Part-of-speech (POS) Tagging of Bengali Written Text Corpus

Niladri Sekhar Dash

Linguistic Research Unit, Indian Statistical Institute, Kolkata, India

Email: ns_dash@yahoo.com

Abstract

In *Natural Language Processing*, *Language Technology*, and *Computational Linguistics*, the idea of part-of-speech (POS) tagging is linked with assigning words to particular parts-of-speech. It is also known as *Grammatical Annotation* (GA) and *Word Category Disambiguation* (WCD), the primary task of which involves the process of marking each word with a tag – manually or automatically – within a piece of natural text as corresponding to a particular part-of-speech, based on its form and function in contexts of its use in larger syntactic frames like phrases, sentences, paragraphs, and texts. An electronically developed language database, after being tagged at the part-of-speech level, becomes a valuable resource for various works of natural language processing, language technology, computational linguistics, machine learning, cognitive linguistics, applied linguistics, and descriptive linguistics. POS tagging is generally carried out on electronic version of language corpora (either manually or automatically) using a set of predefined tags, which are primarily used to assign part-of-speech, linguistic properties, and functions of words, terms, and other lexical items used in corpora. The issue of assigning part-of-speech to words, although appears to be simple, straightforward and one dimensional, is in fact, embedded with several theoretical and technical complexities with regard to identification of actual lexico-semantic entities and syntactico-grammatical functions of words used in a piece of text. Also, it includes defining the basic hierarchical modalities of tag assignment and designing a rule-based schema useful for automatic assignment of tags to words. All these issues ask for a highly synchronized strategy designed elegantly with proper combination of linguistic and extralinguistic knowledgebase and computational expertise for achieving maximum success with minimum enterprise.

Keeping these issues in view, in this paper we have made an attempt to define maxims, principles, and rules we need to follow if we try to design a tagset for tagging words at the part-of-speech level in Bengali corpus. We also define strategies we need to adopt when we try to tag words manually as well as try to formulate algorithms meant for automatic assignment of POS tags to words used in the Bengali corpus. We also address some other issues relating to tag assignment to words in the corpus with reference to some examples elicited from the Bengali corpus. Although we have addressed some important theoretical issues related to the field of inquiry within its general scope as well as have proposed some rudimentary rules to be followed at the time of POS tagging, we have deliberately avoided all technical issues and computational aspects of POS tagging with a clear expectation for making the topic of investigation an open area of general interest and enquiry. However, we have meticulously addressed almost all the major linguistics and extralinguistics issues relating to the process of manual POS tagging elaborately referring to the instances taken from the Bengali written text corpus. We hope that this paper will be treated as a 'genesis paper' the principal goal of which is to clarify the very concept of POS tagging; identify the maxims and principles of POS tagging; highlight the pros and cons of the tagset proposed by the *Bureau of Indian Standard* (BIS); define the strategies and rules for POS tagging of Bengali corpus; and make aware the new generation of scholars who are going to be engaged in the task of POS tagging of Bengali corpus about the invisible quicksand on the path of their journey.

Key Words: annotation, tagging, noun, verb, adjective, adverb, postposition, Bengali, part-of-speech, morphology, syntax, semantics, context

1. Introduction

In the areas of Natural Language Processing (NLP), Language Technology (LT), and Computational Linguistics (CL), the process of Part-of-Speech (POS) tagging is also identified as *Grammatical Annotation* (GA) the primary goal of which is to disambiguate words at the grammatical level and assign them to particular lexical categories or parts-of-speech. Because of this unique goal, POS tagging is called as *Word Category Disambiguation* (WCD), which in essence, involves the process of marking up words – manually or automatically – within a corpus as corresponding to particular parts-of-speech, based on their forms, functions and contexts (i.e., relationship of words with their adjacent and related words) within the larger syntactic strings like phrases, sentences, and paragraphs.

Although in practice, the process of POS tagging is a highly complicated and error-prone process, it cannot be ignored because research and development works of NLP, LT, and CL cannot advance further with just a list of words obtained from a language corpus without any information about the grammatical behaviours of the words in various usage-based contexts. In descriptive and applied linguistics on the other hand, POS tagging of words in corpus becomes equally necessary because it is observed that words are able to represent different parts-of-speech in different textual contexts and the information about parts-of-speech of words provided in traditional dictionaries is fuzzy, implicit, and indeterminate. This feature that words vary in part-of-speech is not a new thing to natural languages, as we find a large number of words in a language, which are ambiguous in form, meaning, and part-of-speech. This feature is prevalent in any living natural language of the world – be it an advanced language like English or an endangered language like Mundari. For instance, in English the word *sound* can be tagged as a noun (e.g., *the sound of music*), an adjective (e.g., *a sound decision*), as well as a verb (e.g., *he sounds rational*) based on the context of its use in sentences. Similarly in Bengali, the word হাত (hāt) 'hand', can be tagged as a noun, as a finite verb, as well as a non-finite verb based on its use in different contexts, as the following examples show:

(1a) তার হাতে কিছু টাকা এসেছে ।
(tār hāte$_{[NN]}$ kichu ṭākā eseche)
"Some money has come to his hand"

(1b) সে কিছু টাকা হাতিয়েছে ।
(se kichu ṭākā hātiyeche$_{[FV]}$)
"He has grabbed some money"

(1c) সে কিছু টাকা হাতিয়ে নিয়েছে ।
(se kichu ṭākā hātiye$_{[NFV]}$ niyeche)
"He has stolen some money"

When we perform POS tagging on the words in the sentences above, we shall indicate that the word হাতে (hāte) in sentence (1a) is a noun (NN), the word হাতিয়েছে (hātiyeche) in sentence (1b) is a finite verb (FV), while the word হাতিয়ে (hātiye) in sentence (1c) is a non-finite verb (NFV), although these words are actually derived from the noun হাত (hāt) 'hand' by adding different verbal inflections and suffixes. While a native Bengali speaker can identify these words as noun, finite verb, and non-finite verb, respectively, and can perform grammatical and semantic analysis of these words innately based on his/her internalised linguistic rules and grammar of the language, a computer system which is being trained to tag words automatically in natural text corpora, needs elaborate linguistic rules and conditions to perform the task of identifying parts-of-speech of words in texts.

Although it is often believed that the list of parts-of-speech presented in grammars, dictionaries, and school texts are enough for a language to tag words in texts, in reality, however, we have noted that there are many fine text categories and text sub-categories in a natural language text, and these fine divisions should be spelt out distinctly if we want to design a POS tagging scheme for words used in a text of a language. For instance, in Bengali, it is usually stated in grammar books that it has only eight parts-of-speech: Noun, Pronoun, Adjective, Adverb, Finite Verb, Non-finite Verb, Postposition, and Indeclinable. It is also assumed that only these parts-of-speech are required for the language, and once we learn these categories, we shall have no problem to identify the part-of-speech of each and every word used in the language. In actuality, however, we have noted that there are many fine text categories, like *demonstratives, infinitives, gerunds, conjunctions, enclitics, quantifiers, punctuations, particles,* etc., the form and function of which we need to identify, analyse and understand properly, if we want to have a total command over parts-of-speech of words vis-à-vis grammar of the language.

Furthermore, we can tag distinction between plural and singular nouns and pronouns; tag words marked with case makers and inflections in texts; tag grammatical gender found with words; tag nouns and adjectives marked with gender and number markers; tag verbs attached with person, number, gender, tense, aspect, modality, honourification, and other markers; tag adjectives marked with suffixes of degree; tag adjectives behaving like nouns in texts, etc. Information of tagging of all these types is required in an algorithm developed for automatic POS tagging of words in a language text by a computer system. Else a POS tagging algorithm will fail to trace varied unique linguistic information of words, while a POS tagged text will fail to exhibit the finer linguistic functions of words the information of which required in subsequent linguistic research and development activities related to both theoretical and applied linguistics.

In this paper I intend to draw attention towards these issues to deal with maxims, principles, and rules of POS tagging while I investigate the methodologies and guidelines proposed in the BIS (the Bureau of Indian Standard) tagset to be applied in POS tagging of Bengali corpora. In Section 2, I define the concept of POS tagging in general; in Section 3, I refer to some of the early POS taggers developed for English; in Section 4, I try to make distinctions between morphological analysis and POS tagging; in Section 5, I refer to some early POS tagsets developed for Bengali; in Section 6, I focus on the maxims of annotation as proposed in Leech (1993); in Section 7, I discuss about the principles of POS tagging as defined in the BIS-2010 tagset; in Section 8, I critically evaluate the usability of the BIS tagset on a Bengali corpus; in Section 9, I illustrate the process of manual POS tagging of Bengali corpus; in Section 10, I propose some rules to be followed at the time of POS tagging of natural texts; and in Section 11, I identify utilities of POS tagged corpora in various works of machine learning, natural language processing, language technology, applied linguistics, and descriptive linguistics. The data, examples and information presented in this paper are the outcomes of investigation I have undertaken for developing a POS tagged corpus of modern Bengali written text for various works of linguistics and language technology. Since the POS tagset and the process of POS tagging (either manual or automatic) are yet to be formalized for the language, the observations and arguments presented in this paper are clearly open for modification necessitated from the types of research initiated in linguistics and language technology.

2. What is Part-of-Speech (POS) Tagging?

In principle, part-of-speech (POS) tagging is a process of assigning part-of-speech tags to each and every word used in a piece of text after the word is passed through the stages of morphological analysis and grammatical interpretation (Garside 1995). Generally, a set of specially designed codes (known as 'tags') carrying grammatical information are assigned to words to indicate their parts-of-speech with regard to their use in the text (Leech and Garside 1982). In most cases, a well-defined set of linguistic rules are used to identify and assign POS tags to words to determine their lexico-semantic entities and syntactico-grammatical functions in the text. The immediate advantages of POS tagging can be realized in the following three levels:

(a) **Lexical level:** It allows to analyse morphological structure of words represented in their surface forms,
(b) **Orthographic level:** It draws distinction among the homographic forms used in the same text or similar other texts to make distinctions in their semantic roles, and
(c) **Syntactic level:** It allows identification of syntactico-grammatical functions of words to assign their POS entities accordingly.

The POS tagging is the commonest form of text annotation, which is considered to be the first stage of a more comprehensive process where multiword expressions, such as, *compound words, reduplicated forms, idiomatic expressions, proverbial expressions, set phrases*, and others used in a text are assigned with chunking markers leading to eventual assignment of phrase markers to each of the sentences used in a text. Although the use of POS tags on a text makes the text very difficult to read and comprehend for human beings, the text becomes maximally suitable for providing linguistic information needed by a computer for differentiating between words used in different parts-of-speech (Leech and Eyes 1993). Moreover, from application point of view, POS tagging is a useful technique that increases specificity in data retrieval from corpora and provides basic grammatical information about words required in semantic annotation, discourse annotation, parsing, dictionary compilation, grammar development, language teaching, and language planning[1].

In simple words, the process POS tagging on a piece of text, in a systematic manner, may be carried out through the following eight steps:

(a) Identification of words within a piece of text,
(b) Identification of their orthographic forms and appearances,
(c) Analysis of their morphological structures and formation,
(d) Identification of their syntactic (grammatical) functions in sentence,
(e) Determination of their grammatical roles and parts-of-speech,
(f) Identification of their semantic roles in the sentences,
(g) Assignment of POS tags – either manually or automatically, and
(h) Final verification and validation of the tags assigned to words.

Following the steps stated above, the process of POS tagging may be carried out on a piece of text at three separate stages as the followings:

Stage 1: Manual or automatic pre-editing of a corpus text,

Stage 2: Manual or automatic tag assignment to words, and
Stage 3: Manual post-editing of the tagged text database.

At the **pre-editing stage**, a language text database is converted into a suitable digital format for carrying out a POS tagging programme. At this stage, the entire text database is meticulously checked to verify if there is any typographical and/or orthographical error of any type within the text database, and if there is any, it is manually or automatically corrected in accordance with the physical source text before the digital text is put to POS tagging (Dash 2004). Moreover, if needed, a selected text database has to pass through the stages of text normalization and tokenization to make the database maximally suitable for POS tagging – both manually and automatically.

The **tag assignment stage** begins with the assignment of just one and only one POS tag to each and every word used within a sentence after considering specific syntactico-grammatical function of the words in sentence. For achieving higher level of accuracy at this stage, one can also use a lexical database where words are previously assigned with possible parts-of-speech for reference purposes. Such a lexical database is open-ended in the sense that it can be updated time-to-time with addition of new words obtained from various sources of language use. Also to deal with the newly found words in the corpus, which are not available in the previously made lexical database, one can adopt various methods such as the lists of common affixes and case markers with their possible parts-of-speech for achieving greater accuracy in POS tagging (Biber, Conrad, and Reppen 1998: 258-259).

At manual **post-editing stage**, the entire tagged text database is manually post-edited to verify if words are rightly tagged and if any error is made in POS tag assignment. In case of large corpora, where manual verification of the entire database appears to be highly time consuming, tedious and error-prone, a probability matrix may be devised from the tagged database to deal with the problems of ambiguous tagging and dubious tag assignment (Leech, Garside, and Atwell 1983). The matrix will help to specify transition probabilities underlying between the adjacent tags. For example, if a given word is tagged as noun (W$_{\_N}$), the probability of its immediately preceding word to be an adjective (W$_{\_JJ}$) is quite high in a language like Bengali[2].

Usually a human being, who is engaged in assigning POS tags to the words manually, can do the work quite successfully if she is well versed in morphology and grammar of the language. On the other hand, a computer programme can also do this task quite successfully if it is supplied with adequate amount of linguistic information and rules for POS tag assignment as well as it is trained properly to do the work with less percentage of errors. That means a system designer who is engaged in designing a tool for automatic POS tag assignment to words should be highly well-equipped with adequate linguistic and grammatical knowledge of the language so that she can develop a robust and accurate computer system to assign correct POS to words, terms, and other lexical items used in the text (Kupiec 1992).

However, before the task of POS tagging is executed on a written text corpus database – either manually or automatically – there arises an urgent need for a hierarchically well-defined and globally accepted standard POS tagset, which will be used in a uniformed manner by one an all engaged (or going to be engaged) in POS tagging for a language.

3. Early History of POS Tagging: A Birds' Eye View

In the early years of corpus processing, POS tagging was mainly done manually where two or three people were trained in the grammar of a particular language and were engaged in the task. These people used to assign POS tags to each and every word by hand after reading the whole sentence where the words have been used. However, since this process was considered to be highly lengthy, tedious, time-consuming, and error-prone, POS tagging is, at present, done mostly automatically using algorithms and rules, based on which a computer system tries to associate discrete terms as well as hidden parts-of-speech to each and every word in a text, in accordance with a set of descriptive POS tags designed beforehand keeping in mind the language-specific lexical and grammatical categories of a particular language (Toutanova and Manning 2000).

The issues of designing tagsets to be assigned with words as well as the research on the process of POS tagging have been an area of intensive exploration in corpus linguistics, and this has been closely linked with the birth of the *Brown Corpus* (Francis and Kuchera 1964). The *Brown Corpus* was carefully and painstakingly 'tagged' with part-of-speech markers over many years – initially manually and then automatically. The first approximation was done with a program developed by Greene and Rubin (Greene and Rubin 1971), which contained a large handmade list of rules and probabilities, such as, what categories of words could co-occur at all in the sentences of the corpus. For instance, in English, a noun has greater probability to occur after an article than a verb (arguably). The execution of such rule-based programs produced nearly 70% accuracy in proper identification of parts-of-speech of words automatically. The results were repeatedly reviewed and corrected manually to develop a set of fine-tuned rules, which were finally implemented into the system. By the late 70s, this tagging program was nearly perfect giving highest possible accuracy leaving aside some stray cases on which even human annotators might not like to agree (Toutanova, *et al.* 2003).

