DEVELOPING LANGUAGE RESOURCES

FOR ENGLISH/□□□□□ MACHINE TRANSLATION

A Thesis

Submitted to the Department of Computer Science and Engineering

of

BRAC University

By

Rabia Sultana Ummi

Student ID: 03101037

Fahmina Huda

Student ID: 04301006

In Partial Fulfillment of the Requirements for the Degree

of

Bachelor of Science in Computer Science and Engineering

August 2008

DECLARATION

We hereby declare that this thesis is based on the results found by ourselves. Materials of work found by other researchers are mentioned by reference. This theis, neither in whole nor in part, has been previously submitted for any degree.

Signature of Supervisor

Signature of Authors

ACKNOWLEDGMENTS

Special thanks to Dr. Mumit Khan for his support, teachings and supervision during the entire period of work on this paper. Special thanks to Mr. Altaf Mahmud, the developer of Bangla tagset used in this research, for supplying us with the initial resource needed for this thesis and for helping us understand the tagset better.

ABSTRACT

We developed English-Bangla parallel corpora for statistical machine translation. By hand we tagged 20,000 words of our Bangla corpus according to their particular part of speeches. In our work we also suggested a method for identifying word correspondence in parallel English-Bangla text using a translation model based on part of speech and n-gram model.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

1. Introduction

Machine translation investigates the use of computer software to translate text or speech from one natural language to another [3]. There are three approaches to developing machine translation. These are rule based, example based and statistical.

Statistical machine translation tries to generate translations using statistical methods based on bilingual text corpora, such as the Canadian Hansard corpus, the English-French record of the Canadian parliament. Where such corpora are available, impressive results can be achieved translating texts of a similar kind, but such corpora are still very rare [3].

In linguistics, a corpus (plural corpora) or text corpus is a large and structured set of texts (now usually electronically stored and processed). They are used to do statistical analysis, checking occurrences or validating linguistic rules on a specific universe.

A corpus may contain texts in a single language (monolingual corpus) or text data in multiple languages (multilingual corpus). Multilingual corpora that have been specially formatted for side-by-side comparison are called aligned parallel corpora.

In order to make the corpora more useful for doing linguistic research, they are often subjected to a process known as annotation. An example of annotating a corpus is part-of-speech tagging, or POS-tagging, in which information about each word's part of speech (verb, noun, adjective, etc.) is added to the corpus in the form of tags.

Machine translation for English-Chinese, English-Arabic, English-French, and many other pairs of languages are relatively quite advanced. On the other hand, very little work has been done in developing machine translation software for translating English to Bangla. Whatever work has been done, it’s for the Bangla of Western Bengal in India.

In response to the lack of resources for English-Bangla machine translation, our paper describes the development of a parallel English-Bangla Corpora and a recommendation of a statistical method to word-align Bangla text to corresponding English text in the corpora, using a translation model based on part of speech (POS) and n-gram model. In the process, a 20,000 words part of speech tagged Bangla corpus was developed. These resources are required to develop machine translation software for translating Bangla text or speech to English text or speech or vice versa.

To begin with, we had only a Bangla tagset which was under-development. With this meager resource, we set out to developing other resources required for implementing an English-Bangla statistical machine translation as mentioned above. In Section 2 of this paper, we describe the Bangla tagset and the major modification we have instigated to finalize documentation for the tagset. In Section 3, we describe the methodologies we have used in developing the resources mentioned above. Finally, in the next two sections, there are brief discussions on the outcome of this thesis and the future steps needed in the development of English-Bangla machine translation.

2. The tagset and Suggestions

The initial Bangla tagset, we started working on had 50 tags. After scrutinizing it, we suggested some modifications. Based on these suggestions, the developer made some other modifications to the rules. Additionally, he increased the number of tags to 55. After these modifications, a final documentation has been developed.

The final tagset has 55 tags and 17 categories. Each category includes a set of subcategories. All the 55 tags, their categories and subcategories, and examples have been included in Table 1. Level 1 of the table refers to the category of the tag while Level 2 refers to subcategory of the tag.

Table 1
55-Tag Bangla Tagset

Level 1	**Level 2**	**Tag**	**Word**
Noun	Common	NN	□□□□□, □□□□
	Proper	NNP	□□□□□, □□□□□□□, □□□□, □□□□□□□□□, □□□□□□
	Compound Common Noun	NNC	□□□□/NNC □□□□/NN, □□□□□□□□□□/NNC □□□□□□□□□□/NN
	Compound Proper Noun	NNPC	□□□□□□/NNPC □□□□□/NNPC □□□□□□□/NNP
	Verb Root	NNV	□□□□, □□□
	Temporal	NNT	□□□□□, □□□□□□□□, □□
	Question Temporal	QNT	□□□, □□□
	Locative	NNL	□□□, □□□, □□□
	Question Locative	QNL	□□□□□, □□□□□, □□□□□□

Level 1	**Level 2**	**Tag**	**Word**
Pronoun	Personal Pronoun	PRP	□□□, □□□□, □□□□, □□□□□, □□, □□□□, □□□□, □□□□, □□□, □□□□
	Question Pronoun	QPR	□□, □□□□, □□, □□□□
Adjective	Simple	JJ	□□□□□□, □□□, □□□, □□□□□□□, □□□□□□□□□, □□□□□□□□□
	Verb Root	JJV	□□□, □□□□□□
	Question Adjective	QJJ	□□□□, □□□□
Vocative	Vocative	VOC	□□□, □□□, □□□
Verb	Main Finite Verb	VB	□□□, □□, □□□, □□□□, □□□□□, □□□□, □□□□□□, □□□, □□□□
	Nonfinite Nominal	VBM	□□□, □□□□□, □□□, □□□□□
	Nonfinite Conditional	VBC	□□□□, □□□□□
	Nonfinite Perfective	VBT	□□□, □□□□
	Nonfinite	VBF	□□□□, □□□□□
	Finite Existential	VBE	□□, □□□
	Nonfinite Existential	VBEF	□□□
Adverb	Adverb	RB	□□□□□, □□□□, □□□□□, □□, □□□, □□□, □□□□, □□□□
	Question Adverb	QRB	□□□, □□□□□□, □□□□□□
Conjunction	Coordinating	CC	□□□, □, □□□□□, □□□□, □□□□□
	Compound Coordinating	CCC	□□/CCC □□/CC
	Suspicion	CN	□□□, □□□□
	Eternal Joining	CET	□□□□/CET ... □□□□/CET, □□□/CET ... □□□/CET, □□□/CET ... □□□/CET
	Subordinating	CS	□□, □□□□□, □□□, □□□□□□
	Compound Subordinating	CSC	□□□/CSC □□□/CS, □□/CSC □□□□□/CS
Postposition	Postposition	ON	□□□□□□, □□□□□□, □□□, □□□□, □□□□, □□□□, □□□□□
Interjection	Interjection	UH	□□□□!, □□□!, □□□!
Particle	Particle	RP	□□, □□, □□□
	Question Particle	QRP	□□
Determiner	Common	DT	□□□, □□□□, □□□, □□□□□, □□, □
	Singular	DTS	□□□, □□□
	Question Determiner	QDT	□□□□□, □□□□, □□□□□□□, □□□□□□, □□□□□
Quantifier	Quantifier	QF	□□, □□□, □□□, □□, □□□□
	Quantifier Number	QFNUM	□, □, □□, □□□, □□□□, □□□□□□
	Question Quantifier	QQF	□□, □□, □□□□□□

Foreign Word	Foreign Word	FW	□□□□□ □□□□□□ □□□□
Symbol	Symbol	SYM	□□□□□□□□□ □□ □□□□□□□□□□□□ □□□□□ □□□□□, □□□□□□□□
List Item Marker	List Item Marker	LS	a, b, (a), 1, 2.3.1, □, □.□□
Level 1	**Level 2**	**Tag**	**Word**
Suffix	Postpositional	SFON	□, □, □□
	Accusative	SFAC	□□, □□, □□□, □□□□□, □□□□□□
	Possessive	SF$	□□, □□□
Punctuation Mark	Sentence Final Punctuation	.	|, ?, !
	Comma	,	,
	Colon, Semi-colon	:	:, ;
	Dash, Double-Dash	-	-, --
	Left Parenthesis	(	({ [
	Right Parenthesis	)	) }]
	Opening Left Quote	LQ	', "
	Closing Right Quote	RQ	', "

There have been various issues where we felt modifications are needed in the tagset. These issues and some of our suggestions are discussed below.

In certain situations where a title is associated in a name such as *Dr.* in *Dr. Rabeya Khatun,* we had suggested using a new tag for the title preceding the actual name. Our suggestion was based on the tagset of C7 where the *Dr.* is tagged NNB since it is before the actual name. Similarly, we suggested using a new tag for a noun following the actual name, as in example *Dr. Rabeya Khatun M.B.B.S.* In C7, *M.B.B.S* would be tagged as NNA since it's after the actual name. Therefore, it would be very useful to use a separate tag for nouns such as the ones mentioned above. In arguments against our suggestion, explanations terming increase in the size of the Bangla tagset as problematic have been provided. Since there was no rule assigned for these cases, the developer of the tagset modified the tagset to tag these titles as NN.

In cases of name of degree, award, etc used with a name (following it) such as *M.B.B.S* in *Dr. Rabeya Khatun M.B.B.S* and □□□ □□□□□ক in □□□□□□ □□□□ □□□ □□□□□□, no rules have been suggested in the documentation. We suggest tagging these words as NN.

Additionally, there is no tag suggested when a title is added to a name (preceding it) such as □□□-□-□□□□□ in □□□-□-□□□□□ □ □□ □□□□□ □□ and □□□ □□□□□□□ in □□□ □□□□□□□ □□□□□□□ □□□□□. So, we

suggest tagging them, along with the whole name, as NNPC. This is so because these titles have become unavoidable part of the name.

According to the documentation, DTS are “Single determiners precede nouns and determinate objects those are singular in number and cannot be inflected by any suffix.” To contradict this, we would like to give examples like □□□ □□□□ □□□□. Here □□□□ □□□□ is a compound common noun and □□□□ is inflicted by suffix. Thus, DTS can also be used when a singular noun is inflected. In □□□ □□□□ □□ □□□ □□□, □□□□ and □□□ are mass nouns. Therefore, a correction has to be made so as to include mass nouns in the definition.

Since in Bangla part of speech, there is no adverb. Therefore, there are arguments favoring tagging words in Bangla, corresponding to adverbs in English as nouns or adverbs. To resolve this situation, we have been advised to tag these words as nouns (NNT or NNL). For example, উপর in তুমি উপরে যাও has been tagged as NNL instead of RB, following the above rule mentioned.

While POS-tagging the Bangla words, we have come across certain words like “□□□□□”, “□□□”, “□□□□□ □□□□□” and “□□□□□□□□”. There were no guidelines in the documentation for such confusing words. After a discussion with the developer of the tagset, the documentation was updated and tags for such words were included.

One other change we suggest is using separate tags for the punctuation marks in Bangla as in the case of C7. Each punctuation mark in Bangla has separate uses. For example, ‘!’ and ‘|’ should not be both be tagged as sentence final punctuation, ‘.’, because the exclamation sign has more use than just terminating a sentence.

Using the final tagset and its documentation, we started developing POS-tagged English-Bangla Parallel Corpora. In the next section, we discuss the methodologies used in reaching our goal.

