MSRI Part-of-Speech Annotation Interface

Sandipan Dandapat

1. Introduction

The POS annotation interface is a GUI for assigning the appropriate Part-of-Speech (POS) for each word in a sentence. This is an annotation tool designed for annotating within the framework of a hierarchical tagset designed at MSRI. This is a generic toll developed for adapting flat tagset, hierarchical tagset (upto level two). The toll also provides the facility to provide the morphological attributes with their values. The interface supports various operations for viewing and editing the POS labels and morphological information associated with the word.

2. System Requirement and Installation

The following software resources are required

1) Microsoft Windows Operating System
2) Microsoft Visual C# 2005 or above

To install the annotation tool, extract the .rar file. Execute the *Annotation Tool.exe* to start the interface.

3. How to Define a Tagset

To adopt a new tagset following resources are required

1) *tag_type.txt* – to store the Tag and the corresponding Types
2) *type_att.txt* – to store the type and the corresponding set of attributes
3) *att_val.txt* – to store the attributes and its values
4) *mapping* – represents the abbreviation of the higher level Tag and Type description

3.1 'tag_type.txt' – this file stores one *Tag* and the corresponding *Types* in each line. Make sure that the **Tag and Type are separated by a single TAB** and **Types are separated by a comma**. For example

Noun	Common,Proper,Verbal,Spatio-temporal
Verb	Main,Auxiliary

The first field (e.g. Noun, Verb etc.) will appear at the first level of the hierarchical tagset. Further, each element in the second field (separated by comma) will appear in the second level of the hierarchical tagset. If a tag does not have any Type (e.g. Postposition or Punctuation); specify the first field only and don't put any TAB after that. The ordering of the file will be maintained in the menu strip in the annotation tool.

3.2 'type_att.txt' – this file stores the lists of attribute corresponds to a particular *Tag-Type*. Here also two fields are separated by a TAB. The *Tag and Type* are specified by a single space among them. For example

        Noun Common   Num,Csm,Defin,Emph
        Verb Main   Per,Tns,Asp,Mood,Fin,Emph,Neg,Hon

The second field of the file specifies the names of *attributes* corresponding to the particular *Tag-Type*. The ordering of the second field will be maintained for storing the attributes corresponding to the specific Tag-Type. This is also will be maintained as the default display order in the annotation tool for setting the attributes. Further, one can specify the display order of his own by adding a third field which contains the ordering of attributes (separated by comma). For example

        Noun Common   Num,Csm,Defin,Emph   3,2,1,4#

This essentially signifies that in interface the display order will be Definiteness first, then Case Marker, then Number and finally Emphatic Marker. Please note that the third field only changes the display order but the storing order will be same as defined in the second field. **Make sure the spelling of *Tag-Type* in this file is same as defined in the '*tag_type.txt*' file**.

3.3 'att_val.txt' – this file stores an attribute and the corresponding values in each line. This file specifies all the attributes that has been used to in the 'type-att.txt' file. Each line of the file is specified by two fields which are separated by a TAB. First field defines the name of an attribute and the second field defines the all possible values separated by a comma. For example

        Num   sg,pl,du,0,x
        Per   1,2,3,0,x
        Tns   prs,pst,fut,0,x

The first value of a particular attribute will be set as the default value. Make sure the attribute names are same as defined in the 'type_att.txt'.

3.4 'mapping.txt' – this file represents the abbreviation of the Tag-Type which will be used for storing and viewing. This is required to reduce the length of the Tag-Type which otherwise difficult to view a very large Tag associated with every word after annotation. Here also each line of the file is represented by two fields and the fields are separated by a TAB. For example

        Noun Common   NC
        Verb Main   VM
        Pronoun Wh-   PWH

The sample file for Bengali tagset is given in Appendix C.

Important:
1) **Make use the same name (as described above) for incorporating a new tagset in the annotation tool.**
2) Maintain the format and structure as described above
3) Once the above four files are compiled put them in the same directory where the .exe file exists. Run the .exe to start the interface of annotation Tool.