The tagged *Brown Corpus* has been extensively used over the years for innumerable linguistic studies relating to frequency of word use and frequency of use of parts-of-speech of words in texts. In fact, the tagged *Brown Corpus* inspired for development of similar 'tagged' corpora in many other languages across the world, particularly the English corpora which were developed by this time in England and other countries of Europe, Australia and New Zeeland (Garside, Leech and McEnery 1997, Kennedy 1998). The statistical data, and the patterns and information derived from the analysis of the tagged *Brown Corpus* also provided a baseline direction for development of later POS tagging systems, like the *CLAWS tagger* and the *Brill Tagger*, as discussed below.

3.1 CLAWS POS Tagger

The *Constituent Likelihood Automatic Word-tagging System* (CLAWS) for POS tagging for English texts was first developed at the *Lancaster University*, UK. From the early 1980s, this system has been continuously revised and modified several times to give its present shape (Fligelstone, Rayson, and Smith 1996). The fourth revised version of this tagger (CLAWS4) has been used to POS tag more than hundred million words of the *British National Corpus* with appreciable rate of accuracy. This system has consistently achieved 96 to 97% accuracy in POS tagging even though the precise degree of accuracy varied based on the type of English text. Experiments carried out on the major part-of-speech categories shows that the system has an error-rate of only 1.5%, and nearly 3.3% of words are ambiguous whose ambiguities still remain unresolved. At present a Template Tagger – a tool of rule-based formalism – is built into CLAWS4 to act as post-processor of tagging process. The Template Tagger is created from manual analysis of tagged corpora and from the knowledge of frequent errors created by the CLAWS4. The implementation of the Template Tagger has drastically improved the tagging accuracy in the resulting corpus database (Fligelstone, Pacey, and Rayson 1997).

Till date, several tagsets have been used in CLAWS over the years and these tagsets have been modified time and again to make these maximally appropriate for the modern English texts. The tagset of CLAWS1 included 132 basic tags for words, many of which were identical in form and application to those tags used in the *Brown Corpus*. The revised version of the tagset that was used in CLAWS2 was further enlarged including 166 tags. However, since such elaborate tagset created problems in the task of tagging, the number was reduced for general purposes. Thus the tagset that was used for the *British National Corpus* contained only 60 tags, as this was designed mainly for handling much larger quantities of data than the data that were dealt with up to that point in other specialised works. On the other hand, the sample database of the *British National Corpus* contained more than 160 tags. The current standard tagset of CLAWS is the C7_Tagset, which is more advanced with addition of tags for the punctuation marks. The C7_Tagset has been further upgraded to produce C8_Tagset to make finer distinctions in determiner and pronoun categories as well as for auxiliary verbs in English.

With regard to tagging guidelines, several detailed strategies have been proposed in CLAWS4 to decide how to draw the line of distinction between the correct and the incorrect assignment of tags. This was essential as there was a clear need for detailed guidelines to be used in tagging practice and therefore this has been carefully created to remove confusion of any kind about what is a 'correct' or an 'accurate' tagging in the corpus. The tagging guidelines proposed in CLAWS4 are also useful for Indian language corpora, as these guidelines have some rules which are mostly language independent and hence, relevant for all natural languages. Modification on these guidelines may be incorporated in accordance with the requirement of specific Indian languages, if we think of using these guidelines for POS tagging Indian language corpora.

3.2 Hidden Markov Model

In the middle of 1980s, the researchers in England, Norway, and Sweden, when working to tag a large database of the *Lancaster-Oslo-Bergen Corpus* (LOB), started using Hidden Markov Model (HMM) to disambiguate parts-of-speech of words. The HMM was adopted because it provided an opportunity for counting the cases (mostly from *Brown Corpus*) for developing tables regarding the probabilities of certain sequences of lexical items and as well as the patterns of usage of words in formation of natural sentences. For instance, it helped researchers to trace that if the article 'the' occurs at certain place within a sentence, then the next word is either a noun (40%), or an adjective (40%), or a number (20%). Based on information of this kind, a computer program was developed that could decide that the word 'cook' within a string of *the cook has left for home* is far more likely to be a noun than a verb. This method also helped the researchers to benefit from the knowledge about the part-of-speech of the following words.

The higher order HMM could do many more tasks by learning probabilities and possibilities not only of word pairs, but also about the strings of three or more words combined together to form larger sequences. For instance, if a POS tagging program first encounters an article followed by a verb, then the possibility of the very next word to be a preposition, article or noun is higher than being another verb. This approach has been able to provide the much required break-through in the development of automatic POS tagging programs for English and many other languages.

The application of HMM approach on the *LOB Corpus* showed that in certain situations where several ambiguous words occurred together, the possibility of identification of actual part-of-speech of each word was simply multiplied. However, with the higher order HMM, it was easy to enumerate every possible combination of words and assign a relative probability to each word by multiplying together the probabilities of each choice in turn. The combination with highest probability was then chosen as the most suitable candidate for tagging. The research team of the *Lancaster University*, UK adopted this technique and achieved a considerably high level of accuracy (93-95%) in POS tagging of different English corpora.

However, some scholars have raised criticism against the HMM approach to argue that in the act of dissolving ambiguities merely assigning the most common tag to each known word and the tag 'proper noun' to all unknown words can achieve moderate level of accuracy (nearly 90%), because many such words are actually unambiguous within a text (Charniak 1997).

Another limitation of the HMM-based POS tagging was noted in its 'generous nature' in the act of tag assignment. Although this technique was widely advocated because of its high rate of accuracy in the CLAWS system, it had proved to be quite expensive as it tended to enumerate all possibilities. Moreover, sometimes, it had to resort to backup methods when there were too many possibilities. For example, the *Brown Corpus* contained an extreme case where all words within a string of seventeen words in a row were ambiguous. Moreover, there were some cases where words such as 'still' could be represented in as many as seven different parts-of-speech. Such complexities in identification of ambiguities by HMM program actually made a tagged corpus more deceptive and less reliable than it actually was before it was tagged.

3.3 Dynamic Programming Algorithm

A Dynamic Programming Algorithm (DPA) was developed in 1988 to dissolve the problems of ambiguity in POS tag assignment in corpora (DeRose 1988). This algorithm could execute the task in vastly less time and more or less accurately. Based on Viterbi Algorithm[3] this method used a table of word pairs in an indigenous method for estimating the values for the word-pairs in corpus. Following this strategy it achieved not only high rate of accuracy (over 95%) in trial corpus, but also included within its analysis the results of specific types of error, probabilities, and other related information (Abney 1997). This strategy had also been replicated on a Greek language corpus where it proved to be similarly effective.

This innovative POS tagging method surprisingly disrupted many on-going POS tagging activities of corpora in English and other languages. The rate of accuracy reported in this study (DeRose 1990) was much higher than typical accuracy rates acquired through application of sophisticated algorithms that usually integrated part-of-speech information of words with other levels of linguistic information relating to syntax, morphology, semantics, and so on.

In general, however, all the methods and algorithms (CLAWS and DPA) failed miserably in case of those words where information from the fields of semantics and other domains was required for accurate POS tagging. Strikingly, such cases are not very rare in all natural language texts. Moreover, these systems failed in those cases where information from the domains of discourse, pragmatics, and extralinguistics was required for identification of actual part-of-speech of words.

Such limitations in POS tagging led scholars to realize that part-of-speech tagging should strictly be separated out from other levels of corpus processing, such as parsing. It was also understood that POS tagging and parsing are two different ways of treating language texts with different goals, and hence, they should be treated separately, at least, at the initial stage of corpus processing. It simplified the approach which rejuvenated researchers to separate this area from other areas of corpus processing and to chalk out new methods for POS tagging. Although the HMM was generally accepted as the standard method for POS tagging, new models were also considered with introduction of rule-based, stochastic, neural, and statistical approaches. These approaches produced methods like Brill Tagger, TnT Tagger, etc., which promised to achieve higher accuracy. Some of these methods are briefly discussed in the following sections.

3.4 Brill POS Tagger

The Brill POS Tagger was developed by Eric Brill in 1993 and it is mostly known as an 'error-driven transformation-based tagger', which generates an interface in the act of doing part-of-speech tagging in pre-defined rule-based method (Brill 1995). It is error-driven in the sense that it recourses the process of supervised learning; and it is transformation-based in the sense that a tag may be assigned to a word as well as changed using a set of pre-defined rules. For instance, if a word is already known to the system, the tagger will first assigns the most frequent tag. On the other hand, if the word is unknown to the system, the tagger will naively assign the tag of 'noun' to it. Thus applying the rules over and over and changing into incorrect tags when required, this system is able to achieve high level of accuracy in output. The algorithms used by this method can be summarised in the following six stages:

Stage 1: Start the tagger on a digital corpus database.
Stage 2: Encounter a (new) word (in inbuilt lexicon) and assign the most frequent tag associated to the form of the word.
Stage 3: Encounter an unknown word (out of inbuilt lexicon) and tag it as proper noun if capitalized, else as simple noun (if not capitalized).
Stage 4: Learn or guess tags on the basis of contextual rules.
Stage 5: Change the incorrect tag to a correct one with contextual rules.
Stage 6: Generate the output.

The initial learning phase of Brill POS Tagger involves several sub-stages and strategies as the followings: First, it iteratively computes the error score of each candidate rule to calculate the difference between the number of errors before and after applying the rule. Second, it selects the best (higher score) rule. Third, it adds it to the rule set and applies it to the text again. Fourth, it repeats the process until no rule has a score above a given threshold. That means, if the chosen threshold is zero, it continues the application of rules until it achieves a greater score than the chosen threshold. Once it is achieved, it is then supposed to be the final stage of the tagging.

For achieving greater rate of success it applies two sets of rules: the first set of rules are called Lexical Rules, which are used for initialisation of the process, while the second set of rules are known as Contextual Rules, which are used to remove errors and correct the final tags.

(a) Lexical Rule : Tag W_1 → W_N IF W_1 carries suffix like '-tion', '-ment', etc.
(b) Contextual Rule : Tag W_1 → W_2 IF the preceding or following tag is X.

The Brill POS Tagger has some problems with parts-of-speech belonging to open classes because of their complicated morphological structure. On the other hand, grammatical categories are easier to detect and annotate correctly. It has been shown that it is more difficult to derive information from the words which are annotated with only POS tags, than from the words whose tags include information about their inflectional categories. Therefore, it is suggested that for better evaluation of the system it was necessary to test the tagger on many large corpora. Higher accuracy of the system could probably be achieved by using it on many larger corpora made with different text types. Moreover, the use of larger lexicon would reduce the number of unknown words. Additionally, the accuracy of the system could be greatly improved if the rule generating mechanisms in the Brill POS Tagger would become flexible allowing consideration of different characteristics of languages.

Although it has been claimed that Brill POS Tagger is a system, which is language independent (Brill 1995), in actuality however, it has several difficulties with the languages which are dissimilar in their form and characteristics from that of English. We wonder if adoption of this system for Indian language corpora will become a liability than an advantage.

3.6 TnT POS Tagger

The Trigrams-n-Tagging (TnT) is claimed to be an elegant and efficient system of statistical POS tagging that can be trained for different languages and can virtually be adopted for any tagset usable for a language or a variety (Brants 2000). The competence of this system depends largely on its ability in generating components for parameters through its use on tagged corpora. Moreover, it is claimed to be capable of incorporating several methods of smoothing to handle unknown words not previously encountered by the tagger. Since this tagger is not optimized for a particular language or a variety, it is open for training on a large variety of corpora belonging to different languages. Therefore, adapting this tagger to a new language, new genre of text, or a new tagset is not a great problem. Moreover, it can be optimized in speed so that it can generate quick outputs from all kinds of text corpus on which the tagger is deployed.

The TnT tagger is an outcome of the Viterbi Algorithm for second orders Hidden Markov Model. It processes words by analysing the suffix parts tagged with the words. The primary paradigm used for smoothing is linear interpolation while respective weights are determined by delayed interpolation. In this system all unknown words are handled by a suffix trie and successive abstraction. The tool can directly be applied on language corpora by using three different modes:

(a) In the first mode the input file will contain one token per line.
(b) In the base mode, the tagger adds a second column to each line, containing the tag for the word.
(c) Finally, in the third mode, the tagger optionally emits alternative tags for each token, together with a probability distribution.

In the output database if a word is marked with an asterisk (*) it has to be considered that the words is not in the lexicon used by the tagger. The speed of POS tagging of words depends on the rate of average ambiguity of words and the percentage of unknown words used in the text.

This tagger was applied on a small part of the *Susanne Corpus* and it generated 94.5% accuracy due to small size of the corpus (around 1,50,000 tokens) and the large tagset (around 160 plus multi-token tags). But when it was applied on the large English corpora like the *Penn Treebank* the accuracy of the output was much higher (nearly 97%) as the number of tagset was reduced for the Treebank. It is also claimed that the tagger can be trained on language databases where written words are separated by white space (Brants 2000). The most notable limitation of the TnT tagger is that although it acts well with any tagset that is represented in ASCII, it cannot work properly where tagset is represented in Unicode (UTF-8).

The approaches so far discussed use pre-existing digital language corpora to experiment and apply methods of tagging algorithms. These experiments revealed that it is possible to bootstrap the texts by using ‘unsupervised’ tagging conventions. In most cases, an unsupervised tagging technique uses a small part of untagged corpus for training purposes and thus produces the tagset through induction. That is, it observes the patterns in word use in sample databases and derives part-of-speech categories themselves for words. For example, while statistical information readily reveals that English particles like ‘the’, ‘a’, and ‘an’ can occur in similar syntactic contexts, verb ‘eat’ can occur in a very different context or situations. It is observed that the application of this technique with sufficient iterations and repeated use in varieties of text can generate similarity classes of words which are remarkably similar to those human linguists will expect or design for. Moreover, application of this technique sometimes produces certain differences by itself which provide many valuable and new insights in part-of-speech of words, never presumed by human annotators.

For decades now, POS tagging is being considered as an inseparable part of corpus processing, because it has become indispensable for any system or tool meant to be used in language technology to identify accurately the POS of words used in a piece of text. However, our past experiences showed that not a single full-proof automatic system for POS tagging is developed yet for English or any other advanced languages, even though we have come across some POS tagging systems, which are quite robust and useful.

What is realized from this enterprise is that there are large numbers of words to whom the correct part-of-speech cannot be assigned without understanding their context of occurrence and the meaning they denote in different contexts. We also need to understand the pragmatics and even the discourse of a text to assign right POS tags to words. This is an extremely complex and expensive pre-condition for any automatic tagging system, because analyzing higher levels related to semantics, pragmatics, and discourse of a language text is a cognitive enigma, which an automatic POS tagging tool cannot dispel with its present knowledgebase and operation skill. Therefore, the workable solution is to apply a good POS tagger on text corpora and then manually correct the erroneous outputs, which is still a dream for most of the Indian languages corpora, including Bengali.

4. Morphological Analysis and POS Tagging

Once the basic principles and rules are developed and collectively agreed upon, we can go for tagging a corpus text manually or train a system to tag the text automatically. However, we need to keep in mind that a POS tagging scheme is not a replacement for a tool used for morphological analysis of words, because POS tagging is different in goal and application. The basic goal of a morphological analyser is to identify the morphemes used in formation of a word, while the goal of a POS tagger is to identify grammatical identity or part-of speech of a word used in a text.

A word when used in a sentence usually carries various information, some of which are distinctly evident from its surface form (e.g., *number marker, plural marker, gender marker, person marker, case marker, tense marker, emphatic marker, negative marker,* etc.) while others are obtainable from its contextual environment (e.g., part-of-speech or grammatical function of a word in a given context). That means both kinds of morpheme-related information of a word can come from morphological and grammatical analysis of a word. This signifies that we should look at a word from two levels:

(a) Morphological level, and
(b) Syntactic level.