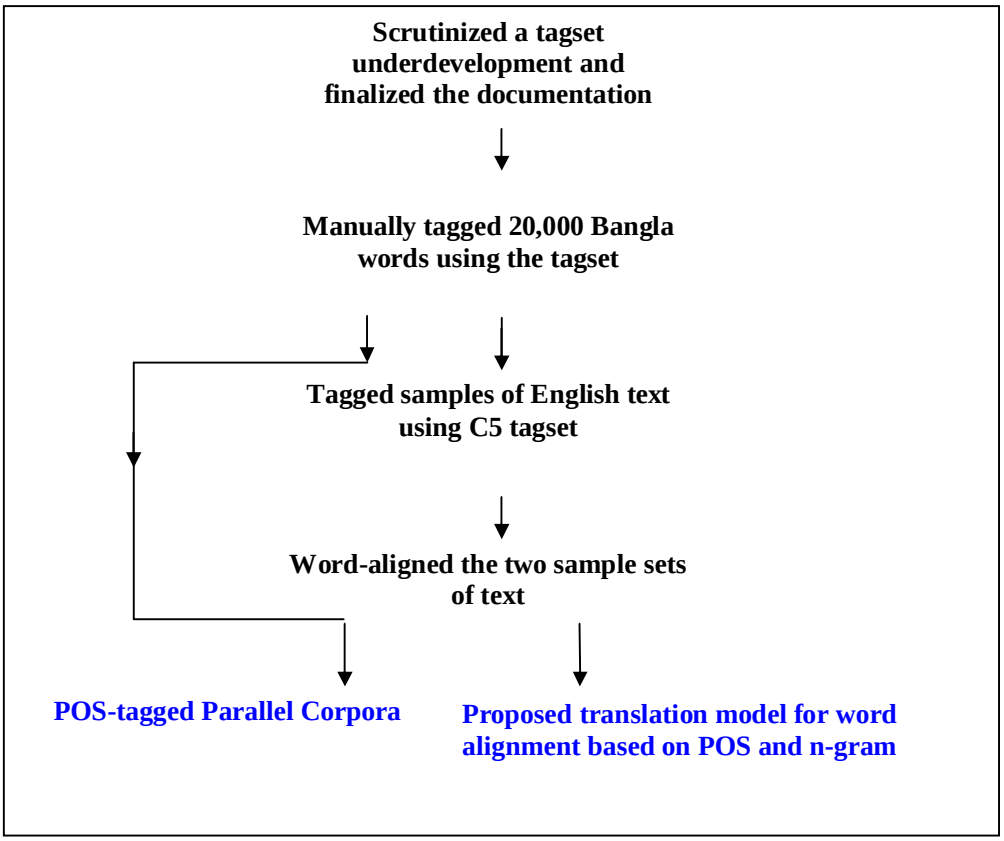
3. Methodology

An overview of the process we used to develop a POS-tagged Parallel Corpora and a proposition of translation model for word alignment based on POS and n-gram model can be seen in the form of Figure 1. As already mentioned in the previous section, we first scrutinized the initial tagset and then used the final tagset to manually tag 20,000 Bangla words. The source of these words is www.bdnews24.com, a

Bangladeshi news website. Various news reports in Bangla and their corresponding reports in English were taken. Then, 20,000 words of the Bangla reports were tagged using the tagset. Combining these, tagged texts with their corresponding English texts, we developed the POS-tagged parallel corpora. To serve as examples, we have included two sets of extracts from the parallel corpora. The first set includes the English sentence “Rupali is now undergoing treatment at Dhaka Medical College Hospital” with its corresponding Bangla sentence “রূপালি র্বতমানে ঢাকা মেডিকেল কলেজ হাসপাতালে চিকি সাধীন রয়েছে।” and the second one consists of “Her sister and another worker suffered injuries in the accident and were sent to hospital” as English text and “র্দুঘটনায় তার বোন ও অপর এক শ্রমিক আহত হন এবং তাদের হাসপাতালে পাঠানো হয়েছে।” as its corresponding Bangla sentence.

Additionally, some samples of English texts were tagged using the C5 tagset for English. The tagged Bangla text was combined with the samples of tagged English text to manually word align the two set of texts. To word align the two sets of text, we used a technique, a proposition we have made for future work in the area, for word alignment using n-gram model and POS.

Figure 1: Annotated English-Bangla Parallel Corpora



In order to word align the Bangla text corresponding to the English text; we have used an algorithm called PosAlign. Chang et al suggest using this algorithm to word align English-Chinese texts. The algorithm has three steps [2]. The steps are:

1. Tag the parallel text with part-of-speeches

2. Initial alignment with the help of cognate lookup

3. Train the translation model iteratively using the unaligned part of the POS sequences

After the tagging of the two sets of texts in Bangla and English, an initial alignment using cognate look up is done. Cognate includes proper nouns, pronouns, numerical expressions and punctuation marks [2]. We also suggest using relationships (father, mother, sister, brother-in-law, etc) as cognates because size of the set of corresponding translations in Bangla is very limited.

Initially, we hand aligned some parts of the texts. Using these parts, we developed an n-gram model to devise a statistical translation model. We used this model on the unaligned part of the POS tagged set of sample texts in English and Bangla. Advantage of using POS alignment is parallel text of limited size is required for training.

To understand the process better, we present a discussion that walks through the process of word alignment of a Bangla sentence to its corresponding English sentence, using PosAlign. The sentences are pre-tagged. This paper will use the following Bangla and English tagged sentences:

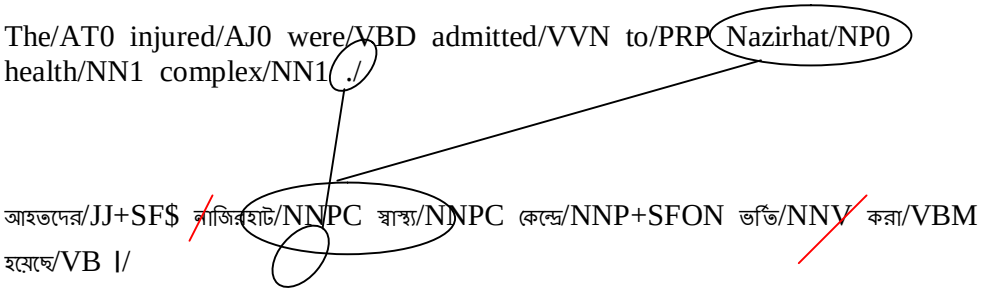
“The/AT0 injured/AJ0 were/VBD admitted/VVN to/PRP Nazirhat/NP0 health/NN1 complex/NN1./.”

and

“আহতদের/JJ+SF$ নাজিরহাট/NNPC স্বাস্থ্য/NNPC কেন্দ্রে/NNP+SFON ভর্তি/NNV করা/VBM হয়েছে/VB ।/.”

Firstly, cognate look up is used to word align the sentences partially. As a result, we align “Nazirhat” to “নাজিরহাট” and ‘.’ to ‘।’ as in Figure 2. Next, we remove the suffix tags, also represented in the figure. Here, we notice that in the example used only the Bangla sentence has suffixes. After these actions, the resulting tag sets of the English and Bangla sentences look like {AT0, AJ0, VBD, VVN, PRP, NN1} and {JJ, NNPC, NNP, NNV, VBM, VB} respectively.

Figure 2: Cognate Lookup and Suffix Removal



Then we use the n-gram model, to word align the rest of the sentences. A subset of the n-gram model is shown in Table 2. We shall use this subset in aligning

the rest of the sentences. The table is a very simple one where there are columns titled "বাংলা/English" meaning the Bangla tag given the English tag and the conditional probability of the Bangla tag given the English tag or the conditional probability of a sequence of Bangla tags given a sequence of English tags. We look up the first tag in the Bangla set, "JJ" in the table. There are three entries for "JJ". The first one is "JJ/ATO,AJO" with a probability of 1.00, the second one is "JJ/AJ0" with a probability of 0.67 and the third one is "JJ/NN1" with a probability of 0.33. Of these three, we choose the one with the highest probability, i.e. "JJ/ATO, AJO". Thus "আহতদের/JJ" is aligned to "The/AT0 injured/AJ0".

Table 2
n-Gram model for Word Alignment

□□□□□/English	**Conditional Probability**	□□□□□/English	**Conditional Probability**
NNPC/NN1	0.40	NNPC/AJ0	0.40
NNP/NN1	1.00		
NNV,VBM,VB/VBD,VVN	1.00		
VBM,VB/VBD,VVN	1.00		
JJ/AT0,AJO	1.00		
JJ/AJ0	0.67	JJ/NN1	0.33
Φ/AT0	0.33	Φ/PRP	0.50

Next we look up "NNPC" and we find two entries, "NNPC/NN1" and "NNPC/AJO" both with equal probability of 0.4. Thus, there is a tie. And still there remains a problem of aligning health with □□□□□□□□□ and complex with কেন্দ্র. To resolve this issue and also the tie, we have devised a rule. The rule suggests that if there is an NNPC aligned with a NP0 in our sentences by using the method of cognate lookup, first count the number of tags in the Bangla compound proper noun sequence (here it is 3; NNPC NNPC NNP) and then check whether NP0 is the starting tag of a sequence in the English sentence that contains the same number of tags and ends with either NP0 or NN1. If we have got a sequence like NP0……NP0, all the tags are to be NP0s there. Else if we have a sequence like NP0……NN1, the sequence may contain other tags like AJ0. If so, then align the two sequences with each other using the same order as they appear in the sentences as following.

AT0 AJ0 VBD VVN PRP NP0 NN1 NN1 .

JJ NNPC NNPC NNP NNV VBM VB .

We also lookup the sequence “NNV,VBM,VB” in the table and locate the entry “NNV,VBM,VB/VBD,VVN” with a probability of 1.00. Thus, we align the rest of the words in English with the rest of the words in Bangla. This way the whole sentence in Bangla is word aligned 100% accurately to the sentence in English.

4. Summary of Result

The outcome of the thesis is the development of resources required to develop English-Bangla machine translation. A 20,000 Bangla word tagged parallel corpora has been built.

Our method of word alignment of sentences using POS and n-gram model has shown good performances, but more tests are required. Encouraged by this good performance of the alignment process, we propose this technique for word alignment of sentences.

5. Future Work and Conclusion

The 20,000 word annotated Bangla corpus will be very useful, not only in the area of machine translation, but also in various other areas of natural language processing. Moreover, the bilingual corpus can be used for statistical machine translation.

This paper encourages with suggestions and guidelines for those who wish to work in Sentence Alignment, Cognate Dictionary and similar other areas. It serves as the foundation for the development of statistical machine translation for English-Bangla text or speech processing.

After this stage of work in this area, future work is required in statistical sentence alignment. From there, an electronic cognate dictionary must be built. Once these two processes are done, advanced work in statistical machine translation using POS alignment can be done.

List of References

[1] Daniel Jurafsky, James H. Martin, *Speech and Language Processing*, University of Colorado, Boulder, Pearson Education, Inc., 2000

[2] Jyun-Sheng Chang, Huey-Chyun Chen, *Using Part-of-Speech Information in Word Alignment*, National Tsing Hua University, Conference of the Association for Machine Translation in the Americas (1994)

[3] *Wikipedia, The Free Encyclopedia*, http://en.wikipedia.org/wiki/Machine_translation

APPENENDICES

Finalized documentation of Mr. Altaf Mahmud's 55-Tag Bangla Tagset has been appended in this section.

Language: Bangla

ID No: 1
Part of Speech: Common Noun
Tag: NN
Category: Noun

Example:
□□/DT □□□□□□**/NN+SFON** □□□□/QF □□□□□□**/NN** □□□□/VB
"Many germs live in this water"

□□□**/NN** □□□**/NN** □□□/VB
"Cow eats grass"

Description and Analysis:
A common noun is one which can be preceded by the definite article and that represents one or all of the members of a class. A common noun also appears as a root of a verb and can be tagged as a verb root (NNV).

Possible confusing tags:

NN or NNP—See **NNP or NN**

NN or JJ—See **JJ or NN**

NN or NNT—See **NNT or NN**

NN or NNV—See **NNV or NN**

NN or DT—See **DT or NN**

NN or VBM—See **VBM or NN**

NN or QF—See **QF or NN**

NN or QFNUM—See **QFNUM or NN**

ID No: 2
Part of Speech: Proper Noun
Tag: NNP
Category: Noun

Example:
□□□□□**/NNP** □□□□/QFNUM □□□□□□/JJ
“Motiur is a warrior”

□□□□□□□□**/NNP** □□□□□□/PRP+SF$ □□□□□□/NN+SF$ □□□/NN
“December is our month of victory”

Description and Analysis:
A noun belongs to the class of words used as names for unique individuals, events or places. It is also called *proper name*.

