity). Demonstratives are the words that indicate the spatial, temporal or discourse location of its referent i.e., the noun which follows the demonstrative has a definite reference in the discourse.

Types and attributes of demonstratives are-

TYPES	ATTRIBUTES
Absolute Demonstrative (DAB)	**Number, Emphatic**
Relative Demonstrative (DRL)	**Number, Emphatic**
Wh Demonstrative (DWH)	**Number, Emphatic**

The attributes for all types of demonstrative are – **Number** and **Emphatic** and both are marked if the respective markers are present in the word.

In Bangla, demonstratives have the same form of the pronouns, but they differ from the pronouns in their distribution. Demonstratives are always followed by a noun, adjective or another pronoun. Relative and wh-demonstratives are non-distinguishable from a relative and a wh-pronoun, except for that a relative or wh-demonstrative is followed by a noun, pronoun or an adjective.

e.g., কোন\DAB.sg.n কিছুরই, এই\DAB.sg.n ব্যাপারে, কোন\DAB.sg.n কোন\DAB.sg.n স্থানে, সেই\DAB.sg.n ভবিষ্যৎবাণীর, সে\DAB.sg.n ব্যাপারে, এমন\DAB.sg.n একটি\JQ.y.n.crd তাত্ত্বিক\JJ.n.n কাঠামোয়, যা\DRL.sg.n কিছু\JQ.n.n.nnm পরীক্ষণযোগ্য, যে\DRL.sg.n সমস্ত\JQ.n.n.nnm কোষে, যেকোনো\DRL.sg.n, ধরনের, যেসব\DRL.pl.n ছেলেরা, কি\DWH.sg.n ধরনের\NC.sg.gen.n.n নমুনা etc.

- If the demonstrative has emphatic marker attached to it, it would mark emphatic marker to be '**y**', e.g.,   যে\DRL.0.n মানুষটি, সেই/DAB.0.y মানুষটি

5.1. ABOSOLUTE DEMONSTRSTIVE (DAB)

Demonstratives have the same form of the pronouns, but they are different in their distribution than the pronouns. Demonstratives are always followed by a noun, adjective or another pronoun, where a pronoun is not followed by another noun, pronoun or adjective. Reduplication of pronouns, such as, *যে যে, যাকে যাকে* etc. are treated as pronouns, and not demonstratives.

5.2. RELATIVE DEMONSTRATIVE (DRL)
Relative demonstratives are non-distinguishable from relative pronouns, except for that a demonstrative is ALWAYS followed by a noun, pronoun or adjective.

5.3. WH DEMONSTRATIVES (DWH)
Wh-demonstratives are non-distinguishable from wh-pronouns, except for that a demonstrative is followed by a noun, pronoun or adjective. The change in the morphological form is found.

6. ADVERB (A)

An adverb belongs to a group of words that modifies a verb, adjective or a sentence.

Types and attributes of adverbs are -

TYPES	ATTRIBUTES
Adverbs of Manner (AMN)	**Case Marker, Definiteness, Emphatic**
Adverbs of Location (ALC)	**Case Marker, Definiteness, Emphatic**

6.1.   ADVERBS of MANNER (AMN)

These adverbs are the adverbs which modifies the way the action is described in the verb.

এইভাবে\AMN.0.n.n কেটে যাবার পর, **যথারীতি**\AMN.0.n.n বাড়ি ফিরে, **আর**\AMN.0.n.n তর সইছিল না, **শুধু**\AMN.0.n.n মনে হলো, **নিশ্চয়ই**\AMN.0.n.y অত\JQ.n.n.nnm ভালো\JJ.n.n হয়না, **প্রায়**\AMN.0.n.n **সঙ্গে**\AMN.0.n.n **সঙ্গেই**\AMN.0.n.y ঘুমিয়ে পড়েছি, **মূলত**\AMN.0.n.n রেনেসাঁস ধারণার পুনর্ঘোষণা, **আবার**\AMN.0.n.n জড়িত হতে হবে, **বারংবার**\AMN.0.n.n ঔপনিবেশিক অত্যাচারের, **হয়তো**\AMN.0.n.n **আস্তে**\AMN.0.n.n **আস্তে**\AMN.0.n.n পরিনত হচ্ছে, **বড়জোর**\AMN.0.n.n বাড়ির উঠোন পার, সংস্কার **বোধহয়**\AMN.0.n.n **এভাবেই**\AMN.0.n.y সম্ভব, **সরকারিভাবে**\AMN.0.n.n সংগঠকদের তরফ থেকে, **যেভাবে**\AMN.0.n.n মৃত্যুবরণ করে, **সেভাবে**\AMN.0.n.n শ্বাসরুদ্ধ হয়েই, **কেবলমাত্র**\AMN.0.n.n ব্যতিক্রম হলো, নেই **সাধারণতঃ**\AMN.0.n.n **সেখানে**\ALC.loc.n.n, বিক্ষোভ দেখানো হয় **এইভাবে**\AMN.0.n.n

- Some post positions like– ভাবে, করে (as in ভালো ভাবে, কিভাবে, যেভাবে, গড়গড় করে) and –এ (as in জোরে, প্রথমে) acts as adverbializer. In the first case, mark the word according to its original category, if it is not written together, e.g., ভালো\JJ.n.n ভাবে\PP.

- In the later case, mark them as adverbs, e.g., কি\PWH.sg.0.n.n.n.n ভাবে\PP.0.n গাড়ি চালাতে হয়?, কিভাবে\AMN.0.n.n গাড়ি চালাতে হয়?