Information of morphological level is possible to extract from morphological structure of a word analysed individually in isolated context. On the other hand, information of syntactic level is possible to extract from its grammatical function within a syntactic construction. While the first function is carried out by a morphological analyzer, the second function is carried out by a POS tagger (and a parser). Thus, morphological analyzer and POS tagger play crucial roles in providing morphological and grammatical information of words – one at their isolated situation and the other at their sentential frames. For elucidation, consider the Bengali word বদলে (badale), which is treated in two different manners in two different systems to elicit two different types of linguistic identity of the word.

Morphological Analysis:

Information Set: 1

Surface Form	: বদলে (badale)
Base Form	: বদল (badal)
Suffix Part	: -ে (-e)
Part-of-Speech	: Non-finite Verb$_{[VNF]}$
Meaning	: “Replacing/changing”

Information Set: 2

Surface Form	: বদলে (badale)
Base Form	: বদলে (badale)
Suffix Part	: Ø
Part-of-Speech	: Postposition$_{[PSP]}$
Meaning	: "In exchange of"

Information extracted from morphological analysis shows that the word বদলে (badale), in an isolated context, can have two types of morphological information. It can be a non-finite verb (VNF) or a postposition (PSP). A morphological analyzer intends to capture this information of a word after analyzing its morphological structure or surface form.

On the other hand, a POS tagger tries to determine the actual grammatical role of the word in a sentential context, as the following examples show (2a-2b). It tries to determine that the word বদলে (badale) in the first sentence (2a) is a non-finite verb (VNF), and in the second sentence (2b) is a postposition (PSP), because in two different sentences it performs two different grammatical roles. Thus a POS tagger, taking into consideration syntactico-semantic roles of a word within a sentential frame, tries to determine if the word is used as a noun or as a verb in a particular sentential context (Leech 1997).

POS Tagging:

(2a) সময় কিন্তু এখন অনেক বদলে$_{[VNF]}$ গেছে।
(samay kintu ekhan anek badle geche)
"Time has indeed changed so much by now"

(2b) তুমি আজকের বদলে$_{[PSP]}$ বরং কাল আসতে পারো।
(tumi ājker badale baraṃ kāl āste pāro)
"You can come tomorrow rather than today.

What is understood from the above discussion is that a morphological analyzer primarily looks at a word in isolation and provides all its possible morphology-related analyses, including multiple parts-of-speech or lexical identities. On the contrary, a POS tagger looks at a word in a sentential context (within small window) and provides the exact part-of-speech of the word in a syntactic frame. Therefore, it can be argued that the task of a POS tagger is primarily to disambiguate the grammatical category information provided by a morphological analyzer and select the most appropriate POS category of words in specific sentential contexts. Taking this argument as a ground truth, a POS tagset for a language is designed primarily with the grammatical categories of words in mind. Any other morphological information, which is obtained from a morphological analyzer, is not usually included at POS tagging level.

One important thing we need to keep in mind that before we apply morphological analysis or POS tagging automatically on a text corpus, we must carry out these tasks manually on sample texts to gather experience, to reduce amount of error, and to save time. Prior experiments are also required for achieving higher accuracy and robustness when morphological analysis and POS tagging are to be done automatically on corpora. In essence, prior trial and correction exercise usually generate all the following advantages.

(a) Experimental morphological analysis and POS tagging of words in a corpus database throw up many new issues and challenges, based on which predetermined tagsets can be revised, modified, and extended.

(b) Elaborate guidelines for morphological analysis and POS tagging are developed during this period based on experimental analysis and tagging exercises.

(c) Morphologically analysed words and POS tagged corpora may be utilized for training automatic morphological analyzer and POS tagger.

(d) Prior experiments supply necessary knowledge and insights required for enhancing robustness of a morphological analyser and a POS tagger.

(e) Information gathered from prior experiments may also be used in language description, dictionary compilation, grammar book writing, and language teaching.

5. Early POS Taggers in Bengali

Although techniques and tools for POS tagging are already developed and used for English and many other advanced languages in the early 1990s (Leech, Garside, and Bryant 1994, Garside, Leech and McEnery 1997), these are yet to be developed and frozen for the Indian languages including Bengali. During last few years, many people have either asked for a tagged Bengali corpus or suggested for developing it at various platforms and forums, but none has so far taken serious initiative in this direction. Nearly, six years ago, arguably the first generic POS tagset for Bengali was designed by an individual to tag manually a text database of nearly hundred thousand words of modern Bengali prose for academic and training purposes (Dash 2005a). Some ideas may be obtained about POS tagging in Bengali from the major POS categories of this tagset (Table 1) and the exercises that were carried out on a sample written Bengali text database (Dash 2005b: 104-108).

No.	POS Categories	Label	Example
1	Noun	[NN]	বালক (bālak), শহর (śahar), কথা (kathā), মানুষ (mānuṣ), etc.
2	Pronoun	[PR]	আমি (āmi), তুমি (tumi), সে (se), তারা (tārā), তুই (tui), etc.
3	Demonstrative	[DM]	যে (ýe), এই (ei), ওই (oi), তাই (tāi), etc.
4	Finite Verb	[FV]	করছি (karchi), করতাম (kartām), গেল (gela), যাবে (ýābe), etc.
5	Non-Finite Verb	[NF]	করলে (karle), করতে (karte), গেলে (gele), গিয়ে (giye), etc.
6	Adjective	[AD]	ভাল (bhāla), মন্দ (manda), সুন্দর (sundar), সাদা (sādā), etc.
7	Adverb	[AV]	হঠাত্ (haṭhāt), বাবদ (bābad), কারণে (kāraṇe), etc.
8	Postposition	[PP]	পরে (pare), কাছে (kāche), আগে (āge), নিচে (nice), etc.
9	Conjunction	[CN]	তবে (tabe), যদি (ýadi), নইলে (naile), যাতে (ýāte), etc.
10	Indeclinable	[IN]	কিন্তু (kintu), অথবা (athabā), বরং (baraṃ), আর (ār), etc.
11	Particle	[PT]	ই (i), ও (o), তো (to), না (nā), নে (ne), নি (ni), etc.
12	Quantifier	[QT]	এক (ek), দুই (dui), প্রথম (pratham), পয়লা (paylā), etc.
13	Reduplication	[RD]	বনে বনে (bane bane), কত কত (kata kata), যে যে (ýe ýe), etc.
14	Punctuation	[PN]	., : ; - / …, !, ? ( ), [ ], { }, etc.
15	Others	[OR]	Mathematical symbols, +, -, x, >, <, $, #, @, ^, &, * etc.

Table 1: The generic POS categories proposed for Bengali (Dash 2005a)

Recently, the *Microsoft Research India*, Bangalore has developed a tagset for Bengali (2009) under the scheme for providing a common tagset framework for Indian languages. This tagset offers flexibility, cross-linguistic compatibility, and reusability across all Indian languages. This tagset has been modified to a certain extent to be adopted and used by the *Central Institute of Indian Languages*, Mysore. This tagset is also used by the *Microsoft Research Labs India* to develop a manually tagged Bengali text database of 7,168 sentences (10,2,933 words) including some Bengali texts obtained from blogs, *Wikipedia, MultiKulti* and a portion of the *EMILLE/CIIL Corpus*. This tagged database is available from *Linguistic Data Consortium* of the *University of Pennsylvania*, USA. The annotated data is structured into two ways: Bengali-I containing 3,684 sentences (51,091 words) and Bengali-II containing 3,484 sentences (51,842 words). The annotated data is provided in XML and text files. The XML files contain metadata about the material, such as language, encoding, and data size. The *Annotation Guidelines for Bengali* included in this release contain a description of annotation methodology (http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2010T16)[4].

তাই$_{[AV]}$ মানুষ$_{[NN]}$ তাহার$_{[PR]}$ সংগৃহীত$_{[AD]}$ দ্রব্যের$_{[NN]}$ মধ্যে$_{[PP]}$ যেইটুকু$_{[PR]}$ প্রয়োজনের$_{[NN]}$ অতিরিক্ত$_{[AD]}$ সেইটি$_{[PR]}$ অন্যের$_{[PR]}$ সংগৃহীত$_{[AD]}$ পৃথক$_{[AD]}$ ধরনের$_{[NN]}$ দ্রব্যের$_{[NN]}$ সহিত$_{[PP]}$ বিনিময়$_{[NN]}$ করিয়া$_{[NF]}$ নিজের$_{[PR]}$ প্রয়োজনীয়$_{[AD]}$ দ্রব্য$_{[NN]}$ সংগ্রহ$_{[NN]}$ করিত$_{[FV]}$ । $_{[PN]}$ মানুষ$_{[NN]}$ যেই$_{[DM]}$ যুগে$_{[NN]}$ গুহার$_{[NN]}$ অভ্যন্তরে$_{[PP]}$ বসবাস$_{[NN]}$ শুরু$_{[NN]}$ করিয়াছিল$_{[FV]}$,$_{[PN]}$ সেই$_{[DM]}$ যুগ$_{[NN]}$ হইতেই$_{[PP]}$ মানবসভ্যতার$_{[NN]}$ ক্রমবর্ধমান$_{[AD]}$ অগ্রগতির$_{[NN]}$ সূচনার$_{[NN]}$ সঙ্গে$_{[RD]}$ সঙ্গে$_{[RD]}$ হিসাবশাস্ত্রের$_{[NN]}$ ভিত্তি$_{[NN]}$ প্রস্তর$_{[NN]}$ স্থাপিত$_{[AD]}$ হইয়াছিল$_{[FV]}$ । $_{[PN]}$ আদিমযুগে$_{[NN]}$ মানুষ$_{[NN]}$ পশু$_{NN]}$ শিকার$_{[NN]}$,$_{[PN]}$ মতস্য$_{[NN]}$ শিকার$_{[NN]}$ ও$_{[IN]}$ বন্য$_{[AD]}$ ফলমূল$_{[NN]}$ সংগ্রহ$_{[NN]}$ করিয়া$_{[NF]}$ জীবিকা$_{[NN]}$ নির্বাহ$_{[NN]}$ করিত$_{[FV]}$ । $_{[PN]}$

Fig. 1: Sample of a POS tagged Bengali written text (Dash 2005b)

Very recently, however, the *Department of Information Technology* of the *Ministry of Information and Communication Technology, Govt. of India*, under the initiative entitled *POS Tag Standardization Committee* formed for the purpose, has been able to mark the *Standards for POS Tag Set for the Indian Languages* after several meetings held at different institutes of the country during last one year (2009-2010). This tagset is known as the BIS tagset, which is critically evaluated in some details in Section 7 and Section 8.

6. Maxims of POS Tagging

The basic purpose of language corpus tagging is to add-up various interpretative (i.e., intralinguistic and extralinguistic) information to a written or spoken text database by some encoding attached to, or interspersed with the electronic version of the corpus text. That means apart from pure language database, a text corpus may also be provided with some additional intralinguistic and extralinguistic information, which may be divided broadly into following two types:

(a) Representational information, and
(b) Interpretative information.

For a written text corpus database, representational information is the actual form of the text that contains *characters, letters, words, terms, phrases, sentences, spellings,* *punctuations*, and some other visual information. Interpretative information, on the other hand, is something which is added to the basic text database, usually by expert linguists presumed to have insight into, or knowledge of, the linguistic features of the text as well as of the language (Leech 1993: 275).

According to Leech and Smith (1999), there are, at least, three basic criteria indispensable in any kind of text annotation or tagging. These are: **consistency, accuracy,** and **speed**. While ***consistency*** demands for a kind of uniformity in the scheme of text annotation throughout the language database stored within a corpus; ***accuracy*** demands for freedom from any kind of error in the tagset or in the process of tagging. It also asks for adherence with definitions and guidelines of the strategy designed for annotation. Finally, ***speed*** directs towards automatic implementation of the scheme on a very large amount of corpus database within a short span of time[5].

Since the scheme of tagging of written text corpus is yet to be standardised for most of the Indian languages including Bengali, and since the acceptance of tagging standards depends heavily on corpus evaluating skill of the experts who are engaged in adding these tags to the texts keeping in mind their usefulness to the scheme they have adopted, Leech (1993) has identified the following seven maxims to be applied strictly in tagging of a text corpus.

(a) It should always be easy to dispense with annotation and revert to the raw corpus. The raw corpus should be recoverable.

(b) The annotations should, correspondingly, be extractable from the raw corpus to be stored independently or stored in an interlinear format.

(c) The scheme of analysis presupposed by the annotations – the annotation scheme – should be based on principles or guidelines accessible to the end-user. The annotation scheme should consist of the set of annotative symbols used, their definitions, and the rules and guidelines for their application.

(d) It should also be made clear beforehand about how and by whom all the annotations were applied.

(e) There can be no claim that the annotation scheme represents 'God's truth'. Rather, the annotated corpus is made available to a research community on a *caveat emptor* principle. It is offered as a matter of convenience only, on the assumption that many users will find it useful to use a corpus with annotations already built in, rather than to devise and apply their own annotation schemes from scratch (a task which could take them years to accomplish).

(f) To avoid misapplication, annotation schemes should preferably be based as far as possible on 'consensual', and 'theory-neutral analyses' of the corpus data.

(g) No annotation scheme should claim authority as the standard, although *de facto* interchange of 'standards' may arise through widening availability of annotated corpora. And this approach should be encouraged.

In general, tagging of intralinguistic information involves encoding words, terms, phrases, and other linguistic items used in a corpus database with their part-of-speech and grammatical information (due to this POS tagging is also called *grammatical annotation*), while tagging of extralinguistic information encodes the same words and terms with information of orthography, semantics, discourse, pragmatics, anaphora, and sociolinguistics, etc. Thus, taking into consideration the nature and scope of information to be tagged with words and other linguistic items used within a text corpus, tagging of a corpus, according to our argument, may be classified into eight broad types:

(a) **Orthography tagging** (or Orthographic annotation): This process aims at representing a text as much as possible as it actually exists in its complete natural state, despite the attachment of multiple extratextual and textual tags. It tags, for example, different orthographic symbols, such as, *single quotes, double quotes, type size, indentation, bold face, italics,* etc., as well as tags the *capital letters, periods, apostrophes, segments, paragraphs, lines, punctuations, abbreviations, postcodes,* etc. used in the text.

(b) **Part-of-speech tagging** (or Grammatical annotation): It involves assigning specific part-of-speech to words after understanding their actual grammatical roles within given sentences. At the sentence level, this information may be tagged for chunks such as multiword expressions[6], local word groups, phrases, and idiomatic expressions, etc. These are actually minimal constituent units which allow a sentence to be parsed at shallow or skeleton level. It may also involve marking of dependencies, constituents, named entities, and predicates and their arguments found within the structures of sentences.

(c) **Prosody tagging** (or Prosodic annotation): It is normally carried out on a spoken text corpus, after a speech corpus is transcribed into written form. In general, it tags all kinds of prosodic features, such as, *pitch, accent, intonation, loudness, length, pause, tone, tonal variation, juncture*, and other suprasegmental features observed in the spoken form of a text. It may also tag information about those features that are noted in normal spoken discourses, such as, *hesitations, repetitions, false starts, fillers, non-beginnings, non-ends, abrupt termination*, *laughter, external noise,* etc. These are normally tagged manually with special set of tags that capture information of a normal spoken interaction recorded in actual socio-cultural setting of linguistic communication.

(d) **Word-sense tagging** (or Semantic annotation): It is applied on a language corpus database to capture the appropriate sense of a particular word within a given context. The basic goal of this type of tagging is to distinguish the lexicographic senses of words – a process used in word sense disambiguation and assignment of semantic domains to words used in texts. Thus, word sense tagging work at identifying the semantic information of words used in corpora and exhibiting semantic relationships underlying between the words within texts. It is a higher level of tagging, which also marks agent-patient relationships of words denoting their particular actions.