Proper nouns are not usually preceded by determiners.

Possible confusing tags:

NNP or NN
Compounds containing proper nouns as their second constituent, such as ‘□□□□-□□□□□’ (mid-March) should be tagged as proper nouns (NNP).

When either quotation marks or italic type emphasizes that an instance of a word refers to the word itself rather than its associated concept, the word should be tagged as proper noun (NNP).

Example:
‘□□□□□□**/NNP’** □□□□ □□□□□ □□□□□ □□□□ □□□ □□□□□□ □□□□□
‘□□□□□**/NNP’**
“ ‘Kedara’ is an old Bangla word which is called ‘chair’ in English”

But,
□□□□ □□□□□□ □□’□□ □□□□□□**/NN** □□□
“I have two chairs in my room”

The name of a degree or a title that appears at the starting of a name of a person should be tagged as common noun (NN).

Example:
□□□□□□□**/NN** □□□□□□/NNPC □□□□□/NNP
“Advocate Abdur Rahman”

□□**/NN** □□□□□/NNPC □□/NNP
“Dr. Nurul Huq”

But sometimes the name of a degree or a title appears at the end of a name to become a part of the whole name. In that case, it should be tagged as proper noun (NNP).

Example:
□□□□□□/**NNPC** □□□□□/**NNPC** □□□□□□□/**NNP**
“Abdur Rahman Advocate”

NNP or NNT—See **NNT or NNP**

NNP or JJ—See **JJ or NNP**

NNP or NNP+SFON—See **NNP+SFON or NNP**

ID No: 3
Part of Speech: Compound Common Noun
Tag: NNC
Category: Noun

Example:
□□□□**/NNC** □□□□**/NN** □□□□/QF □□□□□/VBF □□□□□□/VB
"All boys and girls have gone to play"

□□□□□□□□□□/NNP+SF$ □□□□□□□□□□/DT □□□□□/NN+SFON
□□□□□□/VB □□□□/QFNUM □□□□**/NNC** □□□□□□□**/NN**
"There is one district commissioner in each of the districts of Bangladesh"

Description and Analysis:
This part has been taken from [2].
There is no separate tag for Compound Common Noun in the Penn tagset. But in this tagset, the tag NNC is used for compound nouns. This tag has been introduced in order to identify unhyphenated compound words as one unit. All words except the last one, of compound words will be marked as NNC. Thus, any NNC will always be followed by another NNC or an NN. This category helps identify these words as one unit although they are not conjoined by a hyphen.

Possible confusing tags:

NNC or JJ—See **JJ or NNC**

ID No: 4
Part of Speech: Compound Proper Noun
Tag: NNPC
Category: Noun

Example:
□□□□□□**/NNPC** □□□□□**/NNPC** □□□□□□□**/NNP**
“Abdur Rahman Bishwas”

□□□□□□□□□□□□□□□**/NNPC** □□□□□□□□**/NNPC** □□□□□**/NNP**
“The Government of People’s Republic of Bangladesh”

Description and Analysis:
This part has been taken from [2].
This tag is also an addition. All words, excluding the last one, in a compound proper noun will be marked as NNPC. Just as the NNC tag, this tag too helps identify a compound proper noun as one unit and not confuse it with a list of proper nouns.

ID No: 5
Part of Speech: Verb Root
Tag: NNV
Category: Noun
Example:

□□□/PRP □□□□**/NNV** □□□□□/VB
“I have taken a bath”

□□/PRP □□□/NNT □□/NN □□□**/NNV** □□□□/VB
“He is drinking tea now”

Description and Analysis:
This part has been taken from [2].
This tag has been introduced to account for the concept of “kriyamuls” of any Indian language. “Kriyamuls” are verbs formed by combining a noun or an adjective with a helping verb. The “kriyamuls” formed by joining a noun will be NNV. As in the above example '□□□□' (bath) is a noun which is joined to the verb '□□□□□' (done) to express the sense of the verb 'to bathe'. So here '□□□□' (bath) is marked as NNV and the main verb is marked as VB.

Possible confusing tags:

NNV or NN
A ‘kriyamul’ (verb root) is generally followed by a helping verb and in that case, it will be tagged as a verb root (NNV) as well, but when it occurs as a separate word or takes a possessive marker, it acts as a common noun. In this case, it will be common noun (NN).

Example:
□□□□ □□□□ □□□□**/NNV** □□□□/VB
“We are all taking bath”

But,
□□□□□□**/NN+SF$** □□□□/NN □□□ □□□□□
“There is no water here for taking bath”

□□ □□□□□□ □□□□□**/NN+SFON**
“He went to take a bath”

Moreover, when it is also in possessive relation (the possessed of possessor), it tends to be tagged as common noun (NN).

Example:
□□□□/PRP+SF$ □□□□**/NN** □□□□ □□□□
“I haven’t yet taken bath”

NNV or VBM—See **VBM or NNV**

ID No: 6
Part of Speech: Temporal Noun
Tag: NNT
Category: Noun

Example:
□□/PRP □□/**NNT** □□□□/VB
"He is coming today"

□□□□/PRP □□□□□□□□/**NNT** □□□□□/NNL+SFON □□□□□/VB
"He will go home tomorrow"

Description and Analysis:
It is a noun that belongs to a class of words denoting a time span. It has been distinguished from noun due to a slight different syntactic distribution i.e. it is not necessary to be inflected by a postpositional marker when it contains the time of event occurred. At the example given above, if the word '□□' (today) is inflected by a postpositional marker '□□', both the syntactic and semantic distribution of the sentence won't be altered.

Example:
□□ □□□□/**NNT+SFON** □□□□
"He is coming today"

They can be adverbs:
□□ □□/**RB** □□□□

And they can also behave like nouns when followed by postposition
□□ □□/**NN** □□□/ON □□□□ □□

Above examples show that these are most likely to be nouns, but they are not pure nouns. So they have been categorized as 'Temporal Noun' (NNT). If such words are tagged according to their syntactic function, it may decrease the automatic POS tagging (such as HMM or BRILL tagger) performance, or can hamper machine learning in any free word order language such as Bangla. Therefore, in such cases, we shall use NNT rather than RB or NN.

Possible confusing tags:

NNT or NN
Sometimes temporal nouns appear as expressions of a counting event (e.g. something like counting day) and followed by a postposition (ON). In that case, it should be tagged as common noun (NN).

Example:
□□□/**NNT** □□□□ □□□□□□□□□
"I went to Dhaka yesterday"

But,
□□ □□/**NNC** □□□/**NN** □□□ □□□ □□□□□
"He is counting days everyday"

NNT or NNP
When the name of a month or a day itself denotes a time or duration of an event, it will be tagged as temporal noun (NNT) and not as proper noun (NNP).

Example:
□□/JJ □□□/JJ □□□□□□/**NNT** □□□□□□□ □□□ □□□□□ □□□□□□
"Bangla New Year was held on the 14th of April"

□□ □□/JJ □□□□□□/**NNT** □□□□□□
"He came last Monday"

□□□□□□□□/**NNT** □□□□□□/**NNT** □□□□ □□□□ □□□
"I have a holiday tomorrow on Sunday"

But,
□□ □□□□□□/**NNP**
"Today is Saturday"

□□□□□ □□□□□ □□□ □□□□□□□□□/**NNP**
"January is the first month of a year"

NNT or RB—See **RB or NNT**

ID No: 7
Part of Speech: Question Temporal Noun
Tag: QNT
Category: Noun

Example:
□□/PRP □□□**/QNT** □□□□/VB ?/.
“When is he coming?”

□□□□□/PRP+SFAC □□□/VB □□□**/QNT** □□□□/PRP □□□□□/VB
“Let me know when he comes”

Description and Analysis:
When temporal nouns are used to form interrogative sentences or to join relative clauses, they will be tagged as question temporal nouns (QNT). Thus, question and relative categories have been subsumed under a single tag.

ID No: 8
Part of Speech: Locative Noun
Tag: NNL
Category: Noun

Example:
□□□□/PRP □□□□**/NNL+SFON** □□□/VB
"You go upstairs"

□□/VOC □□□□/NN ,/, □□□□□**/NNL+SFON** □□□/VB
"Hey boy, come here!"

Description and Analysis:
Following concept and analysis has been taken from [2].
'Locative Noun' is another entirely new tag, introduced to cover an important phenomenon of Bangla language. Words like '□□□', '□□□', '□□□' are used in various ways in Bangla.

1. They act as postpositions e.g.

□□□ □□□□□/NN+SF$ □□□**/ON** □□□ □□□□□□
"I have put the pen on the book"

Here '□□□' is a postposition which is the direct equivalent to the English preposition 'on'.

2. They also act as adverbs e.g.

□□□□ □□□□**/RB** □□□
"You go upstairs"

Here '□□□' is an adverbial of place.

3. These words also take postpositions themselves and so in some sense, behave like nouns e.g.

□□ □□□**/NN** □□□/ON □□□□□
"He has come from upstairs"

4. As pointed out in 3, these words take postpositions and act as arguments of the verb in the sentence. They also take a postposition to join with another noun. So in that sense, they also behave like nouns e.g.

□□□□□**/NN+SF$** □□□/NN
"upper portion"

To tag such words, one option is to tag them according to the category to which they belong in the given sentence. For example, in 1, the word occurs as a postposition and so it can be marked as a postposition. In example 2, it is an adverb and so it can be marked as an adverb and so on.

But we feel that these words are more like nouns as is evident from 3 and 4, and also if we consider words like '□□□', '□□□', etc. as places, then we can tag them as nouns.

But these are not pure nouns as well. They are nouns which indicate a location or time. These also function as adverbs or prepositions in a context. So a new tag NNL is introduced for such words. If such words are tagged according to their syntactic function, it may decrease the automatic POS tagging (such as HMM or BRILL tagger) performance, or can hamper machine learning in any free word order language such as Bangla.

Possible confusing tags:

NNL or NNT
Some words like '□□□', '□□□' are sometimes used as temporal nouns (NNT) rather than locative nouns (NNL). Generally, just looking at the main verb we can determine the part of speech. Nevertheless, in case of distinction between confusing tags, determining the part of speech of a word prominently depends on the entire context. So one should consider the entire sentence containing the words that someone is unsure of and not just the word in isolation since the context is very important. At the example below, a vertical slash '|' is used to denote two confusing tags for the word '□□□'.

Example:
□□/PRP □□□/**NNL | NNT** □□□□□/VB
"He sat in the front" / "He sat earlier"

But,
□□/PRP □□□/**NNT+SFON** □□□□□□/VB
"He saw first"

NNL or RB—See **RB or NNL**

ID No: 9
Part of Speech: Question Locative Noun
Tag: QNL
Category: Noun

Example:
□□□□/PRP □□□□□**/QNL** □□□□□□/VB ?/.
“Where are we going?”

□□□□□□/PRP+SFAC □□□/PRP □□□□□/VB □□□□□□**/QNL** □□□□/VB □□□□□□□□/VB
“I will show you where he came”

Description and Analysis:
When locative nouns are used to form interrogative sentences or used to join relative clauses, they will be tagged as question locative nouns (QNL). Thus, question and relative categories have been subsumed under a single tag.

ID No: 10
Part of Speech: Personal Pronoun
Tag: PRP
Category: Pronoun

Example:
□□□□**/PRP** □□□□□**/PRP+SFAC** □□□□□□□/NNC □□□□/NN
□□□□□□□□□/VB
"He gave me the grammar book"

□□**/PRP** □□□□**/PRP** □□□/VBT □□□□□/NN □□□/VBT □□□□□□/VB
"He himself came and did the work"

Description and Analysis:
One of the basic parts of speech, a pronoun takes the place of a noun, often serving as a subject or as an object in a sentence. Penn tagset for English makes a distinction between personal pronouns and possessive pronouns. But here for Bangla, all pronouns are marked as PRP. In Bangla, all pronouns are inflected for all cases (accusative, dative, possessive) and numbers (singular, plural). Incase we have individual tags for each of these inflection types; the new tagset will be larger which is unnecessary. A separate type 'Suffixes' has been introduced to cover these types of inflections on nouns and pronouns, which will be described at a later section. But now, for example, when a pronoun '□□□' (I) used as an object of a verb and inflected by accusative case marker '□□' as in the above example, the inflected form becomes '□□□□□' and it is tagged using a '+' sign as PRP+SFAC (accusative suffixes).