6.2. ADVERBS of LOCATION (ALC)

All the words denoting time and place come under ALC; hence it includes words denoting words or phrases that spatially and temporally modifies the verb.

আজ\ALC.0.n.n পুলিশের বড় বাহিনী, শেষে\ALC.loc.n.n একটি বই বগলদাবা করে, যতোক্ষণ\ALC.0.n.n না ঘুমে দু চোখ জড়িয়ে আসে ততোক্ষণ\ALC.0.n.n বই পড়া, সেদিনও\ALC.0.n.y যথারীতি\AMN.0.n.n বাড়ি ফিরে, সহজেই যখন\ALC.0.n.n তখন\ALC.0.n.n আমার দরজায় কড়া নাড়তে পারেন, এখানে\ALC.loc.n.n সম্পূর্ণ অন্যরকম, এইবার\ALC.0.n.n অবশেষে\ALC.loc.n.n কিছুক্ষণ\ALC.0.n.n কম্পিউটার সায়েন্স নিয়ে, মুখোমুখি\ALC.0.n.n হয়ে, এখানকার\ALC.gen.n.n জগদ্ধাত্রী পূজা খুবই\JQ.n.y.nnm বিখ্যাত, সেদিন\ALC.0.n.n করেছি, আজও\ALC.0.n.y করছি, কালও\ALC.0.n.y করবো, যেখানেই\ALC.0.n.y, সেইখানেও\ALC.loc.n.y একটা সময়ের পর খন্ডযুদ্ধ শুরু, কখনও\ALC.0.n.y কখনও\ALC.0.n.y ক্যানক্রাই নামে উচ্চারিত...

Case markers are marked in the cases like-

এখানে\ALC.loc.n.n, এখানটায়\ALC.loc.y.n বসো, এখানেরটা\ALC.gen.y.n নাও

Definiteness and *emphatic* attributes are tagged as (y) if the markers are present in the word. E.g., এখানটায়\ALC.loc.y.n বসো, এখানেরটা\ALC.gen.y.n নাও, এখানেরটাই\ ALC.gen.y.y খাবো, এখানটাকে\ALC.acc.y.n

7. PARTICIPLE (L)

A participle is a derived form of a verb which is used as modifiers of verbs.

The types and attributes of participles in Bangla-

TYPES	ATTRIBUTES
VERBAL PARTICIPLE (LV)	**Emphatic**
CONDITIONAL PARTICIPLE (LC)	**Emphatic, Realis**

7.1. VERBAL PARTICIPLE

In Bangla, -তে form of the verb in reduplication is used in adverbial participles. The attribute is **Emphatic** and is marked only if it is present morphologically in the word.

e.g.,   বাতাসে উড়তে\LV.n উড়তে\LV.n চলে এল

                মেয়েটা নাচতে\LV.n   নাচতেই\LV.y   কথাটা বললো

7.2. CONDITIONAL PARTICIPLE

Conditional participles are used when there are two clauses in a sentence and one clause is conditioned by the other clause, so that the verb of the first clause will be in the participial form.

Attributes – Emphatic, Realis.

Emphatic attribute is marked if the emphatic marker is present in the word.

Realis:

Realis is the attribute which indicates that action is actually realized. A realis modality denotes a state which is known or actualized. In Bangla, grossly, -তে form is used for realis, -লে is used for irrealis, but exceptions are also found, as in –

ধারণাটি ছিলইনা বা থাকলেও\LC.y.rls ছিল একেবারে প্রাথমিক পর্যায়ে;

E.g., মাটিতে   প্রোথিত হলে\LC.n.ils   স্তম্ভমূলের মাটির উপরের অংশ বিটপে পরিবর্তিত হয়;

গ্র্যাজুয়েটরা মিনিটে 35টা ইংরিজি শব্দ টাইপ করার গতি থাকলে\LC.n.ils আবেদন করতে পারেন;

কথাটা বলতে\LC.n.rls শান্তি হল; আর্টিস্ট স্টেজে নামতেই\LC.y.rls সব চুপ

8. POSTPOSITION (PP)

A post position is the functional word that occurs after the word to indicate the inflectional markers. Post positions indicate grammatical relations between two parts of speech.

A post position is the functional word that occurs after the word to indicate the inflectional markers. Post positions indicate grammatical relations between two parts of speech.

The attributes for postpositions are-

TYPES	ATTRIBUTES
Postposition (PP)	Case Marker, Emphatic

Case markers are marked according to the function of the post position, e.g.,

অন্যদের সম্বন্ধে\PP.loc.n ভাবেন, সকলের পক্ষেই\PP.0.y চারদিকের পরিবেশ, উদাহরণ হিসেবে\PP.loc.n কালপুরুষ মন্ডলের, এ থেকে\PP.0.n একই প্রক্রিয়ায়, আলো থেকেই\PP.0.y এদের অবস্থান চিহ্নিত করা, দৃঢ় ভাবে\PP.0.n বিশ্বাস করে থাকে, এর দ্বারা\PP.0.n ভবিষ্যৎবাণী করা, ক্যানক্রি এ-কে\PP.acc.n NASA\RDF Terrestrial\RDF Planet\RDF Finder\RDF mission\RDF এর\PP.gen.n ১০০টি তারকার মধ্যে\NST.loc.n.n ৬৩তম স্থান, ৬০ -\PU এর\PP.gen.n দশকের প্রাথমিক প্রকাশের পর, দৃশ্যমান উপাত্ত নিয়েই\PP.0.y আগ্রহী, ক্রিয়াপদ ছাড়াও\PP.0.y অনুসর্গ ও অব্যয় নামের দুটি পদ আবিষ্কার করেন, আমেরিকা কর্তৃক\PP.0.n সর্বকালের সর্বশ্রেষ্ঠ বিজ্ঞান

- ▪ Sometimes spatio-temporal nouns appear in the place of postpositions; they will be tagged as NST, and **not** PP.