(e) **Discourse tagging** (or Discoursal annotation): This involves marking of discoursal elements, sociolinguistic cues, and many other extralinguistic features embedded within a piece of text. Here a language corpus is tagged at the level beyond the sentence boundaries to explore the discoursal and/or pragmatic relations expressed by the linguistic elements used in the corpus. For instance, proper identification of discourse elements in spoken texts becomes necessary for indicating the conversational structure of normal speech sequences.

(f) **Anaphora tagging** (or Anaphoric annotation): It tags anaphora and anaphoric relations of words used in a text for intra-sentential or intra-textual references. Usually, various pronouns and nouns are co-indexed within a broad framework of cohesion analysis proposed by Halliday and Hasan (1976). It aims at identifying different types of anaphora used in the texts, mark these anaphors, and sort out these forms to dissolve anaphoric complexities and ambiguities.

(g) **Figure-of-speech tagging** (or Rhetoric annotation): It tries to tag various rhetorical devices like *metaphors, metonymies, idioms, foregrounding, hyperboles, zeugmas, proverbs, similes*, etc. that are different from customary construction, order, or significance of a normal language text. The use of a figure-of-speech in a text is a kind of change from the ordinary manner of expression, in which words are used in other than their literal denotative senses to enhance the ways thoughts have been expressed. It is an important process of tagging, because it helps to capture how various figures-of-speech are used in normal text (either prose or poetic) and how do they contribute in construction of information in the text.

(h) **Etymological tagging** (or Source annotation): This process is used to tag the source language of words wherefrom these are obtained and used in a language. This is indispensible for the languages like English and Bengali, since a large chunk of vocabulary of these languages is actually obtained from various other languages. Bengali, for instance, is very much proud to possess a large list of words borrowed from many languages such as English, Arabic, Persian, Portuguese, Hindi, Urdu, Tamil, Dutch, Spanish, German, Japanese, Chinese, etc. At the time of etymological tagging, actual etymological origin of the words used in the Bengali corpus is tagged for future linguistic works.

Although a corpus tagged with various types of linguistic information is considered highly useful for several works of applied linguistics and language technology, the process of tagging, however, asks for long-time involvement of trained experts. Moreover, it asks for pointed efforts for coming up with bench-mark standards which can be adopted in a uniform manner across the language types for creation of tagged corpora. However, anyone who wants to tag a text corpus will have to deal with the following two major issues:

(a) What kind of linguistic information should be tagged in the corpus, and
(b) How it should be tagged.

For the first question, it is envisaged that we should come up with a well-defined scheme which will allow us to tag various linguistic information. As stated above, linguistic information can be of *transcription, orthography, part-of-speech, morphology, grammar, syntax, etymology, anaphora, semantics, discourse, pragmatics, figure-of-speech,* and *sociolinguistics*. Although we may decide to tag only one type of information initially, we should be able to augment, as and when required, other types of information to the already tagged corpus with little effort. Therefore, it is argued that the tagging scheme of one type must support the other types. Moreover, there should be not curtailment or compromise with the amount of linguistic information to be tagged into the corpus. Keeping these issues in mind we have made an attempt to define a set of principles which we have followed at the time of POS tagging of a Bengali written text corpus.

7. Principles of POS Tagging

For most of the advanced languages, well-defined sets of principles are prepared and used to assign POS tags to words to settle their lexico-semantic entities as well as syntactico-grammatical functions in texts (Leech, Garside and Bryant 1994). Since such principles have not yet been developed and adopted for the Indian languages, there arises an urgent need for laying down some basic principles to be adopted and followed for designing standards for POS tagging for these languages. This leads us to formulate the following ten principles of POS tagging for the Indian languages corpora (as well as for Bengali) of all text types and genres (Dash 2011a)

Principle 1: Uniform tagset to be designed for all text types

The POS tagset should not be skewed towards any one particular type of text. It should work elegantly well for all text types of a language. That means a tagset designed for written texts should equally be useful for spoken texts. Similarly, a tagset developed for informative text, should be equally applicable for imaginative texts, scientific text, commercial texts, legal texts, mass media text, and other texts available in a language. Moreover, a tagset designed for prose texts should equally be applicable for poetic texts.

Principle 2: Spatio-Temporal dimension should be present in the tagset

A POS tagset should be compatible with existing the form and texture of a language-specific text as well as with old and new text materials available in a language. That means a POS tagset designed for the modern Bengali texts should equally be applicable for old Bengali texts and the texts produced in dialects or regional varieties of the language. Also, the tagset should work properly for the Bengali texts produced in West Bengal, India as well as for the texts produced in other states of the country and in Bangladesh, UK, USA, or any other part of the globe.

Principle 3: Layered Approach should be used for POS tagging

It is always better to adopt a layered approach for POS tagging, because a piece of natural language text is always full of varieties with intralinguistic and extralinguistic information. When information is tagged, the resultant text becomes a unique kind of text where linguistic information is made explicit to man as well as to machine. Therefore, it is better to tag as much linguistic information as possible within a given text corpus. However, since the linguistic information embedded in a text is quite complex, usually implicit, and diverse in nature, it is almost impossible to capture all kinds of linguistic information in one go. Therefore, it is better to break up the collective linguistic information into layers and tag each layer separately in the text in the following manner:

Layer 1 : Orthographic information (mostly required in spoken text transcription),
Layer 2 : Morphological information (morphological forms and elements of words),

Layer 3 : Part-of-speech information (morpho-syntactic roles of words),
Layer 4 : Lexicological information (related to origin, evolution and function of words)
Layer 5 : Local word grouping information (where two or more words constitute a single meaning),
Layer 6 : Chunking information (for compounds, reduplications, idioms, and collocations),
Layer 7 : Parsing information (syntactic analysis of phrases, sentences, thematic roles, arguments),
Layer 8 : Semantic information (identifying meaning of words and sense disambiguation),
Layer 9 : Anaphoric information (identifying nominal and pronominal referential relations),
Layer 10 : Pragmatic information (identifying pragmatic elements within texts),
Layer 11 : Discoursal information (identifying discourse elements within texts), and
Layer 12 : Sociolinguistic (identifying socio-cultural elements and cues within texts).

Since these layers are interdependent by nature, some layers cannot deal with isolated linguistic elements. They, however, can deal with words which are linked to each other within a text to perform varied functions. Also, the output information obtained from one layer becomes, in some cases, input information for the other layer. However, the degree of interdependence varies based on the type of text and the nature of linguistic information tagged at respective layers.

Principle 4: Hierarchical Formation of tagset is desired

It is always better to design a tagset keeping open the possibility of tagging more than one level of POS information to words at various tiers. This is possible only when the tagset has been designed in hierarchical format. For instance, we may include 'Noun' (N) as the top-level [Level_1] category and tag different sub-types of noun (e.g., *proper noun, common noun, collective noun, abstract noun, material noun,* etc.) as the lower-level categories [Level_2 and Level_3] under the top level, as the following table (Table 2) shows.

SN	Category			Label	Examples
	Level_1	Level_2	Level_3		
1	**Noun**			**N**	
1.1		Common		N_NNC	লোক (lok) 'man',
1.2		Proper		N_NNP	রবি (Rabi)
1.3		Material		N_NNM	কলম (kalam) 'pen'
1.4		Collective		N_NNL	দল (dal) 'party',
1.5		Abstract		N_NNA	ভয় (bhay) 'fear'
1.6		Verbal		N_NNV	গ্রহণ (grahaṇ) taking'
1.7		Nloc		N_NST	উপরে (upare) 'above'

Table 2: A hierarchical tagset proposed for Bengali nouns

Similarly, one can tag 'Verb' (V) as the top-level category, while 'main verb' (VM), 'auxiliary verb' (VAUX), 'finite verb' (VF), and 'non-finite verb' (VNF) may be tagged as the lover-level categories of the top-level 'Verb'.

Principle 5: Extensibility of Tagset is required for specific languages

The tagset should be extendable for including language specific requirements and for adding new tags, if required. It is an important property to make a tagset maximally flexible and adjustable for a language. It means a tagset should be such that if and when a new POS category is found, it can be incorporated in the existing tagset either at the top level or at the intermediate or lower level of the hierarchy, as the case may be. It also implies that if a particular tag is redundant for a language or a text type, it should be made redundant for that particular language.

Principle 6: Metadata Format should be used for tagset

All kinds of linguistic and extralinguistic information encoded and used in the tagset should be stored in the form of Metadata (e.g., *HTML, SGML, XML,* etc.) so that it becomes understandable and accessible by a computer program designed and applied for POS tagging (discussed in Section 9.1).

Principle 7: Dispensability of tagset from a text is mandatory

A POS tagged corpus should be easy to dispense with the tagset and revert quickly to the raw text. The raw corpus should be recoverable, if needed. This is an important pre-condition of POS tagging, as recovering the original raw text is essential for the works where tagged text has little applicational relevance (Leech 1993).

Principle 8: Standardization of tagset is a pre-requisition

A POS tagset should be designed in such a way that it can handle a wide range of applications of language technology and linguistics. That means it should support all linguistic research activities independent of a particular type or domain. Moreover, it should be applicable for various tasks of language description, grammar book writing, dictionary compilation, language teaching, and language planning, etc.

Principle 9: User Friendly tagset should be designed

The POS tagset should be user-friendly as far as it is possible and feasible. It is also an important feature because the primary purpose of a tagset is to tag a text – a task that is normally carried out by human beings, at least, at the initial stage of tagged text development in a language.

Principle 10: Non-ambiguity is must for POS tagset

The POS tagging involves decision making at every stage of the task. If a proposed tagset contains a number of tags, which can have multiple interpretations, this will lead to inconsistency and confusion in the subsequent stages of tagging. Therefore, it is suggested that all kinds of ambiguities should be avoided at the time of tagset designing.

We have followed all the principles stated above at the time of designing a tagset for the Bengali language as well as using this tagset for developing a POS tagged Bengali corpus.

After marking relevant information about sentences and segments within a Bengali corpus, we have tagged words in the corpus in the following manner (Fig. 2).

```
<paragraph>
<sentence>
এখানে ekhāne\DM_DMD\ দেওয়া deoyā\V_VM_VNG\ কিছু kichu\QT_QTF\ সহজ sahaj\JJ\
উপায়ের upāyer\N_NN\ মাধ্যমে mādhyame\PSP\ আপনি āpni\PR_PRP\ আপনার
āpnār\PR_PRF\ দাঁতকে d□tke\N_NN\ পরিষ্কার pariṣkār\JJ\ ওo\CC_CCD\ শ্বাসকে
śvāske\N_NN\ তাজা tājā\JJ\ রাখতে rākhte\V_VM_VINF\ পারবেন pārben\V_VM_VF\
.\RD_PUNC\
</sentence>
 </paragraph>
```

Fig. 2: POS tagging of words in a Bengali text corpus

Once the task of POS tagging comes to an end, the POS tagged corpus becomes ready for verification and authentication. The eventual tagged corpus may be used for chunking and extracting suitable patterns, rules, and features of lexical usage to be utilized for training a computer system for automatic tagging of other text corpora of a language.

8. The BIS POS Tagset for Bengali

The BIS tagset which is designed for all Indian languages also includes some of the principles stated above. In this section we critically evaluate the BIS tagset proposed for Bengali and analyse if this tagset needs to be modified to address many new observations obtained from the Bengali written text corpus.

The number of POS tagset proposed for tagging words in English corpus varies within a range of 50 to 160. For example, while the tag 'NN' stands for singular common nouns, 'NS' stands for plural common nouns, 'NP' stands for singular proper nouns, etc. On the other hand, for tagging parts-of-speech in a corpus of Koine Greek more than 1000 tags are used (DeRose 1990). Even then many words are found to be ambiguously tagged at the part-of-speech level as it has been observed in case of English corpora also.

For the Indian languages, it is not yet decided what will be the total number of tagset although the BIS tagset has proposed a set of 45 tags for all the Indian languages. Moreover, if is not yet settled if a fixed number of tags will suffice requirement for the Indian languages or there should be difference in the number of tagset based on linguistic features and part-of speech of different Indian languages.

Guided by tagging principles the BIS has proposed a super tagset for all Indo-Aryan languages including *Sanskrit, Hindi, Urdu, Punjabi, Gujarati, Rajasthani, Konkani, Marathi, Bengali, Oriya, Assamese*, and *Maithili.* Since these languages register close genealogical-cum-typological affinities, it is assumed that this super tagset may be useful for these languages with minimum language-specific modifications. It is, however, necessary to understand the rationale that works behind the decision of including various lexical categories and their sub-categories as suggested in the BIS tagset. In the following paragraphs we evaluate each top level category to understand why it is included and how it has to be interpreted by the annotators in an unambiguous manner at the time of tagging the Bengali corpus. Since the top-level categories of the BIS POS tagset constitute a generic scheme at the coarse form, only 12 basic types referring to the major parts-of-speech are proposed depending on the nature of a language. These are included at the top-level and a specific tag is assigned to each part-of-speech as the following list (Table 3) shows.

S.N.	POS TAG	Label	Examples
1	Noun	\N\	বালক (bālak), ঘর (ghar), শহর (śahar), কথা (kathā), etc.
2	Pronoun	\PR\	আমি (āmi), তুমি (tumi), সে (se), তারা (tārā), তুই (tui), etc.
3	Demonstrative	\DM\	যে (ýe), এই (ei), ওই (oi), তাই (tāi), etc.
4	Verb	\V\	করছি (karchi), করতাম (kartām), গেল (gela), যাবে (ýābe), ect.
5	Adjective	\JJ\	ভাল (bhāla), মন্দ (manda), সুন্দর (sundar), সাদা (sādā), etc.
6	Adverb	\RB\	কদাচিত (kadācit), বাবদ (bābad), কারণে (kāraṇe), etc.
7	Postposition	\PSP\	পরে (pare), কাছে (kāche), আগে (āge), নিচে (nice), etc.
8	Conjunction	\CC\	কিন্তু (kintu), অথবা (athabā), বরং (baraṃ), etc.
9	Particle	\RP\	ই (i), ও (o), তো (to), না (nā), নে (-ne), নি (-ni), etc.
10	Quantifier	\QT\	এক (ek), দুই (dui), প্রথম (pratham), পয়লা (paylā), etc.
11	Punctuation	\PUNC\	., : ;, ?, !, ( ), &, etc.
12	Residuals	\RD\	Foreign words, echo words, mathematical and geomatric symbols, unknown words and characters, etc.

Table 3: Top-level POS categories of the BIS tagset

All the top-level categories have several sub-categories following the hierarchical principle (Principle 4) adopted for the tagset. For example, the top-level category 'Verb' (V) has two broad sub-categories: main verb (VM) and auxiliary verb (VAUX). The main verb (VM) category has again four sub-categories: finite verb (VF), non-finite verb (VNF), infinitive (VINF), and gerund (VNG) at the lower level. The third sub-category may be included for specific cases with a view that a particular language may choose to keep that level optional.

8.1 Category 1: Noun (N)

The category Noun (N) is a top-level category with four lower-level categories in the hierarchy (Table 4): common noun (NN), proper noun (NNP), verbal noun (NNV), and noun-locative (NST).