Reflexive pronouns are also tagged as PRP. In the second sentence of the above examples, the word '□□□□' is a reflexive pronoun and has been tagged as 'PRP'.

Possible confusing tags:

PRP or QF—See **QF or PRP**

Child Tags:
The following sub-tags will not be used for tagging, but it is important to mention the subject-verb agreement in the case of pronouns. In Bangla, subject-verb agreement occurs only for person variations. These sub-categorizations are recoverable from lexical analysis.

Sub-tag	**Example**	**Description**
PRP-P1	□□□, □□□□	1st person
PRP-P2	□□□□, □□□□□	2nd person
PRP-P3	□□, □□□□	3rd person
PRP-J	□□□, □□□□	2nd person pejorative form

PRP-H	□□□□, □□□□, □□□□□	Honorific form

ID No: 11
Part of Speech: Question Personal Pronoun
Tag: QPR
Category: Pronoun

Example:
□□□□□□/PRP+SFAC □□□□/NN □□**/QPR** □□□□□□/VB ?/.
“Who gave you the book?”

□□□/DT □□□□□/NN □□**/QPR** □□□□□/PRP+SFAC □□□□/VB
“that man who knows me”

Description and Analysis:
Question pronouns are used as question words in an interrogative sentence, and can also act as relative pronouns to join a relative clause. Therefore, question and relative pronouns are subsumed under a single tag QPR. This category is similar to Wh-Pronoun (WP) in Penn tagset for English.

Possible confusing tags:

QPR or CS—See **CS or QPR**

ID No: 12
Part of Speech: Simple Adjective
Tag: JJ
Category: Adjective

Example:
□□□/DTS □□□□/QFNUM □□□**/JJ** □□/NN
"This is a good book"

□□□□□/PRP+SFAC □□□□□□**/JJ** □□□**/JJ** □□□□□□/NN □□□/VB
"Give me that beautiful red dress"

Description and Analysis:
Adjectives are words that describe or modify a noun or a pronoun, or another adjective in the sentence. Unlike adverbs, which often seem capable of popping up almost anywhere in a sentence, adjectives nearly always appear immediately before the noun or the noun phrases that they modify.

In addition to regular adjectives, hyphenated compounds that are used as modifiers are tagged as adjectives.

Ordinal numbers are also tagged as adjectives.

Possible confusing tags:

JJ or NN
Following analysis has been taken from [1].
Nouns that are used as modifiers, whether in isolation or in sequences, should be tagged as nouns (NN), rather than as adjectives (JJ).

Example:
□□□/NN □□□□ vs. □□□□/JJ □□□□
Wool vs. woolen cloth

□□□□/NNC □□□□/NNC □□□□□□□□/NN
"life insurance company"

Hyphenated modifiers, on the other hand, should always be tagged as adjectives (JJ). Thus, we have different part-of-speech assignments in examples like the following, depending on the orthographic convention used:

Example:
□□□□□-□□□□□□□/JJ □□□□□□ □□□□□/NNC
□□□□□□□/NNC □□□□□□/NN
"Anglo-American force" "Anglo-American force"

Prenominal modifiers that are gradable (i.e. they can be modified by a degree adverb or form a comparative or superlative) should be tagged as adjectives (JJ), not as nouns (NN).

Example:
□□□□ □□□/JJ □□□□
“a good story”

□□□□ □□□□/RB □□□/JJ □□□□, □□□□ □□□□ □□□□ □ □□□□□□□ □□□ □□□□□□ □□□□□

“a very good story, one of the best stories I have ever read”

Words denoting colors should be tagged as nouns (NN) when they are used as names since they have the distribution of nouns i.e. they can be modified by adjectives and they have an overt plural.

Example:
□□ □□□□□□/JJ □□□□ □□□/NN
“That beautiful red”

□□□□□□□/NN □□□ □□□□ □□□□□□□□ □□ □□□□□/NN+SF$ □□□□
“The reds won't match well with this blue”

Also note the following contrast:

□□ □□□□□□□ □□□/JJ □□□□/JJ
“These trees are dark green”

□□ □□□□□□□ □□ □□□ □□□□/NN
“These trees are a dark green”

Generic adjectives, if they are modified by adverbs, should be tagged as adjectives (JJ) and not as common nouns (NN) even when they trigger subject-verb agreement.

□□□□/JJ□□□□□/JJ □□ □□□□ □□□ □□ □□□□ □□□
“The richer in this country pay very little taxes”

JJ or NNP
Following analysis has been taken from [1].
Words that refer to languages or nations like '□□□□□' (English), '□□□□□□' (Japanese) can be either adjectives or proper nouns (NNP). In prenominal positions, such words are almost always adjectives (JJ). Do not tag such words as proper nouns just because they occur in idiomatic collocations.

Example:
□□□□□□/JJ □□□□□/NN □□□□□□□□□ □□□□□□ □□□□□□□
“Michael Modhusudan wrote English poetry”

But,
□□ □□□/RB □□□/JJ □□□□□□/NNP □□□□ □□□□
“He is very fluent in English”

Hyphenated compound proper nouns acting as modifiers such as '□□□□□□□-□□□□□□□□ □□□□□□□□□' (German-Rudman Act) as well as compounds containing proper nouns as their second constituent, such as '□□□□-□□□□□' (mid-March) should be tagged as proper nouns (NNP) rather than as adjectives (JJ).

Vexing cases arise in connection with compound adjectives that are spelled as two words such as '□□□□ □□□□□□□□□' (light smelled). Tag both parts of such a sequence as JJ – thus □□□□/JJ □□□□□□□□□/JJ.

JJ or NNC
Generally a determiner or determinative inflections determinate compound common nouns and are tagged as NNC. Otherwise, they are tagged as adjectives (JJ). But most of the times it is very hard to disambiguate. One should carefully look at the whole sentence and context instead of looking at a particular word or collocations, and can use his/her own decision from the perspective of the context.

Example:
□□/DT □□□/NNC □□□□□□□□□□/NNC □□□□□□/NN
“This law enforcement agency”

□□□/NNC □□□□□□□□□□/NNC □□□□□□□□/NN
“The law enforcement agency”

But,
□□□/JJ □□□□□□□□□□/JJ □□□□□□□□□□/NN
“law enforcement agencies”

□□□/JJ □□□□□□□□□□/JJ □□□□□□□/NN+SF$ □□□□□
“People of law enforcement agencies”

JJ or RB – See **RB or JJ**

JJ or JJV—See **JJV or JJ**

JJ or QFNUM—See **QFNUM or JJ**

Child Tags:
Plain adjectives can also be sub-categorized by degree i.e. comparative and superlative. Comparative adjectives are generally preceded by a comparative postposition i.e. □□□□, □□□□□□□ (than).

Sub-tag	**Example**	**Description**
JJ	দ্রুত	Bare Form
JJ-R	দ্রুত□□	Comparative Form
JJ-S	দ্রুত□□	Superlative Form

ID No: 13
Part of Speech: Verb Root
Tag: JJV
Category: Adjective

Example:
□□□□□□□**/JJV** □□□□□/VBF □□□□□□/NN □□□**/JJV** □□□/VBT □□□□□□/VB
“Being heated, the iron is becoming red”

□□□□□□□□□□/NN □□□□□/RB □□□□□/RB □□□□□□**/JJV** □□□/VBT □□□□/VB
“The cyclone is gradually becoming weaker”

Description and Analysis:
As discussed in the section of verb root in noun category earlier, the kriyamuls are combined with a helping or light verb to express the sense of a complete event. The kriyamuls formed by joining an adjective will be JJV. As in the first example above, '□□□' (red) is an adjective which is joined to the light verb '□□□' (been) to express the sense of the verb 'to redden'. So here '□□□' (red) is marked as JJV and the main verb is marked as VBT.

Possible confusing tags:

JJV or JJ
If an adjective co-occurs with a verb to express a complete process, it will be tagged as verb root (JJV) as well. But if it solely modifies a noun in a sentence, it will be a simple adjective (JJ).

Example:
□□ □□□□□□ □□□□□ □□□□□□**/JJV** □□□/VBT □□□□□/VB
“He has become weak due to sickness”

But,

□□ □□□□/RB □□□□□**/JJ**
“He is very weak”

ID No: 14
Part of Speech: Question Adjective
Tag: QJJ
Category: Adjective

Example:
□□□□**/QJJ** □□□□/NN □□/PRP ?/.
“What type of boy is he?”

□□/NNT □□□□□□□/NN □□□/RB □□□□□□/JJ □□/VB □□□□**/QJJ**
□□□/PRP □□□□□□□□□/VB
“Today’s weather is not as beautiful as I thought”

Description and Analysis:
Like other question nouns, question adjectives subsume those adjectives that are used for questions and relative adjectives. Adjectives that are used for question types can be used for both questions and relatives, where relative adjectives are only used for relative categories.

Possible confusing tags:

QJJ or QRB—See **QRB or QJJ**

ID No: 15
Part of Speech: Vocative
Tag: VOC
Category: Vocative

Example:
□□□**/VOC** □□□□/NNP !/. □□□□□/QW □□□□□/VB □□□□/PRP ?/.
“Hey Karim! Where are you going?”

□□□**/VOC** ,/, □□□□/PRP □□□□/VBT □□/VB
“Hey all of you! Listen here”

Description and Analysis:
Vocatives are words in Bangla used to address someone, or get attention from somebody. Vocatives are generally put just before the target noun or pronoun.

Possible confusing tags:

VOC or DT—See **DT or VOC**

ID No: 16
Part of Speech: Finite Main Verb
Tag: VB
Category: Verb

Example:
□□□/PRP □□□□□/VBF □□□**/VB**
"I am going out to play"

□□□□□/NNP □□□/NNT □□□□□□□□**/VB**
"Nafid is sleeping now"

Description and Analysis:
This is the part of speech that describes an action or occurrence or indicates a state of being and usually is the main element of a predicate-argument structure. Any finite verb that constitutes the main verb of a clause has been categorized as main finite verb, and subsumes all the sub-categorizations based on relevant tense, aspect and mood (TAM) features. Bangla doesn't have auxiliary verbs. Thus, this form of verbs related to tense information doesn't change the syntactical structure of a sentence. However, although Bangla has subject-verb agreement only for person variation, this information for morphological form of the verbs according to persons has not been included in the tagset. That is recoverable only by further lexical or morphological analysis. That's why the categorizations of verb in Penn Tagset as VB, VBD, VBG, VBN, VBP and VBZ have been subsumed as a single tag VB.

Possible confusing tags:

VB or VBC—See **VBC or VB**

Child Tags:
In case of agreement, since Bangla has subject-verb agreement only for person information, finite main verb can be sub-categorized by the following five categories:

Sub-tag	**Example**	**Description**
VB-P1	□□□, □□□□, □□□□□	1st Person
VB-P2	□□, □□□, □□□□	2nd Person
VB-P3	□□□, □□□□, □□□	3rd Person
VB-H	□□□□, □□□□□, □□□□□	Honorific
VB-J	□□□, □□□□□, □□□□	Pejorative

ID No: 17
Part of Speech: Non-finite Nominal Verb
Tag: VBM
Category: Verb

Example:
□□□□□□□□□□/NN+SF$ □□□□□□/NN □□□**/VBM** □□□□□/VBE
"The problem is being solved"

□□□□□/PRP+SF$ □□□□□/NN+SFON □□□□□□**/VBM+SF$** □□□□□/NN □□□/VB
"You don't need to go home"

□□□/DTS □□□/PRP+SFAC □□□□□□**/VBM** □□□□□/VBE
"This has been shown to him"

Description and Analysis:
This is an entirely new tag introduced to cover an important phenomenon of Bangla language. These words like '□□□' (doing), '□□□□' (seeing), '□□□□□□' (show) act like a verb when they take a nominal complement such as '□□□□□□' (solution) at the first example. On the other hand, they also act as nouns when inflected by a possessive marker. But this is apparently conceivable that these words are actually verbs because they constitute events in sentences. However, these verbs also syntactically act as nouns and they themselves can be complements of finite main verbs. Thus, a new tag Non-finite Nominal Verb (VBM) has been devised for such kind of verbs.