9. PARTICLE (C)

A particle is a word that does not belong to one of the main parts of speeches, is invariable in form, indeclinable and typically has grammatical or pragmatic meaning. It does not have any attributes. It typically has grammatical or pragmatic meaning.

TYPES	ATTRIBUTES
Coordinating (CCD)	
Subordinating (CSB)	Emphatic
Classifier(CCL)	Emphatic
Interjection (CIN)	
Others (CX)	Emphatic

The only attribute for particles is – **Emphatic** and that is applicable to three of the types mentioned above.

9.1. *COORDINATING* PARTICLE (CCD)

Coordinating particles are those particles which act as conjunctions that link constituents without syntactically subordinating one to the other. These are similar to English- *and, or* and *but*; e.g., ও\CCD, এবং\CCD, আর\CCD, বা\CCD, অথবা\CCD etc.

9.2. SUBORDINATING PARTICLE (CSB)

Subordinating particles are particles those act as conjunctions linking <u>constructions</u> by making one of them a complement of another. E.g., যাতে\CSB.n, বলে\CSB.n, যে\CSB.n, কিংবা\CSB.n , তাহলেও\CSB.y etc.

9.3. CLASSIFIER PARTICLE (CCL)

Classifier particles act as unit nouns, e.g., ৫০০ **কোটি**\CCL.n বছর, **ইত্যাদিও**\CCL.y, ১০০০ **গুন**\CCL.n, ১ **বিলিয়ন**\CCL.n মার্কিন ডলারের, ২২ **জন**\CCL.n নিহত হয়েছে, আলটেকর **প্রমুখ**\CCL.n, 59 **নং**\CCL.n ছবির, পাঁচ **ডজন**\CCL.n মন্ত্রী, একশ **ভাগ**\CCL.n একমত, সত্তর **শতাংশ**\CCL.n আর্দ্রতা

9.4. INTERJECTION (CIN)

Words that express emotion are interjections, e.g., আহা\CIN, বাহ\CIN, শাবাশ\CIN, উফ্\CIN, আরে\CIN, হায়\CIN etc.

9.5. OTHERS (CX)

This tag is used for negative particles and other particles which cannot be grouped under the above mentioned types. E.g.,

আমার তো\CX.n বিশ্বাস..., তখন ৫৬ না\CX.y কি\CX.n ৫৩ ধারা..., তাই\CX.y কুমেরু হচ্ছে..., এ অনুরোধ আর\CX.n করবেন না, এ যে\CX.n কালো মতো\CX.n জানোয়ার..., যদি নাই\CX.y যাও. আমি যাব না\CX.y

10. PUNCTUATION (PU)

Punctuation marks form a limited set of symbols that are used as punctuations and are different from other symbols that are used in a languages. Punctuations are merely symbols and they do not have any attribute. The fixed set of punctuation marks is –

। （fullstop） **,** (comma) **' '** (single quotation marks) **""**(doublequotation marks)
; (semi colon) **?** (interrogation) **!** (exclamatory mark)

11. RESIDUAL (R)

Residuals are the words those cannot be categorized under any category-type described so far. The types of Residual are-

TYPES	ATTRIBUTES
Foreign word (RDF)	
Symbol (RDS)	
Others (RDX)	

11.1.   FOREIGN WORD (RDF)

Foreign words are those words which are written in any **foreign script** other than Hindi. Any borrowed words from another written in the same (Hindi)script should not be confused with RDF. E.g.,

16\RDF, buildings\RDF, Alexander\RDF the\RDF great\RDF, ˈaɪzək\RDF ˈæzɪˌmʌv\RDF, Исаак\RDF Озимов\RDF etc.

11.2.   SYMBOL (RDS)

Symbols are characters which are not used as punctuation marks. They are not used as alphabets of the language also, e.g., *(\RDS,   )\RDS, $\RDS,   &\RDS,   +\RDS, %\RDS,   @\RDS,   #\RDS* etc.

11.3.   OTHERS (RDX)

This tag is given to words that are written in Bangla numerals. E.g., ২০০৮\RDX সাল,   ৫৭৬\RDX, জুলাই ৩১\RDX ১৯৬৫\RDX   etc.

- Onomatopoeic words would be annotated as residual others, e.g., ঠক\RDX   ঠক\RDX

Special Cases
Examples of the special cases are listed in Appendix C.

Conclusion

This guideline focuses on the morpho-syntactic description of Bangla for facilitating the annotator with the tool. Basic description of the tool consists of the definition and descritions of the tags which is instanced with examples from Bangla. Morphology in Bangla is syncretic which results into lack of one-to-one correspondence of tags and morpho-syntactic categories. This eventually leads to tag ambiguity while tagging. The framework adopted here tries to focus on the morpho-syntactic features of words to derive the appropriate attribute sets for the tags. Although natural language properties are quite systematic, exceptions are also found quite frequently. Any effort to categorize and classify natural language is thus a challenge for language technology research. This tagset tries to capture nuances of the language in the tagset, though this is a mammoth task to achieve; we expect documentation of finer nuances from the annotators for perfection of the tagset description.