Sl. No	Category			Label	Annotation Convention	Examples
	Top level	Subtype (level 1)	Subtype (level 2)			
1	**Noun**			**N**	**N**	
1.1		Common		NN	N_NN	কলম (kalam), চশমা (caśmā)
1.2		Proper		NNP	N_NNP	মোহন (Mohan), রবি (rabi), রশ্মি (rashmi)
1.3		Verbal		NNV	N_NNV	Not required for Bengali
1.4		Nloc		NST	N_NST	উপরে (upare), নিচে (nice), ভিতরে (bhitare)

Table 4: Tagset for major sub-categories for Bengali nouns

There is hardly any confusion with the BIS tagset that asks for making distinction between common nouns and proper nouns in the Indian languages (including Bengali) during POS tagging, as Indian languages, unlike English, lack in the practice of using capital letters in proper names. The lack of capitalization in Indian languages will generate difficulties for both man and machine to make distinction between the two types of noun. Therefore, it becomes necessary to develop an extensive guideline based on which we can tag words used as proper nouns in a piece of text. For instance, person names, place names, object names, item names, organization names, institute names, book names, country names, etc. are clearly proper nouns and these should be tagged as proper nouns only. However, difficulty may arise in case of multiword proper names, such as, পথের পাঁচালী (pather pãcālī), হাঁসুলি বাঁকের উপকথা (hãsuli bãker upakathā), মাউন্ট আবু (māunṭ ābu), জহর লাল নেহেরু (jahar lāl neheru), মোহনদাস করমচাঁদ গান্ধী (mohandās karamcãd gāndhī), মধুবনী চিত্রকলা (madhubanī citrakalā), ইন্ডিয়ান স্ট্যাটিস্টিক্যাল ইন্সটিটিউ (inḍiān sṭyāṭisṭikyāl insṭiṭiuṭ), নিউ দিল্লী (niu dillī), ভারত মহাসাগর (bhārat mahāsāgar), গাল্ফ অফ ওমান (gālph aph omān), etc. where each constituent word used to form the nouns is also used as separate lexical item. The frequency of use of such multiword nouns in a language like Bengali is very high, and this unique phenomenon cannot be ignored if one wants to have a full-proof or nearly full-proof POS tagset for Bengali (and for other Indian languages, for that matter). Since we need to tag these words as proper nouns (which would be capitalized in English), we need to develop a set of well-defined principles for achieving this goal through proper study of actual texts of the language.

Furthermore, although there is hardly any confusion with regard to marking common nouns (NN), confusions may arise in case of proper nouns (NNP). There can be debate if proper nouns (NNP) should be included as a separate sub-category in the scheme. For languages such as English the inclusion of this category is important as proper nouns are easily identifiable by orthographic cues like capital letters. Moreover, most of the personal proper nouns in English are the terms, which are actually used only as 'naming terms' for persons. On the other hand, it is well known that in Bengali, there is no specific cue in the script for identifying proper nouns. Moreover, nouns belonging to other sub-categories of nouns can also occur as proper nouns in Bengali, such as, সবিতা (sabitā), সুর্য (surýa), রূপা (rūpā), সোনা (sonā), আকাশ (ākāś), সমীর (samīr), সমুদ্র (samudra), সাগর (sāgar), মমতা (mamatā), মায়া (māyā), বিদ্যা (bidyā), পরি (pari), কমল (kamal), জীবন (jīban), পরশ (paraś), etc.

Furthermore, several adjectival forms can also be used as proper nouns in Bengali. For instance, adjectival forms like সুন্দরী (sundarī), রূপসী (rūpasī), শীতল (śītal), অসীম (asīm), মনোহর (manohar), নীল (nīl), লাল (lāl), কালো (kalo), প্রিয়া (priyā), পরম (param), নির্মল (nirmal), অমল (amal), অরূপ (arūp), বিপুল (bipul), সুধীর (sudhīr), চন্দ্রিল (candril), রূপালী (rūpālī), সোনালী (sonālī), রক্তিম (raktim), বঙ্কিম (baṅkim), চঞ্চল (cañcal), রঞ্জিতা (rañjitā), অঙ্কিতা (aṅkitā), বিনতা (binatā), বর্ণিল (barṇil), অচল (acal), সুনীল (sunīl), কোমল (komal), সুশীল (suśīl), etc. are often used as proper nouns in Bengali.

At the time of automatic POS tagging of words in Bengali corpus the common nouns and proper nouns tagged differently may create confusion that may eventually make the task of machine learning of proper nouns extremely difficult. Even then, we argue that since it is comparatively easier for the human annotators to identify proper nouns in the context of word use, these should be tagged in the corpus itself, and eventually these tags may be merged with common nouns while applying machine learning algorithms.

The third sub-category is verbal noun (NNV), which according the BIS, does not exist in Bengali. In our argument, this category can create confusions as it can be interpreted in several ways, including an alternative term: gerund. The basic points of argument are, therefore, to show how verbal noun is different from gerund, and whether verbal noun does exist in Bengali.

The existence of verbal noun as an important sub-part of the non-finite verb in Bengali is clearly attested in Chatterji (1995: 313), Shaidullaha (1967: 147), Majumdar (1993: 390, 414), Sarkar and Basu (1994: 200), and Thompson (2010:374). While Chatterji uses Bengali term ভাববচন (bhābbacan) or ক্রিয়াবাচক বিশেষ্য (kriyābācak biśeṣya), Shahidullaha calls it ক্রিয়াবাচক বিশেষ্য (kriyābācak biśeṣya), Majumdar identifies it as ভাববচন (bhābbacan), and Sarkar and Basu call it as অসমাপিকার ক্রিয়াবিশেষ্য (asamāpikār kriyā biśeṣya). Finally, Thompson informs, "The verbal noun is the form of verbs given in dictionaries and can therefore be considered the most basic of the non-finite verb forms. The verbal noun can be used like any other inanimate non-count noun. It can function as the subject of sentences. It declines for case and takes modifiers and classifiers but due to its inanimate status the objective case ending is rare. Verbal nouns have no plural forms" (Thompson 2010: 375). For understanding how verbal nouns are used in Bengali, let us consider the following underlined examples:

(3a) এটা আমার মরণ-বাঁচনের প্রশ্ন।
(eṭā āmār maraṇ-bãcaner praśna)
"It is a question of my life and death"

(3b) এই প্রতিজ্ঞা পালন করা সহজ নয়।
(ei pratijñā pālan karā sahaj nay)
"It is not easy to keep this promise"

(3c) তিনি দিন যাপনের গ্লানি নিয়ে বেঁচে আছেন।
(tini din ýapāner glāni niye bẽce āchen)
"He survives with the pain of passing days"

(3d) আজ সেই স্বপ্ন পূরণের সময় এসেছে।
(āj sei svapna pūraṇer samay eseche)
"The time has come for realizing that dream"

(3e) আজ তোমার যাওয়া বারণ।
(āj tomār ýāoyā bāraṇ)
"You are forbidden for going today"

The underlined forms in the above examples clearly show that verbal nouns are very much there in Bengali and therefore should be tagged properly in the corpus.

The fourth sub-type is Nloc (NST) which denotes particular nouns of location, including both space and time. This category is included to register distinctive nature of some locational nouns, which also function as a part of complex postpositions – an important phenomenon of the Bengali language. Certain expressions, such as, উপরে (upare) 'above', নিচে (nice) 'below', আগে (āge) 'before', সামনে (sāmne) 'in front of', etc., which are although used as postpositions, are actually content words that denote spatial and temporal senses. These expressions are also used in various ways and manners to denote sense variations (Dash 2010). For example, these forms may take place as temporal or spatial arguments of a predicate (i.e., verb) in a sentence taking appropriate বিভক্তি (vibhakti) 'case marker', as the following examples show.

(4a) সে উপরে কাজ করছিল।
(se upare kāj karchila)
"He was working upstairs"

(4b) সে আগেই সেখানে হাজির ছিল।
(se āgei sekhāne hājir chila)
"He was there beforehand"

(4c) তুমি এখন বাইরে বসো।
(tumi ekhan bāire baso)
"Now you sit outside"

Apart from functioning like arguments of a verb, these expressions can also act as nouns taking various postpositions, as the following examples show.

(4d) সে তখন উপর থেকে নেমে এল।
(se takhan upar theke neme ela)
"He then got down from upstairs"

(4e) সে আগে থেকেই ঘটনাটা জানত।
(se āge thekei ghaṭanāṭā jānta)
"He know the event beforehand"

(4f) সে এইমাত্র বাইরে থেকে ঘরে ফিরল।
(se eimātra bāire theke ghare phirla)
"He just returned home from outside"

Such complex postpositions ask for special care at the time of POS tagging. For tagging such words, one possible option is to tag them according to their syntactic functions in the context of their usage in sentences. For instance, in first three cases (4a-4c), the underlined words have occurred as postpositions or relation markers. They can, therefore, be marked as postpositions (PSP). On the other hand, in last three cases (4d-4f), these are used as nouns, therefore, should be marked as nouns (N) and not as postpositions. Or alternatively, since these words are more like nouns, they should be tagged as nouns in all their occurrences.

If we follow the above argument, then we shall add up on the fact that this class of words is slightly different from other nouns. These are a type of nouns which indicates location or time. At the same time, they also function as postpositions in certain contexts. If these words are tagged according to their syntactic function, it will increase efficiency in machine learning. Therefore, considering their special status, it is sensible to consider these words to belong to two separate sub-types by introducing two different tags: (a) NST (Nloc) for determining their temporal and spatial entities, and (b) PSP (postposition) to identify their postpositional entity. We also argue that if two different tags are kept for both the usages, the decision making activity of the annotators will be easier. Therefore, a new category (Nloc with the tag NST) is required to be introduced for marking specific syntactic roles of such expressions. The tag NST should be used for a finite set of such words used in Bengali and many other Indian languages[7].

8.2 Category 2: Pronoun (PR)

There is a clear necessity for including a separate top level category for Pronoun (PR), although linguistically, a pronoun is a variable, which acts as a function word and syntactically acts as NP in a piece of text. Therefore, it is better to have separate tags for pronouns that will help both man and machine in resolution of anaphora and referential ambiguities. Although the general argument is that a pronoun should be sub-classified under main category noun (N) rather than having a separate top-level category as Pronoun (PR), it is better to keep it at the top-level category as it is not a sub-type of noun but a variable, which can hold a value that need not necessarily be a noun. Based on this argument, the top-level category pronoun (PR) has been assigned in BIS with five sub-types. Given below (Table 5) is the list of basic sub-types of the top-level category: Pronoun (PR).

S.N	Category			Label	Annotation Convention	Examples
	Top level	Subtype (level 1)	Subtype (level 2)			
2	**Pronoun**			**PR**	**PR**	
2.1		Personal		PRP	PR_PRP	আমি (āmi), সে (se) তুমি (tumi), আমরা (āmrā)
2.2		Reflexive		PRF	PR_PRF	নিজেকে (nijeke)
2.3		Relative		PRL	PR_PRL	যে (ýe), যারা (ýārā), যাদের (ýāder), যাকে (ýāke)
2.4		Reciprocal		PRC	PR_PRC	পরস্পর (paraspar)
2.5		Wh-word		PRQ	PR_PRQ	কে (ke), কাকে (kāke), কারা (kārā), কাদের (kāder)

Table 5: The tagset for major sub-category of Bengali pronouns

(a) Personal pronoun: e.g., আমি (āmi), আমার (āmār), আমাকে (āmāke), আমরা (āmrā), আমাদের (āmāder), আমাদেরকে (āmāderke), তুমি (tumi), তোমার (tomār), তোমাকে (tomāke), তোমরা (tomrā), তোমাদের (tomāder), তোমাদেরকে (tomāderke), তুই (tui), তোর (tor), তোকে (toke), তোরা (torā), তোদের (toder), তোদেরকে (toderke), সে (se), তার (tār), তাকে (tāke), তারা (tārā), তাদের (tāder), তাদেরকে (tāderke), তিনি (tini), তাঁর (tā̃r), তাঁকে (tā̃ke), তাঁরা (tā̃rā), তাঁদের (tā̃der), তাঁদেরকে (tā̃derke), উনি (uni), উনার (unār), উনাকে (unāke), উনারা (unārā), উনাদের (unāder), উনাদেরকে (unāderke), etc.

(b) Reflexive pronoun: e.g., নিজেকে (nijeke), নিজের (nijer), etc.

(c) Relative pronoun: e.g., যে (ẏe), যাকে (ẏāke), যাদের (ẏāder), যারা (ẏārā), etc.

(d) Reciprocal pronoun: e.g., একে-অপরকে (eke-aparke), পরস্পরকে (parasparke), etc.

(e) Wh-words (which are actually interrogative pronouns), e.g., কে (ke), কারা (kārā), কাকে (kāke), কাদের (kāder), etc.

In the BIS tagset the form যেন (ẏena) is suggested to be tagged as a relative pronoun (PRL), while the form কেন (kena) is suggested to be tagged as a Wh-word (PRQ) within the top-level category of pronoun (PR). This argument appears to be incorrect as যেন (ẏena) is hardly used as pronoun, but used as conjunction or particle in Bengali, as the following examples show.

(5a) তোমার কথাই যেন সত্যি হয়।
(tomār kathāi ẏena satyi hay)
"Let your words be proved true"

(5b) যেন খবরটা আমি জানি না।
(ẏena khabarṭā āmi jāni nā)
"As if I do not know the news"

Secondly, even if we assume that কেন (kena) can be used as a Wh-pronoun, it is to be noted that this particular form is not a placeholder for a noun. Moreover, we should also keep in mind that কেন (kena) often acts differently in Bengali to denote different POS functions, and these different usages should be properly tagged according to the function of কেন (kena) keeping in mind the particular sentential context. For instance, consider the following sentence.

(5c) সে কাজটা করবে, কেন না সে এটা জানে।
(se kājṭā karbe, kena nā se eṭā jāne)
"He will do the work, because he know it"

The above example shows that কেন (kena) is used as a conjunction in the sense of 'because', while the following না (nā) is used in the sense of an emphatic particle. Both of them taken together may be tagged as multiword conjunction (similar to Hindi *kyon ki*). Therefore, it can be argued that both যেন (ẏena) and কেন (kena) should not be included in the list of pronouns, but be placed in the list of conjunction.

Another important piece of information is missed in the BIS tagset regarding the POS of the pronominal form কেউ (keu). Although it appears to a Wh-pronoun in surface form, it is not so, as it hardly denotes any question or raises any question about the identity of an individual, as the following examples show.

(6a) কেউ এ খবর জানে না।
(keu e khabar jāne nā)
"No one knows this news"

(6b) কেউ কি এ খবর জানে ?
(keu ki e khabar jāne?)
“Does anybody know this news?”

(6c) সেখানে কেউ ছিল না।
(sekhāne keu chila nā)
“No one was there”

The sentences given above show that the word কেউ (keu) is not used as Wh-pronoun, but as something else, perhaps in the sense of demonstrative (DM). Therefore, this form should be tagged differently, and not as a Wh-pronoun as argued in the BIS tagset. Moreover, at the time POS tagging of pronouns it has to be kept in mind that some pronominal forms can carry out special grammatical and syntactic functions hardly assigned to them in grammar and dictionary. For instance, in Bengali, the pronoun আমি (āmi) can also be used as a noun as the following example shows.

(7a) বাইরের আমির ভেতর আর একটা আমি লুকিয়ে রয়েছে।
(bāirer āmir bhetar ār ekṭā āmi lukiye rayeche)
“Another ‘I’ is hidden inside the external ‘I’”.

8.3 Category 3: Demonstratives (DM)

The category of Demonstrative (DM) is also included as a top-level category as it deserves to be identified in this manner (Table 6). There are two basic issues which need to be discussed in the context of classifying demonstratives as a top-level category: (a) demonstratives are often referred to as demonstrative pronouns, and (b) in Bengali, these are lexically ambiguous in form and meaning with some pronouns like ও (o), ওই (oi), যে (ýe), যেই (ýei), সে (se), সেই (sei), কোন (kona), etc. What is identified as a demonstrative (DM) may also be identified as a pronoun (PR) in specific syntactic functions exerted by a lexical unit of this category.