4. Interface

In this section we will give a brief overview of the each component in the GUI based interface. The detail of the annotation task will be given in the next section. Once the tool is properly installed into the system the main interface window will appear.
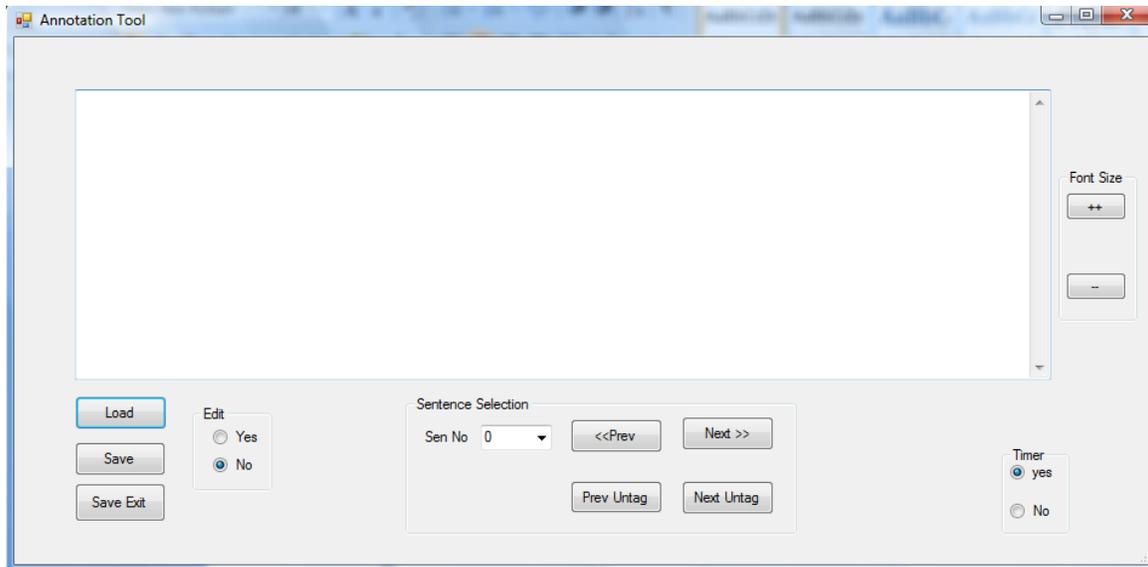

Figure 1: The main interface window of the MSRI Part-of-Speech Annotation tool

The interface has the following components directly visible in the main interface

1) *Text Box* – that supports viewing, labeling and editing the POS tags to each word in a sentence

2) *Load* Button – to load a file/corpus for the annotation task. This also read the corpus into a data structure for the task. Once a file/corpus is chosen by this button the first line of the corpus will appear in the textbox for POS annotation.

3) *Save* Button - This button helps to save the current work into the file so that the work in progress can be saved after some amount of work.

4) *Save and Exit* Button - By pressing this button one can save the annotated corpus and come out from the GUI interface. It is recommended to use this button after every annotation session.

5) In the interface there is a *Sentence Selection* area which provides the user to select a particular sentence for the task. It has the following five components:

     i.   *Sen No* – The user can move to any sentence by selecting the Sen No from the combo box.

    ii.   *'<<Prev'* and *'Next>>'* - move to the previous and next sentences respectively.

   iii.   *Prev Untag* and *Next Untag* –move to the previous and next untagged sentences directly. Untagged sentence refers to a sentence where at least one of the words is not annotated during the previous annotation sessions. Thus, partially tagged sentences are also considered as untagged sentence.

6) *Font Size* - In the right side of the interface has a *Font Size* area and two buttons – '++' and '- -'. These two buttons help to increase and decrease the font size of the text that appears in the *Text Box* respectively.

7) *Timer* – It keeps track of the user time taken to annotate each sentence. The timer is by default set to "enabled". However, one can disable the timer by selecting '*No*' radio button.

8) *Edit* – It toggles between Normal mode and Edit mode facilities. The edit mode is by default set to '*No*'. However, one can enable the edit status to '*Yes*' for correcting the tagged sentences.

5. The Task of Annotation

5.1 Corpus Data format: The first step for the POS annotation task is to prepare the corpus data file. The annotation tool assumes that the user will be using corpus files in Unicode (UTF-8, UTF-16 and UTF-32) encoding. One can also use files with ASCII rendered text e.g. - iTRANS, Roman Readable and wx notation.