Possible confusing tags:

VBM or NN
Nominal Verb (VBM) takes a direct complement (usually a noun), but a noun doesn't.

Example:
□□□□□□□□ □□□□□/VBM
"The flight has been cancelled"

□□□□□□□□□□□□ □□□□□□□/VBM
"Use of firearms"

But remember that the word '□□□□□' is an adjective when it means 'abandoned' e.g.
□□□□□□□□ □□□□□/JJ
"The flight has been abandoned"

VBM or NNV

As per syntactical distribution, nominal verb roots (NNV) are followed by a verb to complete an event. In many cases, they are not followed by a verb while it takes a nominal complement to form a compound noun phrase. In that case, it will be tagged as nominal verb (VBM) and not NNV.

Example:

□□□□□□□□□□□□□/NN □□□□□□□/NNV □□□/VBM □□□□□□

"Use of firearms is illegal"

But,

□□□□□□□□□□□□□/NN □□□□□□□/VBM □□□□/ON □□□□ □□□ □□□□□□ □□□ □□□□□□

"Many laws are going to be imposed regarding use of firearms"

In the second example, the word '□□□□□□□' (use) takes a nominal complement '□□□□□□□□□□□□□' (firearms) and followed by a postposition and is tagged as nominal verb (VBM). But be aware that when they are in possessive relation with a noun, they are tagged as noun.

Example:

□□□□□□□□/NN+SF$ □□□□□□□/NN □□□ □□□□□□

"I don't know the use of the thing"

ID No: 18
Part of Speech: Non-finite Clausal Verb
Tag: VBC
Category: Verb

Example:
□□□□/PRP □□□□□/VBF □□□□**/VBC** □□□□□/CS □□□/PRP □□□/VB
"I will come after you go to play"

□□□□/PRP □□□/DTS □□□□**/VBC** □□/PRP □□□□□/VBF □□□□□/VB
"If you tell him this, he will understand"

Description and Analysis:
Some forms of verbs in Bangla introduce a subordinating clause which has its event possibilities of occurrence depending on that verb, indicating whether it has occurred or not at the preceded clause. A non-finite clausal verb is followed by a clause or a subordinating conjunction '□□□□□' (then) followed by a clause.

Possible confusing tags:

VBC or VB
A non-finite clausal verb is generally followed by a clause or a subordinating conjunction '□□□□□' (then) followed by a clause. If a coordinating conjunction except '□□□□□' (then) joins the two clauses, that verb will be tagged as finite main verb (VB) rather than VBC. Since these vexing cases are very often difficult to distinguish, one should be sure to apply any test cases to the entire sentence containing the word that s/he is unsure of and not just the word in isolation since the context is important in determining the part of speech of a word.

Example:
□□□□ □□□□□ □□□□□ □□□□/VBC □□□ □□□
"If you tell me the thing, I will go"

□□□□ □□□□□ □□□□□ □□□□/VBC □□□□□/CS □□□ □□□
"Once you tell me the thing, I will go"

But,
□□□□ □□□□□ □□□□□ □□□□/VB □□□/CC □□□ □□□
"You have told me the thing and I will go"

ID No: 19
Part of Speech: Non-finite Perfective Verb
Tag: VBT
Category: Verb

Example:
□□/PRP □□□□□□/NN □□□**/VBT** □□□□□□□□□/NN+SFON
□□□□□□/VB
“Having done his breakfast, he went to school”

□□□□/PRP □□□□/NN □□□**/VBT** □□□/VBT □□□□/VBF □□□□□/VB
“Having done office, you can go.”

Description and Analysis:
A perfective verb has a clause following it such that the action of the verb in the clause is accomplished only after action represented by the perfective verb has been completed. It is most likely to be ‘past participle’ verb in English in case of event accomplishment. A perfective verb is followed by a finite clause or by a non-finite verb or by another perfective verb.

Possible confusing tags:

VBT or RB—See **RB or VBT**

ID No: 20
Part of Speech: Non-finite Verb
Tag: VBF
Category: Verb

Example:
□□□□□/PRP+SFAC □□□□/QF □□□□**/VBF** □□□/VBE
“I have to say something”

□□/PRP □□□□□**/VBF** □□□□□**/VBF** □□□□□□/VB
“He is dancing while going”

Description and Analysis:
A simple non-finite verb can be introduced and followed by both finite and non-finite clauses. But a finite main verb must end with a clause which was left unfinished by the non-finite verb. Non-finite verbs are sometimes repetitive such as in the second example above where both of the non-finite verbs are tagged as VBF.

ID No: 21
Part of Speech: Existential Verb
Tag: VBE
Category: Verb

Example:
□□□□□□/PRP+SFAC □□□□/QF □□□□/VBF □□□**/VBE**
"You have to say something"

□□□□□/NNP □□□□□**/VBE** □□□/RB □□□/JJ □□□□/QFNUM □□□□/NN
"Nafid is a very good boy"

Description and Analysis:
Existential or 'Copula' are verbs that link the subject of a sentence with the predicate of the sentence. Although it might not express itself as an action or condition, it serves to equate or associate the subject with the predicate.

Unlike any other finite verbs, existential verbs act as 'Subject Raising Verbs' while the main subject is being inflected with accusative case marker.

For example □□□□□□ □□□□ □□□□ □□□/VBE
"You have to say something"

Here, the existential verb '□□□' (will) raise the subject '□□□□□□' (you) which has been inflected by an accusative case marker '□□'. If there is any other verb rather than the existential verb, the word '□□□□□□' (you) will be a nominal object of the verb '□□□□' (to tell) instead of a subject.

For example □□□□□□ □□□□ □□□□ □□□□/VB
"(Somebody) will go to tell you something"

But note one thing that, except for finite existential verbs (□□, □□□, □□□□□ etc.) and non-finite existential verbs (□□□), they will be tagged as VBE and VBEF respectively, other non-finite forms should be tagged according to their forms (VBT, VBC, VBM etc.).

ID No: 22
Part of Speech: Existential Nonfinite Verb
Tag: VBEF
Category: Verb

Example:
□□□□□□/PRP+SFAC □□□□/QF □□□□/NN □□□□/VBF □□□**/VBE**
□□□□/VB
"You may have to say many things"

□□□□□/PRP+SFAC □□□□/JJ □□□/JJ □□□**/VBEF** □□□/VBE
"I have to be very rich"

Description and Analysis:
Non-finite Existential verb is the non-finite form of existential verb or 'Copula'.

Like existential, non-finite existential verbs also act as 'Subject Raising', while the main subject is being inflected with accusative case marker.

Considering the first example above, if there is any other verb rather than the non-finite existential verb, the word '□□□□□□' (to you) will be a nominal object of the verb '□□□□' (to tell), instead of a subject.

For example □□□□□□ □□□□ □□□□ □□□□/VBF □□□□
"(Somebody) may go to tell you something"

But note one thing that except the finite existential verbs (□□, □□□, □□□□□ etc.) and non-finite existential verbs (□□□) tagged as VBE and VBEF respectively, other non-finite forms should be tagged according to their forms (VBT, VBC, VBM etc.).

ID No: 23
Part of Speech: Adverb
Tag: RB
Category: Adverb

Example:
□□□□□□/NN □□□**/RB** □□□□□**/RB** □□□□□/VB
"The horse runs very fast"

□□/PRP □□□□□□□/NN □□□/VB □□**/RB**
"He doesn't do study"

Description and Analysis:
An adverb can modify a verb, an adjective, a clause, a phrase or another adverb by expressing time, place, manner, degree, cause etc. Bangla adverbs often end in '□□□□'. Unlike adjectives, adverbs can popup at any place in a sentence or a clause.

Possible confusing tags:
The following have been taken from [1].

RB or NN
If the word can be modified by an adjective, it is noun. Otherwise, tag it as an adverb (RB).
Example:
□□□/NN □□□ □□□
"Play after doing work"

□□□□/RB □□□ □□□
"Play from the beginning"

RB or NNT
Locative Nouns that are used adverbially should be tagged as nouns (NNT), not as adverbs (RB).

Example:
□□ □□/NNT □□□□ □□□□□□/NNT □□□□
"He will come today or on Saturday"

RB or NNL
Words denoting the point of compass are tagged as locative noun (NNL) as well and not as adverbs.

Example:
□□□□□□ □□□□□ □□□□□ □□□□ □□□□ □□□ □□□□
□□□□□□□/NNL+SFON
"The nearest town is two miles west from here"

RB or JJ
While most adverbs formed from adjectives end in -□□□□ (ly), not all do. The crucial criterion is whether a word modifies a noun, in which case, it is an adjective (JJ), or it modifies a non-noun, in which case it is an adverb (RB).

Example:
□□□□□/JJ □□□□□/NN
"rapid growth"

□□□□□/RB □□□□/VBT □□□/VBM □□□□□
"rapidly growing tree"

Take care not to tag predicate adjectives as adverbs. Thus, in '□□□□ □□□ □□' (make life simple), '□□□' (simple) is an adjective.

RB or VBT
Put a subordinating conjunction '□□□□□' (then) after the perfective verb. If the sentence still remains grammatical, it will be a perfective verb (VBT). Otherwise, it is an adverb (RB).

Example:
□□□□ □□□□/VBT □□□□
"He has come running"

But,
□□□□ □□□□/RB □□□□
"He has come chasing"

RB or CC—See **CC or RB**

RB or CN—See **CN or RB**

RB or RP—See **RP or RB**

Child Tags:
Bangla does have two prominent sub-categorizations for adverb as other languages do.

Sub-tag	**Example**	**Description**

RB	□□□□□, □□□□□	Plain
RB-N	□□, □□□	Negative

ID No: 24
Part of Speech: Question Adverb
Tag: QRB
Category: Adverb

Example:
□□□□/PRP □□/NNT □□□□□□/NN+SFON □□□□□/VB □□□**/QRB** ?/.
"Why didn't you go to school today?"

□□□□□/NNP □□□□□□**/QRB** □□□□□/NN □□□□□/VB ?/.
"How did Nafid solve the math?"

□□□□□/NN □□□□/VBE □□□/PRP □□□□□□**/QRB** □□□□□□□□/VB
"The work was not done as I had told"

Description and Analysis:
Question adverbs and relative adverbs are subsumed into a tag QRB. Relative adverbs join relative clauses. Unlike relative adverbs, question adverbs can be used as both question words and relative adverbs.

Possible confusing tags:

QRB or QJJ
When distinguishing between these two types, it has to be observed whether that word modifies a noun or a verb. If it modifies a noun, it will be tagged as question adjective (QJJ). Otherwise, it will be question adverb (QRB).

Example:
□□□□/QJJ □□□□□ □□ ?
"What type of man is he?"

□□□ □□□□ □□□ □□□□/QJJ □□□
"I know what type I want"

But,
□□□□/QRB □□ □□□□ ?
"How are you?"

ID No: 25
Part of Speech: Coordinating Conjunction
Tag: CC
Category: Conjunction

Example:
□□□□□/NNP □□/NN □□□□/VB □□□**/CC** □□□□/NN □□□□□/VB
"Nafid is reading book and watching sports"

□□□□/PRP □□□□□□/RB □□□□/VB □□□□□**/CC** □□□□/PRP+SFAC
□□□□/VBF □□□/VBE
"You must go otherwise, he has to come"

Description and Analysis:
An uninflected word used to connect words, phrases, clauses, or sentences; connective: conjunctions may be coordinating. Coordinating conjunctions subsumes all other types of conjunction such as adversative, disjunctive, exclusion, conclusive etc. as Penn Tagset for English.