APPENDIX A: Bangla Tag set

CATEGORY	TYPE	ABBREVIATION	ATTRIBUTE VALUE SET
Noun		N	
	Common		*NC {2, 6, 11, 12}*
	Proper		*NP {2, 6, 11, 12}*
	Verbal		*NV {6, 11, 12}*
	Spatio-temporal		*NST {6, 11, 12}*
Verbs		V	
	Main		*VM {3, 4, 7, 8, 9, 12, 13, 16}*
	Auxilliary		*VA {3, 4, 7, 8, 9, 12, 13, 16}*
Pronoun		P	
	Pronominal		*PPR {2, 3, 6, 10, 11, 12, 16}*
	Reflexive		*PRF {2, 6, 11, 12}*
	Reciprocal		*PRC {6, 12}*
	Relative		*PRL {2, 6, 10, 11, 12, 16}*
	Wh-		*PWH {2, 6, 10, 11, 12, 16}*
Nominal Modifier		J	
	Adjectives		*JJ {11, 12}*
	Quantifiers		*JQ {11, 12, 17}*
Demonstratives		D	
	Absolutive		*DAB {2, 12}*
	Relative		*DRL {2, 12}*
	Wh-		*DWH {2, 12}*
Adverb		A	
	Manner		*AMN {6, 11, 12}*
	Location		*ALC {6, 11, 12}*
Participle		L	
	Relative (Adjectival)		*LRL {12}*
	Verbal (Adverbial)		*LV {12, 18}*
Postposition		PP	*PP{6, 12}*
Particles		C	
	Coordinating		*CCD*
	Subordinating		*CSB {12}*
	Classifier		*CCL {12}*
	Interjection		*CIN*
	Others		*CX {12}*
Punctuation		PU	
Residual		RD	
	Foreign word		*RDF*
	Symbol		*RDS*
	Other		*RDX*

ATTRIBUTE VALUES

Attributes-*Values* and ABBREVIATIONS

1. Number NUM
   - *Singular*   *(sg)*
   - *Plural*   *(pl)*

2. Person   PER
   - *First*   *(1)*
   - *Second*   *(2)*
   - *Third*   *(3)*

3. Tense   TNS
   - *Present*   *(prs)*
   - *Past*   *(pst)*
   - *Future*   *(fut)*

4. Case-marker   CSM
   - *Accusative*   *(acc)*
   - *Genetive*   *(gen)*
   - *Locative*   *(loc)*

5. Aspect   ASP
   - *Simple*   *(smp)*
   - *Progressive*   *(prg)*
   - *Perfect*   *(prf)*

6. Mood   MOOD
   - *Declarative*   *(dcl)*
   - *Imperative*   *(imp)*
   - *Habitual*   *(hab)*

7. Finiteness FIN
   - *Finite*   *(fin)*
   - *Non-finite*   *(nfn)*
   - *Infinite*   *(ifn)*

8. Distributive DSTB
   - *Yes*   *y*
   - *No*   *n*

9. Definiteness   (DEFIN)
   - *Yes*   *y*
   - *No*   *n*

10. Emphatic   (EMPH)
    - *Yes*   *y*
    - *No*   *n*

11. Negative   (NEG)

     Yes   *y*
     No   *n*

12. Honorificity   HON
     Yes   *y*
     No   *n*

13. Numeral   NML
     Ordinal   *(ord)*
     Cardinal   *(crd)*
     Non-numeral   *(nnm)*

14. Realis   (RELS)
     Realis   *(rls)*
     Irrealis   *(ils)*

1. সরকারিভাবে\AMN.0.n.n সংগঠকদের\NC.pl.gen.n.n তরফ\NC.0.0.n.n থেকে\PP.0.n যা\PRL.sg.0.n.n.n.n জানানো\NV.0.n.n হয়েছে\VM.3.prs.pft.dcl.fin.n.n.n ,\PU তা\PPR.sg.3.0.n.n.n.n এরকম\PPR.sg.0.0.n.n.n.n :\PU মিল্লার\NP.0.gen.n.n ছেলে\NC.sg.0.n.n অসুস্থ\JJ.n.n ।\PU

2. হোটেল\NC.0.0.n.n সংলগ্ন\PP.0.n ডলারশপ\NC.0.0.n.n ,\PU যা\PRL.sg.0.n.n.n.n রাত\NC.0.0.n.n এগারোটায়\NC.0.loc.y.n বন্ধ\NC.0.0.n.n হয়েছে\VM.3.prs.pft.dcl.fin.n.n.n ,\PU তাতেও\PPR.sg.3.loc.n.n.y.n উঁকিঝুঁকি\NC.0.0.n.n দিয়ে\PU ভোগ্যবস্তুর\NC.0.gen.n.n হাতছানি\NC.0.0.n.n উপভোগ\NC.0.0.n.n করছে\VM.3.prs.prg.dcl.fin.n.n.n স্থানীয়\JJ.n.n লোকেরা\NC.pl.0.n.n ।\PU