Make sure that the corpus is broken into one sentence per line, which is required for loading one sentence at a time for the annotation task. Also, it is good to separate each lexical

unit by a space during the corpus creation task. This is due to the fact that each lexical unit will be annotated with a separate tag. Otherwise, some manual editing is required during the annotation. For example, punctuations are usually attached with the previous word in a sentence, but it's a separate lexical item for the POS tagging task. Therefore, the punctuations should be separated by a space form the previous word.

5.2 Loading the Corpus Data File: To open a corpus file press the *Load* button. A dialog box will appear, allowing you to find the desired corpus from your computer. After finding the corpus from the dialog box click open. The corpus must be a text (.txt) file. Once the corpus file is opened, the first sentence will appear in the Text Box for annotation as depicted in the figure bellow.
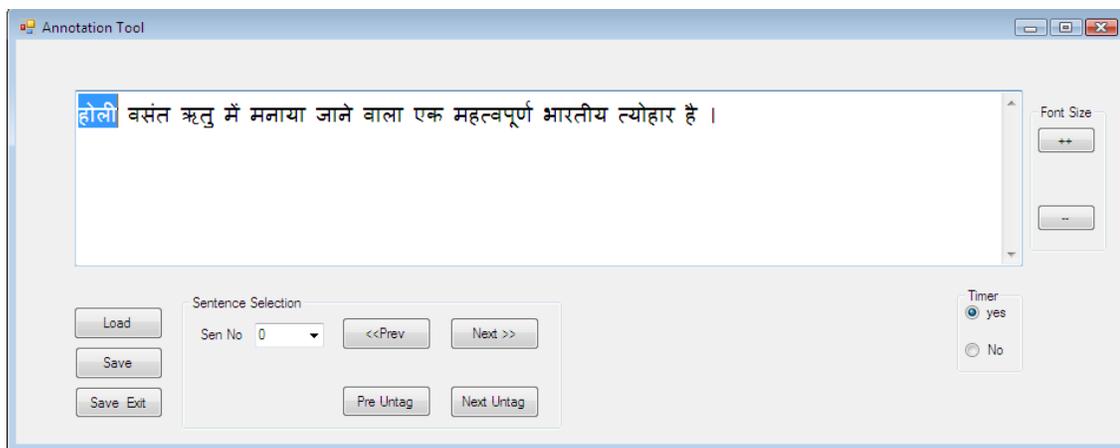

Figure 2: Corpus data file is uploaded and first sentence appears with very first word highlighted for annotation.

Once the sentence is loaded, the first untagged word will be automatically highlighted for annotation. Even if the sentence is partially/fully annotated the first untagged lexical item will be highlighted. Further, the selection module will continue automatically with selecting the next lexical item. Once the user assigns a POS tag to the highlighted word, the automatic selection will move to the next lexical item. The detail of the automatic selection will be discussed in the subsection 4.4.