Sl. No	Category			Label	Annotation Convention	Examples
	Top level	Subtype (level 1)	Subtype (level 2)			
3	**Demonstrative**			**DM**	**DM**	
3.1		Deictic		DMD	DM_DMD	এ (e), এই (ei), সে (se), সেই (sei), ও (o), ওই (o)
3.2		Relative		DMR	DM_DMR	যে (ýe), যেই (ýei) যাহা (ýāhā), যা (ýā)
3.3		Wh-word		DMQ	DM_DMQ	কোনো (kono) কোন (kona)

Table 6: Tagset for major sub-categories of Bengali demonstratives

This may lead us to argue that it is better to put demonstratives in the sub-category of pronouns. But this should not be done, because in our view, the term 'demonstrative' actually refers to those independent demonstrative forms, which are known as demonstrative determiners. In many languages these are distinct lexical items having specific grammatical functions. Grammatically also these forms do not always behave like pronouns. In fact, conceptually, demonstratives are actually specifiers of nouns and NOT variables, while pronouns are variables as grammatically they can take nominal inflections. So demonstratives are a distinct grammatical category and thus they need to be classified separately (Table 6). Similar to pronouns (PR), demonstratives (DM) have three different sub-types:

(a) Diectic[8] demonstratives (DMD), e.g., এ (e), এই (ei), ও (o), ওই (oi), সে (se), সেই (sei), etc.
(b) Relative demonstratives (DMR), e.g., যে (ẏe), যেই (ẏei), etc.
(c) Wh-demonstratives (DMQ) e.g., কোনো (kono), কোন (kona), etc.

The other issue that comes into serious discussion with demonstratives is that since these are functionally similar to adjectives, it is sensible to classify them under adjectives instead of classifying them as a separate top-level category. This argument also becomes nullified when it is realized that although demonstratives are noun modifiers like adjectives, these are mere specific pointers for nouns, and hence, they do not add any property or value to nouns. Since demonstratives are a small set of lexical items in Bengali, keeping these forms under a separate top-level category is a much better proposition in further schemes of sentence parsing and text analysis.

8.4 Category 4: Verb (V)

The top-level category of Verb (V) has two major sub-categories. While the first level sub-category distinguishes between the main verb (VM) and the auxiliary verb (VAUX), the second level sub-category is based on verbal inflections that denote features like finiteness, non-finiteness, infiniteness, etc. Moreover, gerunds (VNG) are classified at this level as a sub-type of the main verb (VM).

(a) Finite verb (VF) sub-category includes verbs found in complete inflected forms with a sense of completeness of action, such as, করেছিল (karechila), করেছিলাম (karechilām), গেল (gela), দিলাম (dilām), etc.
(b) Non-finite verb (VNF) includes verbs which are inflected but denote a sense of incompleteness in action, such as, করে (kare), খেয়ে (kheye), গিয়ে (giye), বলতে (balte), করিয়া (kariyā), খাইয়া (khāiyā), etc.
(c) Infinitive verb (VINF) includes forms like করাতে (karāte), খেতে (khete), গেলে (gele), etc.
(d) Gerunds (VNG) includes forms like যাওয়া (ẏāoyā), আসা (āsā), দেখা (dekhā), খেলা (khelā), করা (karā), etc.
(e) Auxiliary verb (VAUX) includes forms like ছিল (chila), আছে (āche), থাকবে (thākbe), হবে (habe), etc.

The table below (Table 7) shows the sub-types and their tagset for Bengali verbs. What appears from the BIS tagset is that sub-categorization of verbs for Bengali is quite confusing.

Sl. No	Category			Label	Annotation Convention	Examples
	Top level	Subtype (level 1)	Subtype (level 2)			
4	**Verb**			**V**	**V**	
4.1		Main		VM	V_VM	
4.1.1			Finite	VF	V_VM_VF	যাবো (ýābo), গেল (gela), খায় (khāy), করেছিলাম(karechilām),
4.1.2			Non-finite	VNF	V_VM_VNF	করে (kare), করলে (karle), গেলে (gele), খেয়ে (kheye), খেতে (khete)
4.1.3			Infinitive	VINF	V_VM_VINF	করাতে (karāte), যেতে (ýete), খেতে (khete)
4.1.4			Gerund	VNG	V_VM_VNG	যাওয়া (ýāoyā), আসা (āsā), পড়া (paṛā), দেখা (dekhā), খেলা (khelā), করা (karā), বলা (balā),
4.2		Auxiliary		VAUX	V_VAUX	ছিল (chila), হবে (habe), চাই (cāi), আছে (āche)

Table 7: Tagset for major sub-categories of Bengali verbs

While there is little doubt regarding the classification of verbs into two broad categories like 'main' and 'auxiliary' (because it does not go against the traditional grammar of the language), it has to take into serious consideration the issue of identification and differentiation between a main verb (VM) and an auxiliary verb (VAUX), in Bengali where the difference between the two types of verb is quite fuzzy. Unlike English, an auxiliary verb in Bengali can take inflections normally tagged with a main verb to generate a final form that will act as a main verb. That means the finiteness and other morphological properties may be present on the auxiliaries, and not on the main verb. For illustration, consider the following examples:

(8a) তুমি তাকে বইটা দিতে গেলে।
(tumi tāke baiṭā dite gele)
"You went to give him the book"

(8b) সে তাকে বইটা দিয়ে দিল।
(se tāke baiṭā diye dila)
"You gave him away the book"

(8c) আমি তোমায় বইটা দিয়ে দিতে পারি।
(āmi tomāy baiṭā diye dite pāri)
"I can give away the book"

(8d) বইটা তাকে দেওয়া যেতে পারে।
(baiṭā tāke diye deoyā ẏete pāre)
"The book can be given away to him"

If we look at the examples given above, we can see that in the first sentence (8a), the form দিতে (dite) is a non-finite auxiliary verb, while the form গেলে (gele) is a finite main verb; in the second sentence (8b), the form দিয়ে (diye) is a non-finite main verb, while দিল (dila) is an auxiliary finite verb; in the third sentence (8c), দিয়ে (diye) is a non-finite main verb, দিতে (dite) is a non-finite auxiliary verb, while পারি (pari) is a finite auxiliary verb; and in the fourth sentence (8d), দেওয়া (deoyā) is an auxiliary main verb in gerundial form, যেতে (ẏete) is a non-finite auxiliary verb, and পারে (pāre) is a finite auxiliary verb. Therefore, at the time of POS tagging, which is mostly done at the word level, it is strongly recommended that there should be clear-cut distinctions between finite, non-finite, infinitive, and gerunds for both main and auxiliary verbs in Bengali, and these distinctions should be followed and marked accordingly at the time of POS tagging.

Although it is necessary to make distinctions between verbal noun and gerund in the POS tagset, we are not sure if Bengali lacks in verbal noun but has gerund (as proposed in BIS tagset) or it has both the sub-categories. Grammatically, while verbal nouns behave more as nouns, gerunds behave more like verbs, although both can use nominal case markers. Therefore, verbal nouns should be viewed differently from gerunds on the ground that although both are nouns derived from verbs, while gerunds retain their verbal properties and carry arguments with proper grammatical case, verbal nouns often fail in this function. For example, in Bengali the form পড়ানো (paṛāno) is a gerund, because it can function as a noun and can take appropriate case markers, as the following example show.

(8e) তোমার <u>পড়ানোর</u> ধরনটা আমার ভালো লাগে।
(tomār paṛānor dharanṭā āmār bhāla lāge)
"I like the style of your teaching"

The above example (8e) shows that the verb পড়ানো (paṛāno) in Bengali is a gerund, because it takes arguments syntactically, i.e., it takes arguments and these arguments have appropriate syntactic case markings. Therefore, gerunds are conceptually different from verbal nouns because gerunds can take arguments. They are derived from verbs and they are more like verbs.

Moreover, it is often noted that the finiteness and non-finiteness of a verb form is not always determined by its formative elements or by other morphological markers. Since the same set of inflections can be tagged with both finite and non-finite verbs in Bengali, one has to be careful in case of using POS tag to a particular verb form. For instance, in Bengali forms like করে (kare), করলে (karle), করতে (karte), বলে (bale), বলতে (balte), বললে (balle), গেলে (gele), দিতে (dite), দিলে (dile), etc. can either be finite verb form or non-finite verbs based on the context of their usage in the sentence. Therefore, it is absolutely necessary to look at the sentence and the role of a verb in the sentence before the verb is tagged either as a finite or a non-finite one.

Further complications will arise in case of tagging complex predicates in those verb forms where verbs are formed by using two separate word strings, particularly in the following cases:

(a) নাম ক্রিয়া (nām kriyā = conjunct verb): e.g., উপকার করল (upakār karla), হাজির হল (hājir hala), মুখস্থ করত (mukhastha karta), প্রণাম করল (praṇām karla), প্রশ্ন করবে (praśna karbe), উত্তর দেবে (uttar debe), হাত তুলবে (hāt tulbe), etc.

(b) যৌগিক ক্রিয়া (ýaugik kriyā= compound verb): e.g., বলতে থাকল (balte thākla), শুনতে পেল (śunte pela), লেগে পড়ল (lege paṛla), উঠে পড়ল (uṭhe paṛla), শুয়ে পড়বে (śuye paṛbe), জেনে রেখো (jene rekho), কামিয়ে নিল (kāmiye nila), বলে গেল (bale gela), etc.

(c) ক্রিয়া দ্বন্দ্ব (kriyā dvandva = verbal compound): e.g., আসতো-যেতো (āsto-ýeto), দেখেতে-শুনতে (dekhte-śunte), লিখতে-পড়তে (likhte-paṛte), জেনে-শুনে (jene-śune), চলতে-ফিরতে (calte-phirte), ধরত-মারত (dharta-mārta), etc.

In case of POS tagging of conjunct verbs, one has to decide whether the first word should be tagged as a noun and the second word as a main verb, or the first word as a main verb and the second word as a finite auxiliary verb. Similarly, in case of compound verbs, one has to decide whether the first word is a non-finite main verb and the second word is a finite auxiliary verb; and in case of verbal compounds one has to determine if both words should be tagged as finite main verbs.

Negative verbs like নয় (nay), নই (nai), নও (nao), নোস (nos), নাই (nāi), নেই (nei), নন (nan), নহে (nahe), নহ (naha), নহেন (nahen), etc. should be put under separate sub-category of verbs in Bengali.

8.5 Category 5: Adjective (JJ)

There is no doubt that Adjectives used in Bengali texts should be assigned with a separate top-level category (JJ) as these words (e.g., সুন্দর (sundar), ভাল (bhāla), লাল (lāl), কালো (kālo), মন্দ (manda), বড় (baṛa), মহত (mahat), দয়ালু (dayālu), গভীর (gabhīr), নির্জন (nirjan), স্বাধীন (svādhīn), মায়াবী (māyābī), etc.) are used in a piece of text with specific grammatical function and semantic role. This argument leads us to assign adjectives as a separate top-level category as 'adjective' (JJ). In Bengali, adjectives usually occur immediately before (and rarely after) nouns in sentence to determine, specify or qualify some properties of nouns.

The BIS tagset, however, does not specify how the participial adjectives derived from verbs should be tagged in the texts. Should these forms be tagged as adjectives (JJ) or as verbs (VM)? There are large numbers of such words in Bengali, such as, ঘুমন্ত (ghumanta), চলন্ত (calanta), ছুটন্ত (chuṭanta), জ্বলন্ত (jvalanta), উঠতি (uṭhti), পড়তি (paṛti), পতিত (patita), ধৃত (dhṛta), হৃত (hṛta), etc., which should be tagged as adjectives (JJ) based on their roles in the sentence where these are used.

8.6 Category 6: Adverb (RB)

In general, the term Adverb (RB) refers to a set of lexical items, which usually modify the action of a verb to denote mode of operation of the main or the supporting verb, such as, ধীরে (dhīre), আস্তে (āste), জোরে (jore), দ্রুত (druta), etc. However, in the present BIS tagset the term adverb (RB) refers to only the manner adverbs, such as, তাড়াতাড়ি (tāṛātāṛi), হঠাত্‌

(haṭhāṯ), দৈবাত্ (daibāṯ), কদাচিত্ (kadāciṯ), বিশেষভাবে (biśeṣbhābe), যথাযথ (ýathāýatha), মূলত (mūlata), বিশেষত (biśeṣata), etc. The adverbs of time and space are not included in the BIS tagset, as these lexical items can be tagged in different ways with postpositions.

The BIS tagset does not speak anything specifically about how words like আজ (āj) 'today', কাল (kāl) 'tomorrow', পরশু (parśu) 'day after tomorrow', এখন (ekhan) 'now', তখন (takhan) 'then', etc. should be tagged in the texts. Although these words are used quite regularly as adverbs in Bengali, the BIS tagset does not give any scope to tag these forms accordingly, as the tagset includes the manner adverbs only. Moreover, it is to be noted that these adverbial forms are also distinct from noun locatives (NST), and hence, these cannot be tagged as postpositions in the texts. Besides, these forms can also be used as nouns in the text. So there should be provision in the BIS tagset to tag these forms in proper manner.

Furthermore, the BIS tagset fails to refer to the cases where apparent adverbial forms are actually used as nouns in Bengali as the underlined words of the following examples show:

(9a) তোমার হিসেবে অনেক গোলমাল আছে।
(tomār hisebe anek golmāl āche)
"There are discrepancies in your accounts"

(9b) সে সম্বন্ধে আমার কোনো ধারণা নেই।
(se sambandhe āmār kono dhāraṇā nei)
"No idea about that matter"

(9c) সম্পর্কে সে আমার ছোট ভাই।
(samparke se āmār choṭa bhāi)
"In relation, he is my younger brother"

(9d) না আসার কোনো কারণ ছিল না।
(nā āsār kono kāraṇ chila nā)
"There was no reason for not coming"

(9e) স্থান বদলের কোন সুযোগ নেই।
(sthān badaler kona suýog nei)
"No scope for changing places"

(9f) এ বিষয়ে আমি কিছু জানি না।
(e biṣaye āmi kichu jāni nā)
"I know nothing about this"

If we look at the examples given above, we can observe that the words হিসেবে (hisebe), সম্বন্ধে (sambandhe), সম্পর্কে (samparke), কারণ (kāraṇ), বদলের (badaler), and বিষয়ে (biṣaye) are actually used as nouns in the sentences although these are often treated as adverbial forms in the language. If we adhere to BIS tagset, we have to tag these forms as N_NST (in NLoc) showing that these are functioning as adverbs denoting space, time, situations, etc. However, in the sentences mentioned above these are not functioning as adverbs but acting as nouns. Therefore, clear options needs to be provided for tagging these forms as nouns in the corpus.

8.7 Category 7: Postposition (PSP)

Postpositions used in Bengali carries special linguistic advantages in understanding the language in its present form. These are functionally similar to prepositions used in English but while prepositions are used before nouns and pronouns in English, postpositions in Bengali are usually placed after nouns and pronouns in sentence, and in most cases, nouns and pronouns are found to belong to possessive or genitive case. Although, according to scholars, Bengali postpositions are not a closed word class, these are not, in fact, quite large in number. In Bengali, there are nearly hundred postpositions, most of which are derived from nouns and verbs. It is, therefore, useful to treat postpositions as a separate word class in Bengali because many of the locative nouns or perfective participles change or expand their meaning in their use as postpositions (Thompson 2010: 229).

Functionally, majority of Bengali postpositions belong to postpositional or adverbial phrases so that they can act as functional adverbs in sentence—a very common pattern of their usage noted in modern Bengali texts. Since postpositions are functionally capable of revealing spatial, temporal, situational, locational, directional, and conditional information, (e.g., আগে (āge), পিছনে (pichane), থেকে (theke), অবধি (abadhi), কাছে (kāche), দুরে (dure), দিকে (dike), নিচে (nice), উপরে (upare), মাঝে (mājhe), সামনে (sāmne), পাশে (pāśe), দিয়ে (diye), পর্যন্ত (paryַanta), নাগাদ (nāgād) etc.,) they are rightly assigned with a separate top-level lexical category as Postposition (PSP).

8.8 Category 8: Conjunctions (CC)

In the BIS tagset Conjunction (CC) has been assigned with separate top-level category with two sub-categories as shown below:

(a) Coordinative conjunction (CCD), e.g., আর (ār), বা (bā), এবং (ebaṃ), অথবা (athabā), কিংবা (kiṃbā), etc. and

(b) Subordinative conjunction (CCS), e.g., কিন্তু (kintu), নইলে (naile), নতুবা (natubā), নচেত্ (nacheṯ), নাহলে (nāhale), বরং (baraṃ), ছাড়া (chāṛā), etc.