3. রমানাথন\NP.0.0.n.n কৃষাণ\NP.0.0.n.n এরকম\PPR.0.0.0.n.n.n.n অঘটন\NC.0.0.n.n ঘটিয়েছেন\VM.3.prs.pft.dcl.fin.n.n.y ,\PU বিজয়\NP.0.0.n.n অমৃতরাজ\NP.0.0.n.n ঘটিয়েছেন\VM.3.prs.pft.dcl.fin.n.n.y ,\PU এখন\ALC.0.n.n ঘটাচ্ছেন\VM.3.prs.prg.dcl.fin.n.n.y রমেশ\NP.0.0.n.n কৃষাণ\NP.0.0.n.n ।\PU

4. 1938\RDX সালের\NC.0.gen.n.n সেপ্টেম্বর\NP.0.0.n.n ও\CCD.n অক্টোবরে\NP.0.loc.n.n নওচিকে\NP.0.acc.n.n লেখা\NV.0.n.n রবীন্দ্রনাথের\NP.0.gen.n.n পিপলস\NP.0.0.n.n ডেইলিতে\NP.0.loc.n.n তিন\JQ.n.n.crd কলম\NC.0.0.n.n জুড়ে\VM.0.0.0.0.nfn.n.n.n ছাপানো\NV.0.n.n হয়েছে\VM.3.prs.pft.dcl.fin.n.n.n ,\PU সঙ্গে\NST.0.n.n রবীন্দ্রনাথের\NP.0.gen.n.n ছবি\NC.0.0.n.n ।\PU

5. এদিকে\ALC.loc.n.n ভারতীয়\NP.0.0.n.n অলিম্পিক\NP.0.0.n.n য্যাসোসিয়েশন\NP.0.gen.n.n 20\RDX জুলাই\NP.0.0.n.n দল\NC.0.0.n.n পাঠাবার\NV.gen.n.n পরিকল্পনা\NC.0.0.n.n নিয়ে\PP.0.n 95\RDX জন\CCL.n সদস্যের\NC.0.gen.n.n বিমান\NC.0.0.n.n টিকিট\NC.0.0.n.n বুক\NC.0.0.n.n করে\VM.0.0.0.0.nfn.n.n.n রেখেছে\VA.3.prs.pft.dcl.fin.n.n.n ।\PU

6. মানুষের\NC.0.gen.n.n জীবন\NC.0.0.n.n ও\CCD.n সমাজ\NC.0.0.n.n সম্পর্কিত\JJ.n.n যে\DRL.sg.n কোন\PPR.sg.3.0.n.n.n.n ঘটনাকে\NC.0.acc.n.n তার\PPR.0.3.gen.n.n.n.n সঠিক\JJ.n.n বৈজ্ঞানিক\JJ.n.n পারস্পর্য\NC.0.0.n.n ও\CCD.n কার্য–কারণ\NC.0.0.n.n সম্পর্কের\NC.0.gen.n.n ভিত্তিতে\NC.0.loc.n.n বিচার\NC.0.0.n.n করে\VM.0.0.0.0.nfn.n.n.n দেখার\NV.gen.n.n যে\DRL.sg.n দৃষ্টিভঙ্গী\NC.0.0.n.n ,\PU ভাববাদী\JJ.n.n দার্শনিক\JJ.n.n চিন্তাভাবনার\NC.0.gen.n.n অবস্থান\NC.0.0.n.n তার\PPR.sg.3.gen.n.n.n.n বিপরীত\JJ.n.n মেরুতে\NC.0.loc.n.n ।\PU

7. ভারতীয়\NC.0.0.n.n   বৈজ্ঞানিকদের\NC.pl.gen.n.n   মধ্যে\NST.0.n.n   প্রথম\JQ.n.n.ord বরাহমিহির\NP.0.0.n.n   লক্ষ্য\NC.0.0.n.n   করেন\VM.3.prs.sim.dcl.fin.n.n.y   ,\PU প্রাচীনকালে\NC.0.loc.n.n   অশ্লেষানক্ষত্রে\NP.0.loc.n.n   সূর্যের\NC.0.gen.n.n   দক্ষিণায়ন\NC.0.0.n.n হত\VM.3.pst.sim.hab.fin.n.n.n   ,\PU   তাঁর\PPR.sg.3.gen.n.n.n.y   কালে\NC.0.loc.n.n হচ্ছে\VM.3.prs.prg.dcl.fin.n.n.n পুনর্বসু\NP.0.0.n.n নক্ষত্রে\NC.0.loc.n.n ।\PU

8. রমেশচন্দ্র\NP.0.0.n.n বিদ্বান\NC.0.0.n.n ছিলেন\VM.3.pst.sim.dcl.fin.n.n.y ,\PU শক্তিমান\JJ.n.n লেখক\NC.0.0.n.n ছিলেন\VM.3.pst.sim.dcl.fin.n.n.y ,\PU কিন্তু\CSB.n তাঁর\PPR.sg.3.gen.n.n.n.y চিন্তাভাবনা\NC.0.0.n.n কোনো\PPR.sg.0.0.y.n.n কোনো\PPR.sg.0.0.y.n.n দিক\NC.0.0.n.n থেকে\PP.0.n সমালোচনার\NC.0.gen.n.n যোগ্য\NC.0.0.n.n ।\PU

9. ইংরেজ\NC.0.0.n.n   চলে\VM.0.0.0.0.nfn.n.n.n   গেলে\LC.n.ils   জাতীয়তাবাদ\NC.0.0.n.n কতদূর\JQ.n.n.nnm   পজিটিভ\JJ.n.n   আর\CCD.n   কতদূর\JQ.n.n.nnm   নেগেটিভ\JJ.n.n তার\PPR.sg.3.gen.n.n.n.n   পরীক্ষার\NC.0.gen.n.n   সময়\NC.0.0.n.n আসবে\VM.3.fut.sim.dcl.fin.n.n.n ।\PU