5.3 Assigning POS Tag

Once a word is highlighted, click the right button of the mouse (the mouse cursor can be placed anywhere in the textbox) and the first level of the hierarchical tagset (*Category*) will appear. By selecting the appropriate *Category* one can move to the next level in the hierarchy i.e. – *Type* and select the *Type* by making a left click.
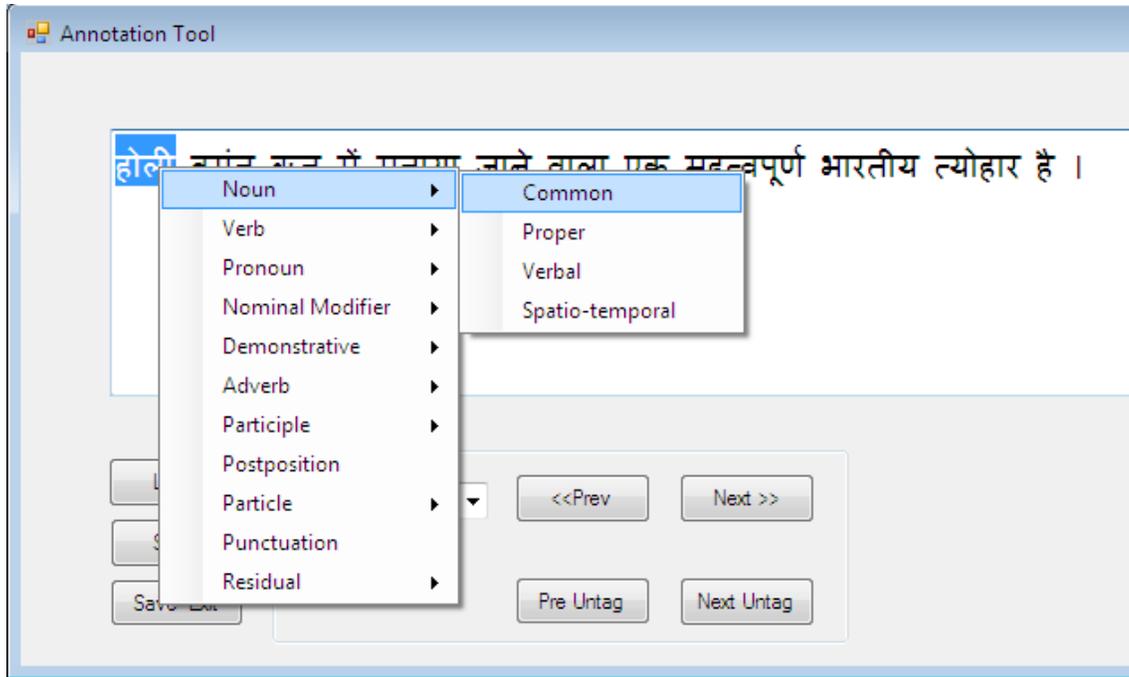
Figure 3: Category and Type selection by mouse right click event

As soon as the *Type* is selected, a new *Form* will appear for the assignment of the *attributes*. All the attributes appear in the new Form with a default value. One can change the value by selecting one item from the drop drown box associated with each attribute. Refer to the annotation guideline for the details of the attributes and their corresponding values.
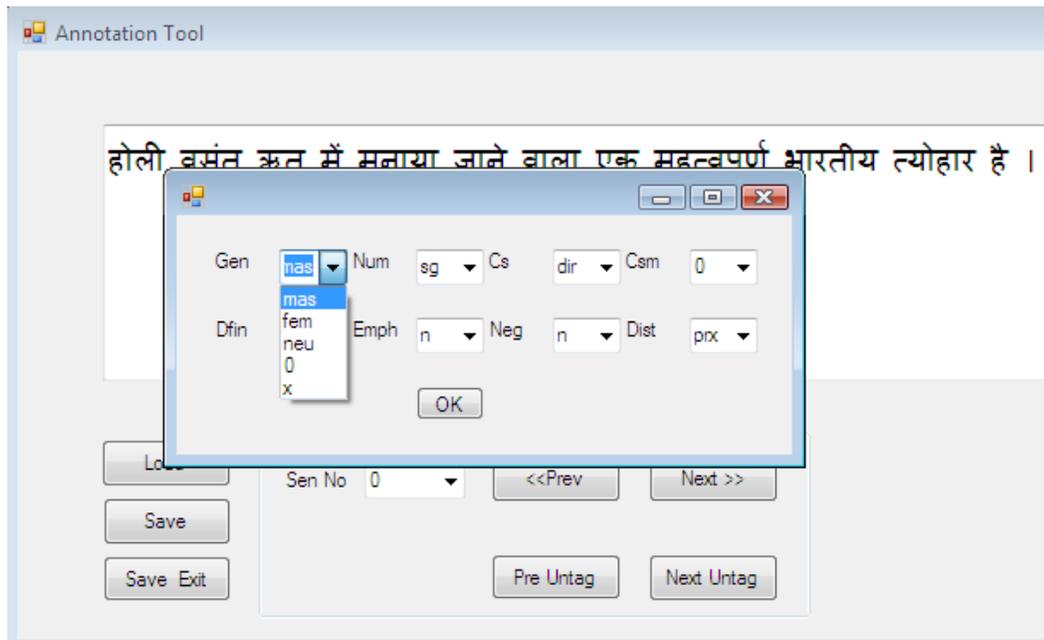

Figure 4: Window for assignment of morphological attribute values

After setting the values for all the attributes, click *OK*. This will assign the write tag at the end of the selected word and will highlight the next word for the annotation.