Possible confusing tags:

CC or RB
The words '□□', '□□□□' are usually used as conjunctions (CC) when they join two clauses, but in some cases they modify verbs to express continuity of actions. In this case, these are tagged as adverbs (RB). It is often very difficult to decide which one is which. Again, it is worth mentioning that in case of distinguishing parts of speech; consider the whole sentence that contains it, not the specific words or collocation. Since, context is a vital issue for deciding relative parts-of-speech.

Example:
□□□□□□ □□□□□□□ □□□□□ □□/CC □□□□ □□□□□□
"The boy has done his study and also has drawn picture"

□□□□ □□□ □□□□ □□□□/CC □□□□□□□□ □□□□?
"You have practiced math and again, have read the poem?"

But,
□□□ □□/RB □□□□□
"I won't come any more"

□□□□□□□ □□□□□□, □□□□/RB □□
"Thank you, come again"

□□ □□□□□ □□□□□□, □□□□/RB □□□□
"He has gone home and will come again"

CC or RP—See **RP or CC**

ID No: 26
Part of Speech: Compound Coordinating Conjunction
Tag: CCC
Category: Conjunction

Example:
□□□□□/NNP □□/NN □□□□/VB □□**/CCC** □□/CC □□□□/NN □□□□□/VB
"Nafid is either reading book or watching sports"

□□□/PRP □□□□/PRP+SFAC □□□□□□/VB □□□**/CCC** □□/CC □□/PRP □□□□□/VB
"I called him yet he didn't come"

Description and Analysis:
This is another new tagset introduced for coordinating conjunction (CC) with same categorization as compound common noun (NNC) and compound proper noun (NNPC) as discussed for nouns. Some words in Bangla are compounded to act as a single unit of coordinating conjunction (CC). All words except the last one will be marked as CCC. Thus any CCC will be followed by another CCC or a CC. This strategy will help to recognize those words as a single unit of conjunction.

ID No: 27
Part of Speech: Suspicion Conjunction
Tag: CN
Category: Conjunction

Example:
□□□/PRP □□□**/CN** □□□/VB □□/PRP □□□□/RB □□□□/VB
"If I go, then he may come"

□□□□□/VBF □□□□□□/VB □□□/NN ,/, □□□/RB □□/NN □□□/RB
□□□/NN □□□□**/CN** □□□□/NN □□□□/NN □□□/VB
"I can't do work, always fear and feel shy wondering people may say something"

Description and Analysis:
Suspicion conjunctions are conjunctions which are used to express a suspected action that may happen or not. Like an adverb, they can pop up anywhere in the clause in which they are used.

Suspicion conjunctions join subordinating clauses, thus acting as subordinating conjunctions.

Suspicion conjunctions also act as eternal joining conjunctions along with coordinating conjunctions.

Possible confusing tags:

CN or RB
In a separate single clause, they are usually tagged as adverb (RB), not as CN.

Example:
□□□/RB □□ □□□ □□□?
"What if he falls down?"

CN or CET—See **CET or CN**

ID No: 28
Part of Speech: Eternal Joining Conjunction
Tag: CET
Category: Conjunction

Example:
□□□**/CET** □□□□□□□□/NN □□□/VB □□□**/CET** □□□□/NN □□□□/NNV □□□/VB
"The patient died when doctor came"

□□□**/CET** □□/PRP □□□□/NN □□□□/ON □□□/NNV □□□/VBE □□□□**/CET** □□□□□□/NN □□□□□/VB
"It started raining just when he came out of home"

Description and Analysis:
When two clauses are joined by double coordinating conjunctions, they are called eternal joined conjunctions.

Eternal joined conjunctions are mostly paired by using a relative word.

Possible confusing tags:

CET or CN
Suspicion conjunctions (CN) that act as eternal joining conjunctions will be tagged as CET, not CN.

Example:
□□□ □□□/CET □□□ □□□□□/CET □□ □□□□
"If I ask, he will come"

□□□□/CET □□□ □□□□ □□□□ □□□□□□/CET □□ □□□□□ □□□ □□□□
"He is talking softly fearing someone may overhear"

CET or DT—See **DT or CET**

ID No: 29
Part of Speech: Subordinating Conjunction
Tag: CS
Category: Conjunction

Example:
□□□□□/NNP □□□□□/PRP+SFAC □□□□□□□/NNV □□□/VB □□**/CS**
□□□/PRP □□□/VBT □□□□□□/VB □□□□/RP
"Rahman asked me whether I was going or not"

□□□□□/NNP □□□□/VBC □□□**/CS** □□□□□/PRP+SFAC □□□/VB
"Let me know when Nafid comes"

Description and Analysis:
A subordinating conjunction usually introduces subordinate clause. A subordinate clause depends on the rest of the sentence for its meaning. It does not express a complete thought so it does not stand alone. It must always be attached to a main clause that completes the meaning.

Subordinating clauses are attached by subordinating conjunctions, suspicion conjunctions or postpositions.

Possible confusing tags:

CS or QPR
When '□□' introduces complements of nouns, it is a subordinating conjunction (CS).

Example:
□□□ □□□□ □□□ □□ □□/CS □□□□□□ □□□□ □□□
"It is claimed that fairies have wings"

But when '□□' introduces relative clauses, it is a question pronoun (QPR).
Example:
□□□□□□□ □□/QPR □□□□□ □□□□
"The man who knows me"

ID No: 30
Part of Speech: Compound Subordinating Conjunction
Tag: CSC
Category: Conjunction

Example:
□□□/PRP □□□□/PRP+SFAC □□□□/QF □□□□□/VB □□**/CSC** □□□□□/CS □□/PRP □□□□□/VB
"He hasn't come because I haven't told him anything"

□□/NNT □□□□□□/NNV □□□□□/VB □□□**/CSC** □□□/CS □□□□□□/NN+SFON □□□□□/VB
"I won't go to school today since it's raining"

Description and Analysis:
Same as compound coordinating conjunction, subordinating conjunctions are also formed by more than one word such as '□□ □□□□□' (for this reason) or '□□□ □□□' (that's why). So, in that case, all words except the last one will be marked as CSC. Thus, any CSC will be followed by another CSC or a CS. This strategy will help to recognize those words as a single unit of subordinating conjunction.

ID No: 31
Part of Speech: Postposition
Tag: ON
Category: Postposition

Example:
□□□/PRP □□□□□/NN+SF$ □□□□**/ON** □□□□/NNP □□□□□□□□□/VB
"I went to Dhaka with father"

□□□□□/PRP+SF$ □□□□□□**/ON** □/DT □□□/NN □□□/VB □□/RB
"This work cannot be done by you"

□□□□□/NNP □□□□**/ON** □□□□/NNP □□□□□/JJ
"Oveek is taller than Nafid"

Description and Analysis:
A word that indicates the relationship between a noun or a pronoun, or a syntactic construction to another element in a sentence is called post-position or postposition in Bangla. Unlike preposition in English that precedes a clause, in case of Bangla, postpositions usually follow nouns, pronouns or clauses. That's why, its named as 'Postposition'.

Possible confusing tags:

ON or NNT
The words '□□□' (at before), '□□□' (at after) will be tagged as NNT+SFON. The uninflected word '□□' (then) will be tagged as postposition (ON), but remember, when they are followed by another postposition, they will be tagged as Temporal Noun (NNT).

Example:
□□□/NNT □□□□/ON

□□□/NNT □□□□/ON

□□/NNT □□□□/ON

ID No: 32
Part of Speech: Interjection
Tag: UH
Category: Interjection

Example:
□□□□**/UH** !/. □□/RB □□□□□□/JJ □□□□□/NN !/.
“Wow! What a beautiful scenery!”

□□□**/UH** !/. □□□□/PRP □□□/QW □□□□/VB ?/.
“Alas! What have you done?”

Description and Analysis:
Interjections are words used to express strong feelings or sudden emotions. They are included in a sentence usually at the start to express a sentiment or a strong emotion such as surprise, disgust, joy, excitement or enthusiasm.

They generally come at the start of a sentence followed by an exclamation point or by a comma, if the feeling's not as strong.

ID No: 33
Part of Speech: Particle
Tag: RP
Category: Particle

Example:
□□□□□□/NN □□**/RP** □□□□/NNT □□□□□/VB
“The boy has not come yet”

□□□/PRP □□□□□/CC □□□□□□/VB □□□□□/CC □□□□**/RP** □□□□/PRP □□□□□/VB
“I don’t know as much as you do”

Description and Analysis:
It is an uninflected item that has grammatical function but does not clearly belong to one of the major parts of speech.

Question and negative particles are also merged into this category.

A particle doesn’t alter the syntactic or semantic distribution of a sentence.

Possible confusing tags:

RP or RB
The distinction between adverb (RB) and particle (RP) is often difficult to make. There are a number of test cases that can be applied to distinguish. But, be sure to apply these tests to the entire sentence containing the word that tends to be confusing, and not just the word in isolation since the context is important in determining the part of speech of a word.

A word is a particle if it bears a stress at any clausal position but doesn't alter the semantic interpretation of a sentence. On other hand, a word is an adverb if it modifies a verb and contributes to the sentence's final semantic interpretation. Moreover, it has same syntactic distribution as other adverbs (generally, appears just before or after the verb that it modifies).

Example:
□□□□ □□/RP □□□□□ □□
“Don’t you tell me”

But,
□□□□ □□□□□ □□ □□/RB
“You don't tell me”

But, be aware that, sometimes '□□' appears to express a request rather than modifying a verb although it's syntactic distribution is same as adverb. In that case, it is a particle.

Example:
□□□□ □□□□□ □□ □□/RP
“Please, tell me”

RP or CC
Particles can also be used as coordinating conjunctions.

Example:
□□□ □□□□ □□/CC □□ □□□□□?
“So what if it’s not done?”

□□ □□ □□□□□ □□/CC □□□□□ ?
“Has he told it or not?”

QRP or QDT—See **QDT or QRP**

ID No: 34
Part of Speech: Question Particle
Tag: QRP
Category: Particle

Example:
□□□□/PRP □□□□□□□□□/NN □□□□/VB □□**/QRP** ?/.
“Have you read the novel?”

□□□□/PRP □□**/QRP** □□□□□/PRP+SFAC □□□□□/VB ?/.
“Do you know me?”

Description and Analysis:
This type has been included mostly in grammatical aspects rather than for pure syntactic distribution. But, the intuition is that looking at the syntactical construction of a sentence, it can be known that the sentence is an interrogative sentence (basic yes/no question) rather than assertive.

Possible confusing tags:

QRP or QDT—See **QDT or QRP**

ID No: 35
Part of Speech: Simple Determiner
Tag: DT
Category: Determiner

Example:
□□**/DT** □□□□□□/NN □□□□/PRP+SF$
“These books are mine”

□□□**/DT** □□□□/NN+SFON □□□□□□/NN □□□□□/VB □□/RB
“I haven’t found the writing in any book”

Description and Analysis:
A word that precedes a noun to determine an object is called a determiner. Determiners can exist syntactically without a head noun.

Simple determiners allow the head noun to be inflected by any suffixes.

Possible confusing tags:

DT or VOC
Determiners are sometimes used as vocatives to call or address someone. So in that case, tag it as VOC, not DT.

Example:
□□/VOC □□□□□, □□□□□□ □□□
“Hey Nafid, go to school”

DT or CET
The following contrasts have been taken from [1].
When there are first members of the double conjunctions "□□□...□□□" and "□□□□□...□□□□", '□□□□□' (either) and '□□□□' (or) are tagged as coordinating conjunctions (CET) not as determiners (DT)

Example:
□□□□□/DT □□□□ □□□□□ □□□□
“Any of the boys could sing”

But,
□□□□□/CET □□□□ □□□□ □□□□/CET □□□□ □□□□□□ □□□□□ □□□□□
“Either a boy or a girl could sing”

Be aware that '□□□□□' can sometimes function as a determiner (DT) even in the presence of '□□□□' or '□□□□□'.