10. মুখিশপুরে\NP.0.loc.n.n   লোহগড়\NP.0.0.n.n   নামে\PP.0.n   একটি\JQ.y.n.crd   দুর্গ\NC.0.0.n.n গড়ে\VM.0.0.0.0.nfn.n.n.n তোলেন\VA.3.pst.sim.dcl.fin.n.n.n ।\PU

11. বিবেকজ্ঞান\NC.0.0.n.n লাভের\NC.0.gen.n.n জন্য\PP.0.n অপরিহার্য\JJ.n.n সাধন\NC.0.0.n.n -\PU পদ্ধতি\NC.0.0.n.n   নির্দেশ\NC.0.0.n.n   করাই\NV.0.n.y   যোগদর্শনের\NC.0.gen.n.n   মুখ্য\JJ.n.n কাজ\NC.0.0.n.n ।\PU

12. এ\PPR.sg.3.0.n.n.n.n যে\CX.n কালো\JJ.n.n -\PU মতো\CX.n জানোয়ার\NC.0.0.n.n লম্বা\JJ.n.n হয়ে\VM.0.0.0.0.nfn.n.n.n   মাটির\NC.0.gen.n.n   সঙ্গে\PP.0.n   মিশে\VM.0.0.0.0.nfn.n.n.n আছে\VA.3.prs.prg.dcl.fin.n.n.n ।\PU

13. জ্যামিতিক\JJ.n.n   আকৃতি\NC.0.0.n.n   ,\PU   রেখা\NC.0.0.n.n   ,\PU   তরঙ্গায়িত\JJ.n.n রেখা\NC.0.0.n.n ,\PU ব্যাস\NC.0.0.n.n ,\PU ব্যাসার্ধ\NC.0.0.n.n ,\PU রঙের\NC.0.gen.n.n ভঙ্গিমার\NC.0.gen.n.n বিচিত্র\JJ.n.n ও\CCD.n জটিল\JJ.n.n ব্যবহারে\NC.0.loc.n.n চেনা\NV.0.n.n - \PU জানা\NV.0.n.n -\PU শোনা\NV.0.n.n বস্তুগুলিও\NC.pl.0.n.y অমৃত\JJ.n.n (\RDS য়্যাবস্ট্রাক্ট\JJ.n.n)\RDS প্রকাশে\NC.0.loc.n.n নতুন\JJ.n.n রূপ\NC.0.0.n.n ও\CCD.n অর্থ\NC.0.0.n.n পরিগ্রহণ\NC.0.0.n.n করে\VM.3.prs.sim.dcl.fin.n.n.n ,\PU পরিবর্তন\NC.0.0.n.n করে\VM.3.prs.sim.dcl.fin.n.n.n   তাদের\PPR.pl.3.gen.n.n.n.n   পুরানো\JJ.n.n আত্মপরিচয়\NC.0.0.n.n ।\PU

14. এদের\PPR.pl.3.gen.n.n.n স্বরবৈচিত্র্য\NC.0.0.n.n মনে\NC.0.loc.n.n রাখবার\NV.gen.n.n চেষ্টা\NC.0.0.n.n কর\VM.3.prs.sim.imp.fin.n.n ,\PU -\PU কোকিল\NP.0.0.n.n ,\PU ময়ূর\NP.0.0.n.n ,\PU পেঁচা\NP.0.0.n.n ,\PU পাপিয়া\NP.0.0.n.n ,\PU কাকাতুয়া\NP.0.0.n.n ,\PU বুলবুল\NP.0.0.n.n ,\PU পায়রা\NP.0.0.n.n ,\PU বসন্তকৌরী\NP.0.0.n.n ,\PU ঘুঘু\NP.0.0.n.n ,\PU কাঠঠোকরা\NP.0.0.n.n ,\PU বেনেবৌ\NP.0.0.n.n ।\PU

15. খেলাধূলায়\NC.0.loc.n.n ,\PU বিশেষত\AMN.0.n.n সাধারণ\JJ.n.n অর্থে\NC.0.loc.n.n যাকে\PRL.sg.acc.n.n.n বলে\VM.3.prs.sim.dcl.fin.n.n স্পোর্টস\NC.0.0.n.n ,\PU তাতে\PPR.sg.3.loc.n.n.n.n আমার\PPR.sg.1.gen.n.n.n.n উতসাহ\NC.0.0.n.n ছিল\VM.3.pst.sim.dcl.fin.n.n না\CX.y ;\PU অনেক\JQ.n.n.nnm খেলাই\NC.0.0.n.y আমার\PPR.sg.1.gen.n.n.n.n কাছে\NST.loc.n.n বিরক্তিকর\JJ.n.n বলে\CSB.n মনে\NC.0.loc.n.n হত\VM.3.pst.sim.hab.fin.n.n.n ।\PU

16. এ\DAB.sg.n ঢং\NC.0.0.n.n -\PU এ\PP.loc.n আমরা\PPR.pl.1.0.n.n.n.n পূর্বে\NST.loc.n.n কোন\DAB.sg.n নাচ\NC.0.0.n.n নাচিনি\VM.1.prs.pft.dcl.fin.n.y.n ।\PU