Note: if no lexical item was highlighted during the selection of the tag, an error message will appear and will be asked to select a lexical item (if a lexical unit is partially selected then the tag will appear at the end of the selection).
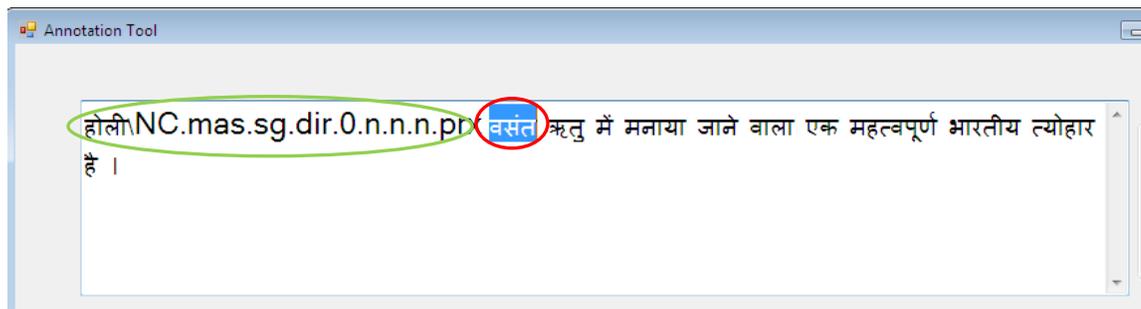

Figure 5: the tagged word (in green oval) and the next highlighted lexical item (in red oval)

Continue the same process for annotating all the words in a sentence. Once a sentence is annotated, one can move to any other sentence using the *Sentence Selection* area (Sec. 3).

After the completion of a Corpus data file, it's important to Save and Exit from the interface by clicking the associated button. If in case, one does not save the corpus data file after the work and exit the interface by *close* option on the right corner of the interface, a *.tmp* file with the same name of the Corpus data file will be generated in the present working directory as a backup of the work. This .tmp file can be used for further annotation.

5.4 Automatic Selection: As mentioned in the previous section, the next lexical unit will be automatically highlighted for POS tag assignment. This automatic selection process can be stopped by a left-click on the text box area. In case the selection module is stopped and needs to be invoked again, select (by mouse click) a lexical unit manually and assign its POS Tag (as described in subsection 4.3). From this point the automatic highlighting process will start again.

5.5 Timer Option: The timer is used to keep record of the time taken to annotate each sentence by a user. If the timer is on, a record will be created in the present working directory with name: *user_log*. Each line of the file has the following information:

< *Username, sentence number, start time, end time, time taken*>

One can stop the timer by selection the 'No' radio button.

6. The Task of Editing

Correction of human/machine annotated POS tagged data is one of the major tasks during the creation of POS labeled corpora. The present version of the tool incorporates the facility of editing the annotated sentence.

6.1 Set the *Edit* mode: Manually select the Edit status to '*Yes*' through the radio button in the Edit frame (by default the status is '*No*').

6.2 Load the corpus: Load the file through *load* button as described in section 5.2. One needs to upload a POS annotated text file in the editing mode.

6.3 Selection of TAG during editing: Once the annotated file is opened through the interface for editing, the first category-type is highlighted automatically (see figure 6). It is often the case that a few tags are erroneous. Thus the annotator needs to go to that particular tag for correcting them. We have introduced two new keys from the key-board for moving back and forth in the annotated sentences.
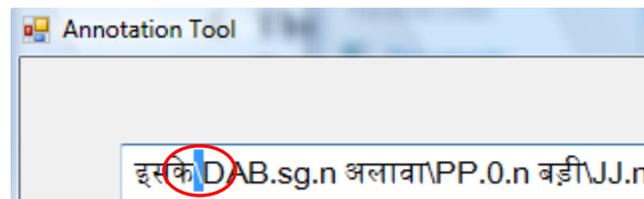