These are actually **indeclinables**, which have specific lexico-syntactic functions in the Bengali language. The subordinators have a further sub-type known as quotative (UT), which occurs in many languages and has the role of conjoining a subordinate clause to the main clause. However, it can be left optional to Bengali as this sub-category does not have any specific linguistic entity or function in the language (Table 8).

In our argument, indeclinables should be in the list of conjunctions because in Bengali language indeclinable are unique lexical items with unique linguistic relevance, independent meaning, and well-defined grammatical function. However, it is to be kept in mind that in some extreme situation an indeclinable can also be used as a noun. For instance, the form কিন্তু (kintu) can be used as a noun as the following example shows.

(10a) আর কোনো কিন্তুর ব্যাপার নেই
(ār kono kintur byāpār nei)
"There is no question of 'but' any more"

In this case, at least, the form কিন্তু (kintu) needs to be tagged as a noun (N), and not just as a subordinative conjunction (CCS).

Sl. No	Category			Labe l	Annotation Convention	Examples
	Top level	Subtype (level 1)	Subtype (level 2)			
8	**Conjunction**			**CC**	**CC**	
8.1		Coordinative		CCD	CC_CCD	আর (ār) এবং (ebaṃ)
8.2		Subordinative		CCS	CC_CCS	কিন্তু (kintu) নইলে (naile)
8.2.1			Quotative	UT	CC_CCS_UT	Not required

Table 8: Tagset for the category of Bengali conjunctions

8.9 Category 9: Particles (RP)

The category Particle (RP) is also a top-level category, which includes classifier (CL), intensifier (INFT), interjection (INJ), and negation (NEG) as its sub-types. The inclusion of sub-categories such as 'classifier' and 'negation' under top-level category of particle (RP) may generate debate, since the term 'particle' is normally defined as a category, which is known as 'indeclinables'. However, since indeclinable are already included in the top-level category of conjunction (CC), it appears sensible to include all these function words in the top-level category of particle (RP). This category includes five sub-categories such as the followings (Table 9):

Sl. No	Category			Label	Annotation Convention	Examples
	Top level	Subtype (level 1)	Subtype (level 2)			
9	**Particles**			**RP**	**RP**	
9.1		Default		RPD	RP_RPD	তো (to), যে (ẏe)
9.2		Classifier		CL	RP_CL	খানা (khānā), খানি (khāni)
9.3		Interjection		INJ	RP_INJ	আরে (āre), হায় (hāy)
9.4		Intensifier		INTF	RP_INTF	খুব (khub), অতি (ati), ভীষণ (bhīṣaṇ)
9.5		Negation		NEG	RP_NEG	না (nā), নি (ni)

Table 9: Tagset for the category of Bengali particles

(a) Default Particles: তো (to), ই (i), ও (o), যে (ẏe), etc.

(b) Classifiers: জন (jan), খানা (khānā), খানি (khāni), খানিক (khānik), টুকু (ṭuku), টুকুনি (ṭukuni), etc.

(c) Interjections: আরে (āre), এই (ei), হায় (hāy), etc.

(d) Intensifiers: ভীষণ (bhīṣaṇ), খুব (khub), সাঙ্ঘাতিক (sāṅghātik), অতি (ati), মারাত্মক (mārātmak), etc.

(e) Negative Particles: না (nā), নে (ne), নি (ni), etc.

In our argument, it is necessary to expand the scope of particles (RP), since moving of these sub-categories to other top-levels will create larger set of sub-categories for other top levels and this may pose serious problems in some NLP applications, such as, language analysis and sentence parsing.

With regard to intensifiers (INTF) it is to be noted that in Bengali, most of the intensifiers have potentials to be used as adjectives or noun modifiers. For instance, forms like খুব (khub), ভীষণ (bhīṣaṇ), etc. can also be used as adjectives, such as, খুব মেজাজ (khub mejaj) 'strong mood', ভীষণ বিপদ (bhīṣaṇ bipad) 'great danger', etc. In these cases, at least, these words cannot be tagged just as intensifiers (INTF) but should be tagged as adjectives (JJ) according to their contextual functions within a sentence.

8.10 Category 10: Quantifier (QT)

Although Quantifiers (QT) are usually identified as cardinal and ordinal adjectives in the general discussion of grammar of a language, these are assigned a separate top-level category in the BIS tagset as these forms need to be tagged separately for various works of mainstream linguistics and language technology. Within a piece of text, in general, one can come across three different types of quantifiers (Table 10), which may be put under three different sub-categories:

Sl. No	Category			Label	Annotation Convention	Examples
	Top level	Subtype (level 1)	Subtype (level 2)			
10	**Quantifiers**			**QT**	**QT**	
10.1		General		QTF	QT_QTF	কিছু (kichu), অল্প (alpa)
10.2		Cardinals		QTC	QT_QTC	এক (ek), দুই (dui)
10.3		Ordinals		QTO	QT_QTO	প্রথম (pratham) দোসরা (dosrā)

Table 10: Tagset for the sub-category of Bengali quantifiers

(a) General Quantifiers, e.g., কিছু (kichu), অল্প (alpa), অনেক (anek), সামান্য (sāmānya), কতক (katak), বহু (bahu), সব (sab), সকল (sakal), সমস্ত (samata), হরেক (harek), etc.

(b) Cardinal Quantifiers, e.g., এক (ek), দুই (dui), তিন (tin), চার (cār), পাঁচ (pā̃c), ছয় (chay), অর্ধেক (ardhek), আধ (ādh), আড়াই (āṛāi) etc.

(c) Ordinal Quantifiers, e.g., প্রথম (pratham), দ্বিতীয় (dvitīya), তৃতীয় (tṛtīya), চতুর্থ (caturtha), পয়লা (paylā), দোসরা (dosrā), তেসরা (tesrā), চৌঠা (cauṭhā), etc.

8.11 Category 11: Punctuation (PUNC)

The top-level category of Punctuation (PUNC) in the BIS tagset includes punctuation marks used in the Bengali text. In Bengali one comes across many punctuation marks used in the text in various ways and manners, and in some contexts these play vital role in identification of sentence boundaries as well as in deciphering actual meaning of words and sentences (Dash 2011: 201-220). Therefore, these unique orthographic symbols (e.g., *pūrṇachhed, comma, semicolon, colon, exclamation mark, question mark, dash, ellipses, hyphen, brackets,* etc.) need special treatment in text at the time of POS tagging for works of automatic text segmentation, word form recognition, morphological processing, machine learning, and language teaching.

8.12 Category 12: Residuals (RD)

Besides these top-level categories found in a text, there are also many other orthographic signs and symbols, which also need to be identified and assigned specific POS values. For instance, based on the nature and source of text one can come across in a piece of written text, various textual and orthographic elements, such as, *foreign words written in foreign script, echo words, mathematical symbols and signs, chemical formulae, unknown letters and characters* (e.g., @, #, $, %, &, <, >, +, =, etc.), *pointers*, etc., which need to be properly identified and tagged accordingly. All these unique characters and words are put under the 'residual' (RD) – a top level category (Table 11).

Sl. No	Category			Label	Annotation Convention	Examples
	Top level	Subtype (level 1)	Subtype (level 2)			
12	**Residuals**			**RD**	**RD**	
12.1		Foreign word		RDF	RD_RDF	Word written in script other than Bengali
12.2		Symbol		SYM	RD_SYM	$, &, *, (,), etc
12.3		Unknown		UNK	RD_UNK	
12.4		Echo-words		ECH	RD_ECH	টল (ṭal) in জলটল (jal-ṭal), টই (ṭai) in বইটই (bai-ṭai)

Table 11: Tagset for residual forms used in Bengali texts

The critical analysis presented above implies that after identification and finalization of tagset and their general acceptance in a language, one can go for using this tagset for tagging words within a text corpus, elaborated in some details in the following section (Section 9).

9. Tagging a Bengali Text with POS Tags

Tagging of a text may be done either manually or automatically. In both cases, it has to be done starting with the lowest level tagset of the hierarchy. Once the lower level tag is selected and assigned, the higher level tags should be identified and tagged automatically. Although tagging may be done with a clear focus on part-of-speech (POS) of words, the long term goals should not be ignored for developing a generic scheme, which will be useful for incorporating all kinds of linguistic information easily at the later stages. Therefore, we argue for a judicious decision for tagging words in texts with the following types of information following the universally accepted norms and rules of marking the metadata within a text.

9.1 Marking Metadata in a Text

Since a language corpus may consist of several kinds of text, it is important to maintain and preserve some meta-level information about each text document included in the corpus (Dash 2011c). Thus, information regarding *language, title, author, source, domain, text type, time, creator of digital text document,* etc. should be marked on each text within a Header File as 'Metadata'. This task should be done manually in the following manners (Table 12).

<Header type>	Information
<Language >	Bengali
<Language Type >	Written Text
<Category>	Aesthetics
<Subcategory>	Literature-Novel
<Text Type>	Imaginative
<Source Type>	Book
<Title>	ভুত আর ভুতো (bhut ar bhuto)
<Volume>	Single
<Issue>	Not Applicable
<Edition>	First
<Headline>	ভুত আর ভুতো (bhut ar bhuto)
<Author>	শুধাংশু পাত্র (Sudhanshu Patra)
<Publisher>	Dey's Publishing
<Publication Place>	Kolkata, India
<Publication Date>	1993
<Index No.>	B0035
<Data Creator Code>	61802
<Date of Creation>	12. 09. 2006
<Data Imputer/Collector>	Anami Sarkar
<Proof Reader>	Aprakash Gupta
<Date of Proofreading>	16. 08. 2007
<Total no of Words>	5017

Table 12: Header File and its information annotated with metadata

The information stored in the **Header File** actually contains extralinguistic information, which becomes necessary and useful for maintaining records of documents, dissolving copyright problems, and for carrying out research in sociolinguistics, language planning, stylistics, and discourse.

9.2 Marking Paragraph and Sentence Boundaries

After completion of the work of metadata preservation in the Header File, the next stage starts with marking the paragraph boundaries, which may be done in the following manners (Fig. 3):

<p>	ভূতো– আমাদের ভূতো বাবু!	</p>
<p>	ঐ যে ছেলেটা – যার দুষ্টুমিভরা ডাগর দুটি চোখ, যার মুখে সব সময় কথার খৈ ফোটে, যার হাত পায়ের বিরাম থাকে না কোন সময়, যে ছড়া বলতে ভালবাসে, গপ্পো শুনতে আরও ভালবাসে, ইস্কুলে রিণা মিনা নান্টু-মিন্টুদের সাথে ছবি ও ছড়ার বই পড়ে, সেইই ভূতো – আমাদের ভূতোবাবু।	</p>
<p>	ছেলেবেলা থেকেই ভূতোর উপর ভূতের নজর। যখন সে একমাসের শিশু – তখনই সে উপরের দিকে কটমট করে তাকাতো, ঘুমিয়ে ঘুমিয়ে আপনিই ফোকলা মুখে হাসতো, হাত মুঠো করে কাঁদতো আর পা ছুঁড়তো, কখনও বা চমকে উঠতো। ওঁয়া ওঁয়া করে কাঁদতে শুরু করলে সহজে কান্না থামতো না।	</p>

Fig. 3: Paragraph boundaries marked in a Bengali text

Marking sentences and segments of sentences with their respective boundaries within a paragraph is the next stage. Here full and complete sentences are marked with a tag called as <sentence>, while incomplete sentences and set phrases are marked with a different tag called as <segment> in the following ways (Fig. 4).

<segment> আপনার দাঁতের যত্ন </segment>

<sentence> তাজা শ্বাস আর ঝকঝকে দাঁত আপনার ব্যাক্তিত্বকে আকর্ষণীয় করে। </sentence>

<sentence> এখানে দেওয়া কিছু সহজ উপায়ের মাধ্যামে আপনি আপনার দাঁতকে পরিষ্কার ও শ্বাসকে তাজা রাখতে পারবেন । </sentence>

<segment> দক্ষিণ ভারত ভ্রমণ </segment>

<sentence> দক্ষিণ রেলওয়ের চেন্নাই স্টেশন থেকে ধনুষ্কোট যাওয়ার পথে প্রধান লাইনে চেন্নাই থেকে ৩৫ মাইল দূরে চঙ্গলপেট স্টেশন পড়ে। </sentence>

<sentence> চঙ্গলপেট স্টেশন থেকে একটা লাইন অরকোনাম অবধি চলে গেছে । </sentence>

Fig. 4: Segment and sentence boundaries marked in a Bengali text

There can be cases where a text has words in a language other than the matrix language of the text. For example, a text in Bengali may contain English words written in Roman script. From processing point of view, it is important to mark this information as well. All of the above information can be marked manually or automatically to a large degree of accuracy. However, the task of the manual annotators is to check and correct the marked information, wherever necessary.

9.3 Marking Words in Texts

After marking relevant information about the sentences and segments within a text, we can begin to mark POS tag to words used in the text. Here, following the BIS standard or any other standard devised for the purpose, we can POS tag a Bengali text in the following ways (Fig. 5).

```
<p>
<sentence> তাজা\JJ\ শ্বাস\N_NN\ আর\CC_CCD\ ঝকঝকে\JJ\ দাঁত\N_NN\ আপনার\PR_PRP\
ব্যাক্তিত্বকে\N_NN\ আকর্ষণীয়\N_NN\ করে\V_VM_VF\ । \RD_PUNC\ </sentence>
<sentence> এখানে\DM_DMD\ দেওয়াV_VM_VNG\ কিছু\JJ\ সহজ\JJ\ উপায়ের\N_NN\
মাধ্যামে\PSP\ আপনি\PR_PRP\ আপনার\PR_PRF\ দাঁতকে\N_NN\ পরিষ্কার\JJ\ ও\CC_CCD\
শ্বাসকে\N_NN\ তাজা\JJ\ রাখতে\V_VM_VINF\ পারবেন\V_VM_VF\ । \RD_PUNC\ </sentence>
<sentence> দক্ষিণ\N_NN\ রেলওয়ের\N_NN\ চেন্নাই\N_NNP\ স্টেশন\N_NN\ থেকে\PSP\
ধনুষ্কোট\N_NNP\ যাওয়ার\V_VM_VNG\ পথে\N_NN\ প্রধান\N_NN\ লাইনে\N_NN\ চেন্নাই\N_NNP\
থেকে\PSP\ ৩৫\QT_QTC\ মাইল\N_NN\ দূরে\N_NN\ চঙ্গলপেট\N_NNP\ স্টেশন\N_NN\
পড়ে\V_VM_VF\ । \RD_PUNC\.</sentence>
<sentence> চঙ্গলপেট\N_NNP\ স্টেশন\N_NN\ থেকে\PSP\ একটা\QT_QTC\ লাইন\N_NN\
অরকোনাম\N_NNP\ অবধি\PSP\ যায়\V_VM_VF\ । \RD_PUNC\ </sentence>
<sentence> যদি\CD_CCS\ আমরা\PR_PRP\ কোনো\DM_DMQ\ মানুষকে\N_NN\ অপারেশন\N_NN\
টেবিলে\N_NN\ অজ্ঞান\N_JJ\ করে\V_VM_VNF\ করাতের\N_NN\ দ্বারা\PSP\ তার\PR_PRP\
মাথার\N_NN\ উপরের\N_NN\ ভাগটা\N_NN\ ধীরে\RB\ ধীরে\RB\ কেটে\V_VM_VNF\ আলাদা\N_JJ\
করে\V_VM_VNF\ দিই\V_VM_VF\ তবে\CD_CCS\ আমরা\PR_PRP\ নিজের\PR_PRF\ চোখে\N_NN\
একটা\QT_QTC\ জ্যান্ত\N_JJ\ মস্তিষ্ককে\N_NN\ দেখতে\V_VM_VINF\ পাবো\V_VM_VF\
।\RD_PUNC\</sentence>
</p>
```

Fig. 5: POS tagged words within sentences within a Bengali text

Once this tagging task comes to an end, a tagged corpus becomes ready for manual verification and authentication. After this stage is complete, the eventual tagged corpus can be used for chunking as well as for extracting suitable patterns, rules, and features to be used for training a computer system for automatic tagging of other corpora.