Example:
□□□□□/DT □□□□ □□□□/CC □□□□□□ □□□□□ □□□□□
“Any of the boys or girls could sing”

DT or QF—See **QF or DT**

DT or QFNUM—See **QFNUM or DT**

ID No: 36
Part of Speech: Single Determiner
Tag: DTS
Category: Determiner

Example:
□□□**/DTS** □□/NN
"This is a ball"

□□□**/DTS** □□□□□/NNL+SFON □□□□/VB □□/RB
"You won't find that here"

□□/QPR □□□**/DTS** ?/.
"Who is that?"

Description and Analysis:
Single determiners precede nouns and determinate objects those are singular in number and cannot be inflected by any suffix. If the noun is inflected, it will change syntactic distribution of the sentence.

Example:
□□□/DTS □□□/NN
"This is a pen"

Generally common nouns are preceded by singular determiners.

Singular determiners can syntactically pop up without the presence of a head noun and can also appear after the head noun.

Possible confusing tags:

DT or NN
When determiners are used pronominally, i.e. without a head noun, they should still be tagged as determiners (DT) and not as common nouns (NN).

Example:
□□□ □□□/DTS □□□ □□□□□ □□□□□□□
"I can't keep this holding"

□□□□□/DT □□□□/NN □□□□ □□□
"Any one will do"

ID No: 37
Part of Speech: Question Determiner
Tag: QDT
Category: Determiner

Example:
□□□/PRP □□□□/VB □□**/QDT** □□□/VB □□□□/PRP
"I know what you want"

□□□□□/PRP+SFAC □□□□□/NN □□/VB □□□□**/QDT** □□□□/PRP □□□□□/VB
"Tell me what you have heard"

Description and Analysis:
Question determiners are used for interrogative sentences and to join relative clauses. Relative determiners and question determiners are subsumed under a single category which is question determiners (QDT).

Possible confusing tags:

RP or QDT
Question particle is used to construct yes-no questions. On the other hand, question determiner determinates a nominal object in an interrogative sentence. Since they have similar syntactic distribution, one has to consider the entire sentence to disambiguate between these two tags.

When '□□' is used as question particle--
□□ □□/RP □□□□□□□□?
"Did he see?"

When '□□' is used as question determiner--
□□ □□/QDT □□□□□□□□?
"What did he see?"

ID No: 38
Part of Speech: Quantifier
Tag: QF
Category: Quantifier

Example:
□□□□□/NNP □□□□**/QF** □□□□/NN □□□□□□/VB
"Nafid has seen some birds"

□□□/PRP □□□□□□/RP □□□□□□□/PRP+ICAC □□**/QF** □□□/NN □□□□□/VB
"I haven't told them everything"

Description and Analysis:
All words that quantify the nominal objects are called Quantifiers and will be tagged as QF.

Quantifiers may occur without a head noun, and can be preceded by a determiner.

Possible confusing tags:

QF or NN
In some constructions, quantifiers are used as nouns. In such cases, they will be tagged as nouns (NN) -

□□□□□/PRP+SF$ □□□□□/NN □□□□□□ □□ □□□□□□□□□□□□ □□□□□
"Many of them are going and being refused"

QF or PRP
The word '□□□□' (all) should be tagged as personal pronoun and not as quantifier (QF) e.g.

□□ □□□□/PRP
"Let us all go"

□□□□□ □□□□/PRP □□□□□ □□□
"All of you come here"

QF or JJ
Quantifiers can also be quantified by another quantifier; those quantifiers should not be tagged as adjectives e.g.

□□□□□□ □□□/QF □□□□/QF □□□□□
"You are getting much more"

□□□□/QF □□□□/QF □□□□/NN □□□□ □□

“Bring a few papers”

ID No: 39
Part of Speech: Quantifier Number
Tag: QFNUM
Category: Number

Example:
□□**/QFNUM** □□□/NN □□□□□□□/VB
"Someone came"

□□/DT □□□□□/NN+SF$ □□□□□/NN □□□**/QFNUM** □□□□/NN
"The price of this book is 200 taka"

Description and Analysis:
Some words that denote numerical numbers are tagged as quantifier numbers. Number-number combinations are also tagged as quantifier numbers if they don't have same distributions as adjectives.

Possible confusing tags:

QFNUM or NN
In some constructions, quantifier numbers are used as nouns. In such cases, they will be tagged as nouns (NN).

□□□□□ □□□□□□/NN □□□□ □□□□/NN □□□□
"One of every five persons will come"

But remember that when there is an expression for counting events where quantifiers appear pronominally, they will be tagged as quantifiers as well and not as nouns. The nominal phrase that consists of quantifiers is followed by a postposition.

Example:
□□□□/QFNUM □□□□/QFNUM □□□/ON
"One by one"

□□□□/QFNUM □□□□/QFNUM □□□□□□/ON
"by groups of two"

Note that, here the word '□□□' is tagged as postposition (ON), which is usually a verb in perfective form (VBT). Please see discussion for the word '□□□' under **specific words and collocation section**.

QFNUM or JJ
Number-number combinations should be tagged as adjectives (JJ) if they have the same distributions as adjectives.

Example:
□□□ □-□□/JJ □□□□□□□□□□ □□□□□
"That 9-11 terrorist attack"

When quantifier numbers are used as ordinal numbers, they will be tagged as adjectives (JJ)

Example:
□□ □/JJ □□□□□/NNT □□□ □□□□ □□□□□□ □□□□□□□
"the 5th of July was my brother's birthday"

QFNUM or DT
If quantifier numbers are used as counterparts of other determiners such as '□□□□' (another), '□□□□' (another one), etc, they will be tagged as determiners (DT).

Example:
□ □□□□□ □□/DT □□□, □□□□/DT □□□ □□□□□ □□ □□□□
"This is one side and the other is here"

□□/DT □□□ □□□, □□□□/DT □□□ □□
"One brother is younger and another is elder"

ID No: 40
Part of Speech: Question Quantifier
Tag: QQF
Category: Quantifier

Example:
□□/DT □□□□□□□/NN+SF$ □□□□□/NN □□**/QQF** ?/.
“What is the price of this dress?”

□□□/PRP □□□□□□/VB □□□□/PRP □□□□□□□/PRP+ICAC
□□□□□□**/QQF** □□□□/VB
“I have heard as much as you have told them”

Description and Analysis:
Question quantifiers are used to form interrogative sentences or to join relative clauses. Thus, question words and relative categories for quantifiers have been subsumed under a single tag (QQF).

ID No: 41
Part of Speech: Foreign Word
Tag: FW
Category: Foreign Word

Example:
‘/` Bengali/FW ’/’ □□□□/RB □□□□/QFNUM □□□□□□/JJ □□□□/NN
“’Bengali’ is actually a foreign word”

□□□□□/NNT □□□□□□□□□□□□/NNP+SF$ □□□□□/JJ
□□□□□□□□/NN+SF$ □□□□□□□□□/NN+SFON □□□□/NN □□□/VBE
International/FW Mother/FW Language/FW Day/FW
“Yesterday, the headline of the first page of Wikipedia was ‘International Mother Language Day”

Description and Analysis:
The words taken from another language used in Bangla and not used by native speakers are tagged as Foreign Word (FW). One should use his/her judgment to identify what is a foreign word.

ID No: 42
Part of Speech: Symbol
Tag: SYM
Category: Symbol

Example:
x ∞ yz**/SYM** ,**/**, □□□□□□**/**NNL+SFON **y/SYM** □□□**/**CC **z/SYM** □□□□**/**DT □□□**/**JJ
"x ∞ yz, where both y and z are variables"

θ =90°/SYM □□□**/**VBC **sin θ =/SYM** □□/QW □□□**/**VB **?/**.
"If θ =90° then sin θ = what?"

□□□□**/**PRP+SF$ □□□□□**/**NN+SF$ □□□□□□**/**NN □□□□□**/**RB □□,□□□/QFNUM □**/SYM**
"My expenditure is around 17,000 taka"

Description and Analysis:
Following contrast has been taken from [1].

This tag should be used for mathematical, scientific and technical symbols or expressions that aren't words in Bangla. Sign of currencies like □ (Taka) and $ (Dollar) are also tagged as symbols (SYM).

It should not used for any and all technical expressions. For instance, the names of chemicals, units of measurements (including abbreviations), etc should be tagged as nouns.

ID No: 43
Part of Speech: List Item Marker
Tag: LS
Category: List Item Marker

Example:
□□□□□□□/NNP+SF$ □□□/NN+SFON □□□□□□/NN □□□/VBE –/-
□**/LS ./.** □/QFNUM □□□/NN □□□□□□□□/NN ,/, □□□/CC
□**/LS ./.** □□/QFNUM □□□/NN □□□□□□□/NN
"According to Edison, success is
1. 1 percentage of inspiration, and
2. 99 percentages of perspiration"

□□□□□□□/RB □□□□/QFNUM □□□□□□/NN+SFON □□□□□/NNPC
□□□□/NNP □□□□□□□/VBM -/-
□/LS ./. □□□□□□□□/NNP
□/LS ./. □□□□□□□□□□/NNP ,/, □□□□/NNP
"Bangla is used mainly in two regions –
1. Bangladesh
2. West-Bengal, India"

Description and Analysis:
Following description has been taken from [1].

This category includes letters and numerals used to identify items in a list.

ID No: 44
Part of Speech: Postpositional Suffix
Tag: SFON
Category: Suffix

Example:
□□□□□□/NN □□□□□□□**/NN+SFON** □□□□□□□□/VB
“The child is sleeping on the bed”

□□□**/DT+SFON** □□/QW □□□□□□/NN □□□□/VB ?/.
“Will this be enough for all?”

Description and Analysis:
Bangla is comparatively morphologically rich language. A suffix is added with a noun or pronoun to change its some features like number and case. Since these suffixes inflect a nominal object and convert it to have different features and syntactical categories, these can be assigned to a single POS category which is 'Suffixes'. The motivation behind this is that if a different tag is introduced for each inflected nominal object, the size of the tagset will be large. Rather, it will be a decision to concise the tagset by having another POS category for suffixes and putting a '+' sign when a noun or pronoun is inflected. This will make a compact tagset and information can be easily retrieved from the assigned tag.

Postpositions are attached as suffixes with nouns or pronouns to form phrases that are categorically and structurally equivalent to postpositional phrases. These postpositional phrases have same syntactic distribution as those which are constructed with a normal postposition placed after noun.

Possible confusing tags:

NNP+SFON or NNP
In Bangla, sometimes the subject of a verb is inflected with a postpositional marker (usually a common noun or a proper noun) to denote that the subject is of ‘instrumental’ case form. Only in this case, the nominal subject should be simply tagged as NNP not NNP+SFON.

Example:
□□□□□□/NNP □□ □□□□□

ID No: 45
Part of Speech: Accusative Suffix
Tag: SFAC
Category: Suffix

Example:

□□□/PRP □□□□□□**/PRP+SFAC** □□□□□□□/NN □□□□□□/VB
“I have given you the bag”

□□□□□□□□□□**/NN+SFAC** □□□□/QF □□□□/NN □□□/VB
“Give some money to the beggars”

Description and Analysis:
Accusative inflectors are added as suffixes to denote the case of a noun or pronoun as either accusative or dative. Its corresponding tag is SFAC. An accusative nominal object is the main direct object of a transitive verb.

Some inflectors also denote the case of a nominal object as accusative (or dative) while they also denote the number to be plural, such as ‘□□□’. For simplicity, this suffix has been included only into accusative inflector category, omitting the number information.

Possible confusing tags:

SFAC or SF$**-- See **SF$ or SFAC

ID No: 46
Part of Speech: Possessive Suffix
Tag: SF$
Category: Suffix

Example:
□□□□□□**/NNP+SF$** □□□□□/NN □□□/PRP □□□□□□/VB
"I have taken Karim's pen"

□□□□□□□/NN □□□/VBT □□□□□**/PRP+SF$** □□□□□/NN
□□□□□/PRP+SFAC □□□□/VB
"Please, tell me your name"

Description and Analysis:
According to linguistics, possessive suffix is a suffix attached to a noun to indicate its possessor, much in the manner of possessive adjectives. Since Bangla is referred to be in the Indo-Aryan language family, it also supports possessive suffix.