Figure 6: the automatically selected category-type for editing

(a) Shift key – pressing shift key automatically points to the next category-type/attribute for editing. One *shift key press* selects the next editable potion (either the Category-type or Attribute) of the sentence. If the current selection is a Category-type then a *shift key press* will points to the attribute (if exists) of the word. Thus, one can traverse the whole sentence by pressing shift keys to select the particular tag which needs to be corrected. This selection goes from left to right in the sentence.
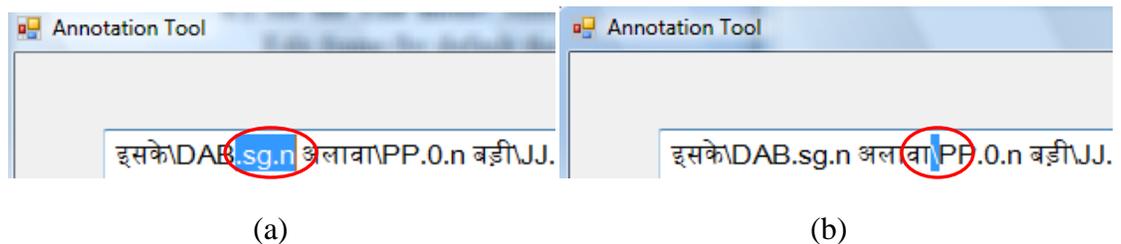

(a)   (b)

Figure7: (a) One *shift key press* select the next editable item (attribute). Next *shift key press* selects the next category-type as in (b)

(b) Ctrl key – *Ctrl key* has the similar functionality of *shift key*. *Ctrl key* press moves the selection form right to left. This is required to go back in the sentences automatically by pressing key. However, Ctrl key moves from category-type to category-type. *Ctrl key* does not select attribute. This is due to the fact that a backward movement essentially denotes correcting the previous word's POS annotation (i.e. Category + Attribute).

6.4 Correcting Category: Once a category is selected (which is highlighted by '\'), assign the new POS tag as described in section 5.3. The change in POS tag signifies the changes in attribute also. Thus after selecting the correct POS label a new *Form* will appear for assigning morphological attributes.

6.5 Correcting Attribute Values: This is to handle the situation when the POS category is correct for a particular word in a sentence but the attribute values are wrongly assigned. One need to select the attribute set by pressing a number of shift keys. Once the attribute set is selected, press right mouse button to change the values of the attribute. A new form will appear after the right mouse button press. This new form will appear with the entire attribute and the old values (as described in the tagged sentence). One can select the correct values for an attribute from the drop drown box associated with each attribute (see section5.3). Once a correction is made the automatic selection module will points to the next editable entity in the sentence.

The selection of '*Tag-Type*' or '*Attribute*' can also be done manually for editing with the facility as described above. Make sure that during manual selection the tag is selected including the '\' for editing both Tag-type and attribute. The attribute (the whole set or a portion of it but it should contain at least one '.') is selected including '.' without the TAG for editing the attribute values.