17. এবার\ALC.0.n.n ইউনিভার্সিটি\NC.0.0.n.n ইন্সটিটিউটে\NC.0.loc.n.n `\PU দাঁড়াও\NP.0.0.n.n পথিকবর\NP.0.0.n.n '\PU ,\PU `\PU দুঃস্বপ্নের\NP.0.gen.n.n নগরী\NP.0.0.n.n '\PU ,\PU `\PU আজকের\NP.0.0.n.n সাজাহান\NP.0.0.n.n '\PU এবং\CCD.n `\PU টিনের\NP.0.gen.n.n তলোয়ার\NP.0.0.n.n '\PU -\PU এই\DAB.sg.n চারটে\JQ.y.n.crd নাটকের\NC.0.gen.n.n অংশ\NC.0.0.n.n তুলে\VM.0.0.0.0.nfn.n.n.n নেব\VA.1.fut.sim.dcl.fin.n.n.n ।\PU

18. অতীত\JJ.n.n দিনের\NC.0.gen.n.n বিভিন্ন\JJ.n.n রাজবংশের\NC.0.gen.n.n ইতিহাস\NC.0.0.n.n ,\PU শাসন\NC.0.0.n.n ব্যবস্থার\NC.0.gen.n.n বিভিন্ন\JJ.n.n অধ্যায়\NC.0.0.n.n ,\PU আন্দোলন\NC.0.0.n.n ,\PU বিপ্লব\NC.0.0.n.n ,\PU বিভিন্ন\JJ.n.n রাজনৈতিক\JJ.n.n প্রতিষ্ঠানের\NC.0.gen.n.n বিবর্তন\NC.0.0.n.n প্রভৃতি\CCL.n রাষ্ট্রবিজ্ঞানের\NP.0.gen.n.n আলোচনায়\NC.0.loc.n.n নূতন\JJ.n.n তথ্য\NC.0.0.n.n সরবরাহ\NC.0.0.n.n করে\VM.3.prs.sim.dcl.fin.n.n.n এবং\CCD.n তাহার\PPR.0.3.gen.n.n.n.n ভিত্তিতে\NC.0.loc.n.n রাষ্ট্রবিজ্ঞানের\NC.0.gen.n.n নূতন\JJ.n.n আদর্শ\NC.0.0.n.n ও\CCD.n তত্ত্বের\NC.0.gen.n.n উদ্ভব\NC.0.0.n.n হইতে\VM.0.0.0.0.ifn.n.n.n পারে\VA.3.prs.sim.dcl.fin.n.n.n ।\PU

19. সিসি\NC.0.0.n.n (\RDS c.c.\RDF)\RDS ও\CCD.n মিলিলিটার\NC.0.0.n.n (\RDS ml.\RDF)\RDS একই\JJ.n.y ।\PU

20. (4)\RDF ভারতীয়\JJ.n.n সমভূমি\NC.0.0.n.n :\PU অবস্থান\NC.0.0.n.n -\PU আরাবল্লী\NP.0.0.n.n পর্বতের\NC.0.gen.n.n পশ্চিমে\NST.loc.n.n রাজস্থানে\NP.0.loc.n.n ভারতের\NP.0.gen.n.n মরুভূমি\NC.0.0.n.n অবস্থিত\JJ.n.n ।\PU

1. প্রায়\AMN.0.n.n ৩২৫\RDX মিলিয়ন\CCL.n কপি\NC.0.0.n.n

2. যা\PRL.sg.0.n.n.n.n তাকে\PPR.sg.3.acc.n.n.n.n বই\NC.0.0.n.n লিখে\VM.0.0.0.0.nfn.n.n.n বড়লোক\NC.0.0.n.n হওয়া\NV.0.n.n প্রথম\JQ.n.n.ord বিলিয়নিয়ার\NC.0.0.n.n (\RDS আমেরিকান\JJ.n.n ডলারে\NC.0.loc.n.n)\RDS হিসেবে\PP.0.n প্রতিষ্ঠা\NC.0.0.n.n করেছে\VM.3.prs.pft.dcl.fin.n.n.n ।\P

3. শিকাগো\NP.0.0.n.n ধর্ম\NC.0.0.n.n মহাসম্মেলনে\NC.0.loc.n.n

4. ভারত\NP.0.0.n.n পরিব্রমণরত\JJ.n.n স্বামিজীকে\NP.0.acc.n.n ক্ষেত্রীর\NP.0.gen.n.n মহারাজ\NC.0.0.n.n বিবেকানন্দ\NP.0.0.n.n নাম\NC.0.0.n.n অর্পণ\NC.0.0.n.n করেন\VM.3.prs.sim.dcl.fin.n.n.y ।\PU

5. রামকৃষ্ণ\NP.0.0.n.n পরমহংস\NP.0.0.n.n -\PU এর\PP.gen.n নামে\NC.0.loc.n.n রামকৃষ্ণ\NP.0.0.n.n মিশন\NP.0.0.n.n প্রতিষ্ঠা\NC.0.0.n.n করেন\VM.3.prs.pft.dcl.fin.n.n.n ।\PU

6. তিনি\PPR.0.3.0.n.n.n.y জীবনে\NC.0.loc.n.n ৫০০\RDX '\PU র\PP.gen.n -\PU ও অধিক\JJ.n.n গ্রন্থ\NC.0.0.n.n রচনা\NC.0.0.n.n ও\CCD.n সম্পাদনা\NC.0.0.n.n করেছেন\VM.3.prs.pft.dcl.fin.n.n.y ,\PU লিখেছেন\VM.3.prs.pft.dcl.fin.n.n.y প্রায়\AMN.0.n.n ৯,০০০\RDX চিঠি\NC.0.0.n.n ও\CCD.n পোস্টকার্ড\NC.0.0.n.n ।\PU