Possible confusing tags:

SF$ or SFAC
Sometimes, it is not possible to identify the case by only looking at the word or collocation. It is very important to analyze the whole sentence that contains it and its context in order to disambiguate the case. Context is the only factor to consider for distinguishing between these two tags. For example, the inflector '□□□' can be both an accusative or dative marker and a possessive marker.

Example:
□□□□□□□□□□/NN+SFAC □□□□□□□□ □□□□ □□□
"Give the money to the beggars"

But,
□□□□□□/PRP+SF$ □□□□□□□□ □□□□ □□□
"Give us our money"

ID No: 47
Part of Speech: **।, ?, ! (**Sentence Final Punctuation Mark)
Tag: .
Category: Punctuation Mark

Example:
□□□/PRP □□□□□□/VB **।** **/ .**
“I am leaving”

□□□ **!/.** □□□/CN □□□□/VBF □□□□□□/VB **!/.**
“Uh! I wish I could go”

□□□□/PRP □□/QDT □□□/VB **?/.**
“What do you want?”

Description and Analysis:
Sentence final punctuation marks are the punctuations commonly placed at the end of several different types of sentences in Bangla and many other languages.

Sentence final punctuation marks consist of . (full stop), ! (exclamation sign) and ? (question mark) which are placed at the end of a sentence or a text, to denote the end of a sentence or an expression.

ID No: 48
Part of Speech: , (Comma)
Tag: ,
Category: Punctuation Mark

Example:
□□□□/PRP □□□□□□/VB **,/,** □□□/PRP □□□□/VB
"Wait, I am coming."

□□□/PRP **,/,** □□□□/PRP **,/,** □□/PRP □□/CC □□□□□/NNP
□□□□□□□□/NN+SF$ □□□□/ON □□□□/VBF □□□/VB
"You, he, Nafid and I will go to the teacher to study"

Description and Analysis:
Comma is a mid-separator and is used principally for separating objects and clauses in many contexts.

ID No: 49
Part of Speech: ; (Semicolon), : (Colon)
Tag: :
Category: Punctuation Mark

Example:
□□□□/NNP □□□□□/VB □□/QFNUM □□□□□□/NN **;/:** □□□□□□/NNP ,/, □□/QFNUM **;/:** □□□/CC □□□□/NNP, □□/QFNUM □/.
“Sumee has scored 70 points; Rehana, 60; and Deepa, 40.”

□□□□□□□□□□/NNP+SF$ □□□□□□□□□/NN □**/:** □□/QFNUM □□□□□□□□□/JJ □□□□□□□□□/NN+SF$ □□□/NN
“The Independence of Bangladesh: An outcome of a gory struggle”

Description and Analysis:
Colons and semicolons are kinds of punctuation marks which have distinct set of contributions of their own within a context. But here, they have been subsumed to have same syntactical category. The reason behind this is that they actually tend to bind two sentences better than they would be if separated by a full stop.

ID No: 50
Part of Speech: - (Dash), -- (Double Dash)
Tag: -
Category: Punctuation Mark

Example:
□□□□□□/NN □□□□□□□□□□/JJ :/: □□/QFNUM **-/-** □□/QFNUM
“Page reference: 11-28”

□□□□□□□/NNPC □□□□□□□□□□/NNP (/(□□□□/QFNUM **-/-** □□□□/ QFNUM)/)
2nd World-War (1939 – 1945)
“World War II (1939–1945)”

Description and Analysis:
Dashes (-) or Double Dashes (--) are widely used in different languages to illustrate a relationships or connections between two things, or to denote a range of values. Both of the categories are subsumed using a single tag (-).

When Dashes are used to define only ranges, they are explicitly tagged. In any other cases, they act as connectors like compound adjectives.
For **example:**
□□□□-□□□□□/JJ □□□□□□□/NN
“Father-son relationship”

But Double Dashes (--) are most likely to be tagged explicitly because they are used to join sentences, indicating parenthetical thoughts or some similar interpolations.

ID No: 51
Part of Speech: ((Left Parenthesis), { (Left Curly Brace), [(Left Square Bracket)
Tag: (
Category: Punctuation Mark

Example:
□□□□□□□/JJ □□□□□/NNPC □□□□□□□□□/NNPC **(/(**
□□□□□□□□/NNT+SFON □□□□□□□□/NNP)/) □□□□□/NNP
“Former East Pakistan (Bangladesh at present) Government”

□□□/DT □□□□□□□/JJ □□□/NN **(/(** □□□□/ QFNUM -/- □□□□/ QFNUM)/)
“That British period (1757 – 1947)”

Description and Analysis:
Brackets are punctuation marks used in pairs to set apart one interjecting text within another text. Left brackets are the left or the starting parts of that pair.

In this tagset, all kinds of left brackets are tagged using left parenthesis [(].

ID No: 52
Part of Speech:) (Right Parenthesis), } (Right Curly Brace),] (Right Square Bracket)
Tag:)
Category: Punctuation Mark

Example:
□□□□/NN+SF$ □□□/NN □□□□□/DT □□□□□□□□/NN+SFON □□,□□,□□,□□□/ QFNUM □□□□□/NN (/(□□□□□□/NN+SFON □,□□,□□,□□,□□□/ QFNUM □□□□□□□□□/NN **)/)**
"The speed of light is 29,97,92,458 meters per second (1,079,252,849 kilometers per hour)"

□□□□□□□□□□□/NNP (/(Linguistics/FW in/FW English/FW **)/)** □□□□/DT □□□□/JJ □□□□□□□/NN
"Linguistics is a difficult field"

Description and Analysis:
Right brackets are the right or the ending parts of the bracket pair.

In this tagset, all kinds of right brackets are tagged using right parenthesis [)].

ID No: 53
Part of Speech: ‘, “ (Opening Left Quotes)
Tag: LQ
Category: Punctuation Mark

Example:
‘/LQ □□□□□/NNP ’/RQ □□□□□□/NN □□□□□/VB □□□□□□/JJ □□□□/NN □□□□/ON
“The word ‘chair’ has come from the English language”

□□/PRP □□□/VB ,/, **“/LQ** □□/NNT □□□□□□/NN □□□/VBF □□□□/VB ”/RQ
“He said, “It may rain today” ”

Description and Analysis:
Quotation marks or inverted commas (occasionally speech marks) are punctuation marks used in pairs to set off speech, a quotation, a phrase or a word. The pair consists of an opening quotation mark and a closing quotation mark.

Opening quotation marks are tagged as LQ.

ID No: 54
Part of Speech: ’, ” (Closing Right Quotes)
Tag: RQ
Category: Punctuation Mark

Example:
‘/LQ □/NNP **’/RQ** □□□□□□□□/NN □□/QRP □□□□□/JJ ?/.
“Is the number ‘2’ a prime?”

“/LQ □□□□/CET □□□□/NN ,/, □□□□/CET □□/NN **”/RQ** □□□□/QFNUM □□□□□□□/NN
“You reap as you sow” is a phrase”

Description and Analysis:
Closing quotes are the right or ending quotes of quotation pais. Both single and double quotation marks will be tagged as RQ.

Specific words and collocations

□□ **:** It is a question determiner (QDT), but it also acts as a coordinating conjunction to join two verbal phrases.

Example:

□□□ □□□ □□□□/VB □□/CC □□/RB □□□□□□/VB

"I don't know whether I will get it or not"

Sometimes it is used as a question adjective (QJJ) to describe or modify a situation.

Example:

□□/QJJ □□□□□□□□□□□ □□□ □□□ ?

"In what situation, did this happen?"

□□□ : The word '□□□' is a postposition when it is preceded by an expression that used to count and has same distribution as postposition.

Example:

□□'□□/QFNUM □□□/ON □□□□

"Go two persons at a time"

□□□□/QFNUM □□□□/QFNUM □□□/ON □□□□ □□□

"We will go one by one to see"

□□□□ : It is tagged as a possessive inflected personal pronoun (PRP+SF$).

□□□ : It is a personal pronoun (PRP) when used in a simple assertive sentence.

Example:

□□□/PRP □□□□□

But, it will be tagged as a question pronoun when used in interrogative sentences.

Example:

□□□/QPR □□□□□□□?

□□□□□□□□ : Generally a common noun, denoting a process or event. But sometimes it also pops up as a postposition (ON), because in that case, it has the same distribution as other postpositions.

Example:

□□□ □□□□□□□□/NN □□□ □□□□□

"They are copying each other's script"

But,

□□□□□ □□□□□□□□/ON □□□ □□□□□□

"Following you, he has gone too"

□□□ : Although it's a simple verb (VB), it has also different syntactic distributions. For example, it is a postposition (ON) when preceded by a nominal phrase.

Example:
□□□ □□□□ □□□/ON □□□ □□□□□
"That seems like a bird"

It is a subordinating conjunction (CS) when preceded by a clause.
Example:
□□□□□□ □□□□ □□□/CS □□□ □□□□□
"The bird seems to be flying"

□□□□□ : It is a locative noun (NNL) and tagged as NNL+SFON when referring to a location or a place.
Example:
□□□□□□□□□□ □□□□□/NNL+SFON □□□□ □□□□ □□□
"There are some papers in the packet"

But when it acts like a postposition (ON) as when following a nominal phrase (mostly headed by a quantifier), it will be tagged simply as ON, not as NNL+SFON.
Example:
□□□□□□/QFNUM □□□□□/NN+SF$ □□□□□/ON □□□□□□ □□□□□□□ □□
"Three of the five books are grammar books"

□□□□ : It is tagged as QDT+SFON when referred by a postpositional phrase such as '□ □□□□' (for this), '□□ □□□□' (for that) etc.
Example:
□□ □□□□/QDT+SFON □□□□ □□□□ □/DT □□□□/ON □□□ □□□□ □□□□ □□□□□□
"I have opened the door so that he can come in"

It will be tagged as eternal joining conjunction (CET) when it co-occurs with '□□□□□', '□□□□□□' etc.
Example:
□□ □□□□/CET □□□□ □□□□ □□□□□/CET □□□ □□□□ □□□□ □□□□□□
"I have opened the door so that he can come in"

□□□ : Generally, it is a common noun, but when it contains the occurring time of an event or action, it is a temporal noun (NNT).
Example:
□□□/NN □□□/VBT □□□
"Come with time in hand"

But,

□□□□/VBM+SF$ □□□/NNT □□□ □□□□ □□
“Don’t talk while studying”

□□/DT □□□/NNT □□□□ □□□ □□□□□□□
“At that time, a man came”

□□□□□ □□□□□ : This kind of repetitive words will be tagged as adverbs (RB) because they describe ‘manners’ of verbs.
Example:
□□□□□ □□□□□/RB □□□□□/RB □ □□□□ □□□□□□ □□□□
“The government is considering this internally”

□□□□ □□□□ : Same as ‘□□□□□ □□□□□’.

□□□□ □□□□ : Same as ‘□□□□□ □□□□□’.

□ □□□□□ : It will be tagged as □/CSC □□□□□/CS when it joins two clauses, but when this phrase appears at the starting of a sentence, it will be tagged as □/DT □□□□□/NN+SFON.
Example:
□□□ □□□□□ □□ □/CSC □□□□□/CS □□ □□□□□
“He didn’t come because I was not present”

But,
□□□ □□□□□ □□□ □/DT □□□□□/NN+SFON □□ □□□□□
“I wasn’t here. That’s why, he didn’t come”

References:

[1] "Part-of-speech Tagging Guidelines for the Penn Treebank Project (3rd Revision, 2nd printing),"
Beatrice Santorini, June 1990.

[2] "A Part of Speech Tagger for Indian Languages (POS Tagger)," *Tagset developed at IIIT –*
Hyderabad after consultations with several institutions thsrough two workshops, SPSAL 2007.