7. Bug list
Following are the list of bugs in the present version of the MSRI POS annotation tool

1. If a lexical unit is partially selected (manually by the user), then the tag will appear at the end of the selection.

Appendix A: Hindi Tagset and Abbreviation

CATEGORIES and *TYPES with Attribute applicable*

Noun	N	
	Common	*NC {1, 2, 5, 6, 11, 12, 13, 14}*
	Proper	*NP {1, 2, 5, 6, 11, 12, 13, 14}*
	Verbal	*NV {5, 6, 11, 12}*
	Spatio-temporal	*NST {5, 6, 11, 12, 13}*
Verbs	V	
	Main	*VM {1, 2, 3, 4, 7, 8, 9, 12, 13, 16}*
	Auxiliary	*VA {1, 2, 3, 4, 7, 8, 9, 12, 13, 16}*
Pronoun	P	
	Pronominal	*PPR {1, 2, 3, 4, 5, 6, 10, 11, 12, 13, 15, 16}*
	Reflexive	*PRF {1, 2, 5, 6, 11, 12}*
	Reciprocal	*PRC {1, 5, 6, 11, 12}*
	Relative	*PRL {2, 5, 6, 10, 11, 12, 13, 15, 16}*
	Wh-	*PWH {2, 5, 6, 10, 11, 12, 13, 15, 16}*
Nominal Modifier	J	
	Adjectives	*JJ {1, 2, 5, 13}*
	Quantifiers	*JQ {1, 2, 5, 11, 12, 13, 17}*
Demonstratives	D	
	Absolutive	*DAB {2, 5, 12, 14}*
	Relative	*DRL {2, 5, 12, 14}*
	Wh-	*DWH {2, 5, 12, 14}*
Adverb	A	
	Manner	*AMN {5, 6, 11, 12, 13}*
	Location	*ALC {5, 6, 11, 12, 13}*
Participle	L	
	Relative (Adjectival)	*LRL {1, 2, 4, 12, 13}*
	Verbal (Adverbial)	*LV {12, 13}*
	Nominal	*LN {1, 2, 4, 7, 12, 13}*
	Conditional	*LC {12, 13, 18}*
Postposition		PP *{1, 2, 6, 11, 12}*
Particles	C	
	Coordinating	*CCD {13}*
	Subordinating	*CSB {13}*
	Classifier	*CCL {13}*
	Interjection	*CIN {13}*
	Others	*CX {13}*
Punctuation	PU	

Residual		RD
Foreign word		*RDF*
Symbol		*RDS*
Other		*RDX*

<u>ATTRIBUTES</u> (The default attribute values are marked with a '*')

Attributes   and   *values*

1. Gender GEN
   - *Masculine*   *(mas)**
   - *Feminine*   *(fem)*
   - *Neuter*   *(neu)*

2. Number   NUM
   - *Singular*   *(sg)**
   - *Plural*   *(pl)*
   - *Dual*   *(du)*

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

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

5. Case   CS
   - *Direct*   *(dir)**
   - *Oblique*   *(obl)*

6. Case-marker   CSM
   - *Ergative*   *(erg)*
   - *Accusative*   *(acc)*
   - *Instrumental*   *(ins)*
   - *Dative*   *(dat)*
   - *Genetive*   *(gen)*
   - *Sociative*   *(soc)*
   - *Locative*   *(loc)*
   - *Ablative*   *(abl)*
   - *Benefective*   *(bnf)*
   - *Vocative*   *(voc)*
   - *Purposive*   *(pur)*

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

8. Mood   MOOD
        Declarative   *(dcl)**
        Subjunctive   *(sbj)*
        Conditional   *(cnd)*
        Imperative   *(imp)*
        Presumptive   *(psm)*
        Abilitative   *(abt)*
        Habitual   *(hab)*

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

10. Distributive   DSTB
        Yes   *y*
        No   *n**

11. Definiteness   (DEFIN)
        Yes   *y*
        No   *n**

12. Emphatic   (EMPH)
        Yes   *y*
        No   *n**

13. Negative   (NEG)
        Yes   *y*
        No   *n**

14. Distance   (DIST)
        Proximal   *(prx)*
        Distal   *(dst)*
        Sequel   *(seq)*

15. Incl/Excl   (SET)
        Inclusive   *(inl)**
        Exclusive   *(exl)*

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

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

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

Appendix B: Bengali Tagset and Abbreviation

CATEGORIES and *TYPES with Attributes applicable*

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, 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 {6, 12}*
Particles	C	
Coordinating		*CCD {13}*
Subordinating		*CSB {13}*
Classifier		*CCL {13}*
Interjection		*CIN {13}*
Others		*CX {13}*
Punctuation	PU	
Residual	RD	

Foreign word	*RDF*
Symbol	*RDS*
Other	*RDX*

ATTRIBUTES

(The default attribute values are marked with a '*')

Attributes   and   *values*

2. Number   NUM
Singular	*(sg)**
Plural	*(pl)*
Dual	*(du)*

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

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

6. Case-marker   CSM
Ergative	*(erg)*
Accusative	*(acc)*
Instrumental	*(ins)*
Dative	*(dat)*
Genetive	*(gen)*
Sociative	*(soc)*
Locative	*(loc)*
Ablative	*(abl)*
Benefective	*(bnf)*
Vocative	*(voc)*
Purposive	*(pur)*