7. গ্র্যান্ড\JJ.n.n মাস্টার\NC.0.0.n.n আখ্যায়\NC.0.loc.n.n ভূষিত\JJ.n.n করা\NV.0.n.n হয়েছে\VA.3.prs.pft.dcl.fin.n.n.n ।\PU

8. বিজ্ঞান\NC.0.0.n.n বিষয়ক রচনায়\NC.0.loc.n.n

9. আইজাক\NP.0.0.n.n আসিমভ\NP.0.0.n.n (\RDS জানুয়ারি\NP.0.0.n.n ২\RDX ?\PU ,\PU ১৯২০\RDX ?\PU [১] –\PU এপ্রিল\NP.0.0.n.n ৬\RDX ,\PU ১৯৯২\RDX)\RDS ,\PU আইপিএ :\PU /\PU ˈaɪzək\RDF ˈæzɪˌmɑv\RDF /\PU ,\PU মূলত\AMN.0.n.n Исаак\RDF Озимов\RDF

10. ডিউই\NP.0.0.n.n দশমিক\NC.0.0.n.n পদ্ধতি\NC.0.0.n.n দ্বারা\PP.0.n চিহ্নিত\JJ.n.n ১০টি\JQ.y.n.crd প্রধান\JJ.n.n বিষয়ের\NC.0.gen.n.n ৯টির\RDX\NC.0.gen.y.n উপরেই\NST.loc.n.y তিনি\PPR.sg.3.0.n.n.n.y লিখেছেন\VM.3.prs.pft.dcl.fin.n.n.y ।\PU

11. অন্য\JJ.n.n দুজন\NC.0.0.y.n ছিলেন\VM.3.pst.sim.dcl.fin.n.n.y

12. বিগ\JJ.n.n থ্রি\NC.0.0.n.n -\PU এর\PP.gen.n একজন\NC.sg.0.y.n ছিলেন\VM.3.pst.sim.dcl.fin.n.n.y ।\PU

13. ১০০টি\JQ.y.n.crd বিষয়ের\NC.0.gen.n.n

14. একটির\NC.sg.gen.y.n উল্লেখ\NC.0.0.n.n

15. যে\DRL.sg.n দুটিকে\NC.sg.acc.y.n পরবর্তিতে\NC.0.loc.n.n তিনি\PPR.sg.3.0.n.n.n.y ফাউন্ডেশন\NC.0.0.n.n সিরিজের\NC.0.gen.n.n অন্তর্ভুক্ত\JJ.n.n করেছেন\VM.3.prs.pft.dcl.fin.n.n.y

16. যাঁরা\PRL.pl.0.n.n.n.y এই\DAB.sg.n গবেষণায়\NC.0.loc.n.n রত\JJ.n.n

17. তাঁরা\PPR.pl.3.0.n.n.n.y আগ্রহী\JJ.n.n নন\VM.3.prs.sim.dcl.fin.n.y.y ।\PU

18. দাঁড়\VM.0.0.0.0.0.n.n.n করাতে\VA.0.0.0.0.ifn.n.n.n চেষ্টা\NC.0.0.n.n করেন\VM.3.prs.sim.dcl.fin.n.n.y

19. খ্রিস্টপূর্ব\JJ.n.n ১ম\JQ.n.n.ord শতাব্দীতে\NC.0.loc.n.n

20. সমাজভাষাবিজ্ঞান\NP.0.0.n.n ,\PU মনোভাষাবিজ্ঞান\NP.0.0.n.n ,\PU গণনামূলক\NP.0.0.n.n ভাষাবিজ্ঞান\NP.0.0.n.n ,\PU ইত্যাদির জন্ম\NC.0.0.n.n দিয়েছে

21. একটি\JQ.y.n.crd অতি\JQ.n.n.nnm প্রাচীন\JJ.n.n শাস্ত্র\NC.0.0.n.n হলেও\LC.y.rls কেবল\AMN.0.n.n ১৯শ\JQ.n.n.ord শতকে\NC.0.loc.n.n এসেই\VM.0.0.0.0.nfn.y.n.n

22. নন\VM.3.prs.sim.dcl.fin.n.y.y

23. বেশির\NC.0.gen.n.n ভাগ\NC.0.0.n.n

24. ব্যাকরণ\NC.0.0.n.n সম্বন্ধে\PP.0.n বলা\NV.0.n.n হয়ে\VM.0.0.0.0.nfn.n.n.n থাকে\VA.3.prs.sim.dcl.fin.n.n.n

References

1. Dandapat, S. 2008. *MSRI Part-of-Speech Annotation Interface*. Draft. Microsoft Research India Private Limited: Bangalore.
2. *Framework for a Common Parts-of-SpeechTagset for Indian Languages*. Draft. Microsoft Research India Private Limited: Bangalore.
3. Paul, S. 2004. *An HPSG Account of Bangla Compound Verbs with LKB Implementation*. Ph.D. Dissertation. CALTS: University of Hyderabad.
4. Chattopadhyay, Suniti. 1939/2003. ভাষা প্রকাশ বাঙ্গালা ব্যাকরণ *(bhAShA prakAsh bAngAlA byAkaraNa)*. Rupa: New Delhi.

Contact: Priyanka Biswas
t-pribis@microsoft.com,

biswas.priyanka@gmail.com