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

8. Mood MOOD
Declarative	*(dcl)**
Subjunctive	*(sbj)*
Conditional	*(cnd)*
Imperative	*(imp)*
Presumptive	*(psm)*

Abilitative	*(abt)*	
Habitual	*(hab)*	

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

10. Distributive   DSTB
        Yes   *y*
        No   *n**

11. Definiteness   (DEFIN)
        Yes   *y*
        No   *n**

12. Emphatic   (EMPH)
        Yes   *y*
        No   *n**

13. Negative   (NEG)
        Yes   *y*
        No   *n**

14. Distance   (DIST)
        Proximal   *(prx)*
        Distal   *(dst)*
        Sequel   *(seq)*

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

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

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

Appendix C: Sample files for Bengali tagset

'tag_type.txt'

Noun   Common,Proper,Verbal,Spatio-temporal
Verb   Main,Auxiliary
Pronoun   Pronominal,Reflexive,Reciprocal,Relative,Wh-
Nominal Modifier   Adjective,Quantifier
Demonstrative Absolutive,Relative,Wh-
AdverbManner,Location
Participle   Verbal,Conditioanl
Particle   Coordinating,Subordinating,Classifier,Interjection,Other
Residual   Foreign Word,Symbol,Other
Punctuation
Postposition

'type_att.txt'

Noun Common   Num,Csm,Defin,Emph   3,2,1,4#
Noun Proper   Num,Csm,Defin,Emph
Noun Verbal   Csm,Defin,Emph
Noun Spatio-temporalCsm,Defin,Emph
Verb Main   Per,Tns,Asp,Mood,Fin,Emph,Neg,Hon
Verb Auxiliary   Per,Tns,Asp,Mood,Fin,Emph,Neg,Hon
Pronoun Pronominal   Num,Per,Csm,Dstb,Defin,Emph,Hon
Pronoun Reflexive   Num,Csm,Emph
Pronoun Reciprocal   Csm,Emph
Pronoun Relative   Num,Csm,Dstb,Defin,Emph,Hon
Pronoun Wh- Num,Csm,Dstb,Defin,Emph,Hon
Nominal Modifier Adjective Defin,Emph
Nominal Modifier Quantifier Defin,Emph,Nml
Demonstrative Absolutive   Num,Emph
Demonstrative Relative   Num,Emph
Demonstrative Wh-   Num,Emph
Adverb Manner   Csm,Defin,Emph
Adverb Location   Csm,Defin,Emph
Participle Verbal   Emph
Participle Conditional Emph,Rels   2,1#

Particle Coordinating Neg
Particle SubordinatingNeg
Particle Classifier   Neg
Particle Interjection   Neg
Particle Other Neg
Postposition   Csm,Emp

'att_val.txt'

Num   sg,pl,du,0,x
Per   1,2,3,0,x
Tns   prs,pst,fut,0,x
Csm   acc,gen,loc,0,x
Asp   sim,prg,pft,0,x
Mood   dcl,sbj,cnd,imp,psm,abt,hab,0,x
Fin   fin,nfn,ifn,0,x
Dstb   n,y,0,x
Defin   n,y,0,x
Emph   n,y,0,x
Neg   n,y,0,x
Hon   n,y,0,x
Nml   ord,crd,nnm,0,x
Rels   rls,ils,0,x

'mapping.txt'

Noun Common   NC
Noun Verbal   NV
Noun Proper   NP
Noun Verbal   NV
Noun Spatio-temporalNST
Verb Main   VM
Verb Auxiliary   VAUX
Pronoun Pronominal   PPR
Pronoun Reflexive   PRF
Pronoun Reciprocal   PRC
Pronoun Relative   PRL
Pronoun Wh- PWH
Nominal Modifier Adjective JJ

Nominal Modifier Quantifier	JQ
Demonstrative Absolutive	DAB
Demonstrative Relative	DRL
Demonstrative Wh-	DWH
Adverb Manner	AMN
Adverb Location	ALC
Participle Verbal	LV
Participle Conditional	LC
Particle Coordinating	CCD
Particle Subordinating	CSB
Particle Classifier	CCL
Particle Interjection	CIN
Particle Other	CX
Residual Foreign Word	RDF
Residual Symbol	RDS
Residual Other	RDX
Punctuation	PU
Postposition	PP