10. Rules of POS tagging

POS tagging of a natural text corpus is not an easy task. It is full of quick-sands and loopholes. The thing that we need to keep in mind that POS tagging, either done manually or automatically, has to be carried out on a text corpus, which contains large number of words, which are actually occurring within sentences with specific syntactic functions and semantic values assigned to them. Therefore, it is not easy to identify the specific POS of words, until and unless actual syntactic roles of words are properly understood and defined. Moreover, there are several other linguistic and technical issues related to POS tagging, such as, *text sanitation, text normalization, tokenization, orthographic error correction, spelling error correction, real word-error correction, grammatical error removal, punctuation errors removal*, etc. (Dash 2009: 40-42).

All these works need to be carried out on a text corpus database before POS tagging takes place. In subsequent stages, this will remove many unwanted problems in *corpus processing, frequency calculation of words, type-token analysis of words, lemmatization, morphological processing, lexical sorting, machine translation, synset generation, WordNet design, dictionary compilation, and language teaching,* etc. Taking these factors into consideration we propose here a few Rules, which we should follow when we try to tag words in a text corpus (Dash 2011b)

Rule 1: Tagging should be done at sentence level

POS tagging should be carried out at the sentence level only – not on the words standing alone in a syntagmatic or paradigmatic distribution in different lexical databases. That means assignment of particular POS tags to words should be done after evaluating their linguistic roles in sentences where they actually occur. In other words, allocation of POS tags to words should be made in accordance with their lexico-grammatical roles within sentences. This is mandatory because a word is often found to be used in different part-of-speech from the part-of-speech assigned to it in a dictionary. For instance, the word সুন্দরী (sundarī) "beautiful$_{\text{[Fem.]}}$" may be identified as an adjective in a dictionary [< *sundar* "beautiful" + -ī$_{\text{[Fem-Suffix]}}$], but in actual text, it may be used as a noun, as the following examples show:

(11a) চিড়িয়াখানায় সুন্দরীকে দেখতে খুব ভিড় হয়েছে।
(ciṛiyākhānāy sundarīke dekhte khub bhiṛ hayeche)
"There is a huge crowd in the zoo to see Sundari".

(11b) স্কুলে সুন্দরী আমাদের সঙ্গে এক ক্লাশে পড়ত।
(skule sundarī āmāder saṅge ek klāśe paṛta)
"In school Sundari used to study in the same class with us".

These examples signify that intra-textual, contextual, and extra-textual information become indispensable for accurate POS tagging of the words.

Rule 2: Words should be normalized before POS tagging

All broken word strings should be joined together before the words are put to POS tagging. This is known as **normalization** (Sproat *et al.* 2001). In many situations it has been observed that a word is broken into two parts where the first part is a base form and the second part is a particle, case marker, or enclitic, etc. Although these have been used separately, these should to be attached to their respective base forms to produce the acceptable final forms, as the following examples show:

(12)	পেরে ছিল (pere chila)	>	পেরেছিল (perechila)	"had done",
	কথা গুলো (kathā gulo)	>	কথাগুলো (kathāgulo)	"the words",
	দিয়ে ছিলেন (diye chilen)	>	দিয়েছিলেন (diyechilen)	"had given",
	নরেন কে (naren ke)	>	নরেনকে (narenke)	"to Naren",
	মেয়ে দের (meye der)	>	মেয়েদের (meyeder)	"to girls", etc.

If such errors are not repaired, then all the broken parts of words will be tagged separately, which is a real distortion of the linguistic truth of the language as well as a fatal step towards decreasing the referential authenticity of the tagged text corpus.

Due to inconsistency in representation of words in Bengali, most of the compound words, adjectives, adverbs, multiword units, reduplicated forms, and idiomatic forms, etc. are written in the Bengali corpus in three different ways: with space, without space, and with a hyphen between the constituent forms, as shown below:

(13)	নৌকা বিহারে	(naukā bihāre)	"in boating",
	দেখে- শুনে	(dekhe-śune)	"seeing-hearing"
	কাল ক্রমে	(kāl krame)	"in course of time",
	কোন রকমে	(kona rakame)	"by any chance",
	এক জন	(ek jan)	"one person",
	চলন- বলনের	(calan-balaner)	"of moves and movements",
	ছেলে- মেয়েদের	(chele-meyeder)	"of boys and girls",
	কোনো- কোনো	(kono-kono)	"some",
	করে- করে	(kare-kare)	"having done",
	অন্দর মহলের	(andar mahaler)	"of inner house", etc.

Such words need to be normalized either as single word units or multiword units, as the case may be, based on their lexicosyntactic entities in the sentences before these are put to POS tagging.

Rule 3: Words should be tokenized before POS tagging

In the reverse manner, it is observed that sometimes some separate words are orthographically joined with their immediately preceding and succeeding words. In these cases, these words have to be separated from their neighbouring members before these are used for POS tagging. For instance, let us consider some examples presented below:

(14)	রামও সীতা (Rāmo Sītā) >	রাম ও সীতা (Rām o Sītā) "Ram and Sita",
	গেলেননা (gelennā) >	গেলেন না (gelen nā) "did not go",
	সেইইচ্ছা (seiicchā) >	সেই ইচ্ছা (sei icchā) "that will",
	সমগ্রজীবনকাল (samagrajībankāl) >	সমগ্র জীবনকাল (samagra jībankāl) "entire life".

This is generally called as **tokenization,** in a very loose manner (Huang *et al.* 2007). Moreover, white space needs to be provided consistently before and after each word (considered as individual lexical unit) which will ensure separate linguistic entity of a word.

Tokenization also involves word segmentation. Every word to be assigned with a particular POS is a single lexical unit and not a token, which internally contains more than one lexical item as the result of Sandhi or similar other euphonic combinations. For example, Bengali has a productive Sandhi process by which words of different part-of-speeches are concatenated into single word units, as the examples show:

(15) ভাল্লাগেনা (bhāllāgenā) [< ভাল (bhāla) + লাগে (lāge) + না (nā)] “is not liked”
যাচ্ছেতাই (ýācchetāi) [< যা (ýā) + ইচ্ছে (icche) + তাই (tāi)] “as one likes”, etc.

It is obvious that a token, which is internally made of words belonging to different part-of- speech, should not be assigned to either any one of the parts-of-speech. It is important to split such tokens into two or more tokens, as the case may be, before a POS tag is assigned to each one of them. The basic argument is that the POS is to be assigned to a simple grammatical entity and not to a complex one.

Rule 4: Exact POS tag should be assigned to words

A word should invariably be tagged at the POS level exactly in what part-of-speech it is used in a sentence. There should be no confusion or controversy in this regard. No attention should be given to lexicographic status or etymological antiquity of the words. For instance, in an English sentence like *“The plural form of ‘you’ is ‘you’”*, both *you* should be tagged as nouns (NN), and not as pronouns (PN), because *you* although is a pronoun, is actually used here as a noun. Similarly, in a Bengali sentence like সোনালী স্বপ্ন দেখতে ভালোবাসে (sonālī swapna dekhte bhālobāse) “Sonali loves to dream” the word সোনালী (sonālī) should be tagged as a noun (NN), and not as a adjective (JJ), even though the word সোনালী (sonālī) “golden” is an adjective in standard Bengali dictionary; and morphologically it can be derived as an adjective due to the presence of the adjectival suffix - লী (-lī) which is tagged to the noun সোনা (sonā) “gold”.

Rule 5: Context should carry utmost importance in POS tagging

It is not at all advisable to POS tag words solely based on POS categories as proposed in the dictionaries and grammars, as it may lead to problems in identification of actual POS roles of the words in a text. Therefore, tagging of words should be entirely context-based and this will instruct and guide a POS annotator about how words are to be tagged in specific contexts taking into consideration lexical, semantic, and syntactic functions of words. Although any general document on POS tagging, such as grammars and morphology of a language, can provide some basic ideas, it is almost certain that several context-specific issues will arise that will eventually lead for modification of existing POS tagsets and/or POS tagging guidelines.

Rule 6: Existing POS categories should be used on a text

It is advisable that POS tagset for a language should be designed in accordance with the existing and accepted set of parts-of-speech proposed in grammars and other reference guides, which has been understood for generations by the language users. Additional POS category can be assigned only when it is found that accepted POS tagset is not adequate enough to address new functions of words noted in texts. Also, it needs to be justified why a new POS tag has to be introduced and how does it supersede the existing POS tagsets of the language[9].

Rule 7: Support system should identify MWUs used in a text

There should be a support system for identification of multiword expressions (Sag *et al.* 2001) in the corpus. The examples of multiword units in Bengali include the followings:

(16) (a) **Compound words**, e.g., বেদনা প্রসূত (bedanā prasūta) "generated through pain", জীবন কল্প (jīban kalpa) "like life", ভ্রমর কৃষ্ণ (bhramar kṛṣṇa) "black as bumble bee", ভাব গম্ভীর (bhāb gambhīr) "serene with dignity", রৌদ্র দগ্ধ (raudra dagdha) "burnt with sun rays", সরকার নিযুক্ত (sarkār niẏukta) "appointed by government", etc.;

(b) **Idiomatic expressions,** e.g., চোখের মণি (cokher maṇi) "apple of one's eye", আষাঢ়ে গল্প (āṣāṛhe galpa) "cock and bull story", দেওয়াল লিখন (deoyāl likhan) "writing on the wall", উভয় সঙ্কট (ubhay saṅkaṭ) "horns of a dilemma", etc.);

(c) **Complex verb forms,** e.g., উঠে পড়া (uṭhe paṛā) "rise", শুয়ে পড়া (śuye paṛā) "lie", চলে যাওয়া (cale ẏāoyā) "leave", ফেলে আসা (phele āsā) "leaving", দেখে নেওয়া (dekhe neoyā) "seeing", গিলে ফেলা (gile phelā) "swallow", etc.; and

(d) **Proverbial expressions,** e.g., কাটা ঘায়ে নুনের ছিয়ে দেওয়া (kāṭā ghāye nuner chiṭe deoyā) "to add insult to injury", বিড়ালের গলায় ঘণ্টা বাঁধা (biṛāler galāy ghaṇṭā bā̃dhā) "to bell a cat", তেলা মাথায়া তেল দেওয়া (telā māthāy tel deoyā) "to carry coal to New Castle", etc..

The support system may be initiated before POS tagging starts or after it is complete. Since POS is a lexical level annotation process, any unit that involves more than one lexical item should be captured with the text database. These MWUs should be tagged as **chunks** and treated with utmost importance because there is valuable lexicosemantic information involved in these lexical items, which asks for separate investigation vis-à-vis treatment for future works of linguistics and language technology.

Rule 8: Multi-tagging approach should be strictly avoided

Multi-tagging approach should be strictly avoided. Although it may appear that a particular word can have more than one POS tag, one should invariably assign that particular tag, which the word under investigation exerts in particular context of its occurrence. For instance, if a word like ভাল (bhāla) 'good' occurs as a noun in a sentence, one must tag it as a noun (N), and not as an adjective (JJ) just because it is identified so in the dictionary. Similarly, it should not carry double tags (e.g., ভাল bhāla\NN\+\JJ\) just because it is used in both parts-of-speech in language and recorded such in the dictionary.

Rule 9: Morphological Processing must be separated from POS Tagging

Morphological processing and POS tagging of words are two different processes and therefore should be treated separately[10]. For instance, the conjugated Bengali verb form বলেইছিলাম (baleichilām) "I had indeed said" should be morphologically analysed in this manner:

(17)	বল (bal)	[FV-Root]
	- ◌ে (-e)	[Aspect]
	- ◌ি (-i)	[Particle_Emphatic]
	- ছ (-ch)	[Auxiliary]
	- ◌িল (-ila)	[Tense_Past]
	- ◌াম (-ām)	[Person_First + Number_Sing/Pl.]

We can retrieve from here all kinds of morphological information of the word to identify its form, class, function, and meaning. In fact, the information extracted from morphological analysis may be used in POS tagging of word as well as in lexical form generation, machine learning, information extraction, and parsing. But it should never be mixed up with the task of POS tagging.

Rule 10: Hyphenated words needs special attention

In case of hyphenated words, we have to be seriously careful in POS tagging as there are several morpho-semantic issues involved in them. We have adopted two different approaches to deal with such words. In case of those words where a formative element (e.g., inflection, particle, case marker, etc.) is separated from the word with a hyphen, it is better to tag the entire hyphenated word as a single lexical unit, since these formative elements are actually the part of the base form, as the examples given below:

(18) হো- য়াট (ho-yāṭ) “what”, মা- ই (mā-i) “mother herself”, কালিদাস- এর (kālidās-er) “of Kalidas”, স্টেটসম্যান- এ (sṭeṭsmyān-e) “in Statesman”, সোমবার- এ (sombār-e) “on Monday”, পদ- এর (pad-er) “of lexeme”, দেশ- এর (deś-er) “of Desh”, মা- র (mā-r) “of mother”, চা- টা (cā-ṭā) “the tea”, পা- টি (pā-ṭi) “the leg”, etc.

On the other hand, in case of those word forms, where hyphen is used between two potentially individual lexical items, which are capable of independent use, it is sensible to tag the words as well as the hyphen as separate entities, because here hyphen is just a functional connector between the words. For instance, consider the following types:

(19) (a) ভূ- প্রকৃতি (bhū-prakṛti) “geo-nature”, কু- স্বভাব (ku-svabhāb) “bad habit”, ছু- মন্তর (chu-mantar) “by a single breath”, etc.

(b) পিক- আপ (pik-āp) “pick up”, বাই- পাস (bāi-pās) “by pass”, মেক- আপ (mek-āp) “make up”, ফলো- অন (phalo-an) “follow on”, etc.

(c) উ- কার (u-kār) “u-allograph”, এ- কার (e-kār) “e-allograph”, ও- কার (o-kār) “o-allograph”, etc.

(d) চোর- ডাকাত (cor-ḍākāt) “thief and robber”, রোগা- মোটা (rogā-moṭā) “thin and thick”, মন- গড়া (man-gaṛā) “fancy-made”, সেই- দিন (se-din) “that day”, দু- বেলা (du-belā) “two times”, শেলী- কীটস (śelī-kīṭs) “Shelley and Keats”, টাকা- পয়সা (ṭākā-paysā) “penny and pie”, কৃষি- মন্ত্রী (kṛṣi-mantrī) “agriculture minister”, স্কুল- মাষ্টার (skul-māsṭār) “school teacher”, বার্লিন- অলিম্পিক (bārlin-alimpik) “Berlin Olympic”, উত্তর- পশ্চিম (uttar-paścim) “north-west”, etc.

(e) দক্ষিণের- দোলা- লাগা- পাখি- জাগা- বসন্ত- প্রভাতে (dakṣiner-dolā-lāgā-pākhi-jāgā-basanta -prabhāte) “in spring morning shaken by swinging breeze of south and awaken by bird’s call”, হাজার- হাত- কালী (hājār-hāt-kālī) “Goddess Kali with thousand cut-off hands”, কথায়- কথায়- রাগ- করা- মেজাজ (kathāy-kathāy-rāg-karā-mejāj) “to-be-angry-with-every-word-temper”, etc.

In such cases, hyphen itself needs to be tagged separately with a separate tag meant for punctuation.

Rule 11: Ambiguity should be dissolved at the time of POS tagging

In a natural language corpus there are several words, which are ambiguous in sense denotation when used in text. For instance, in Bengali ভাবে (bhābe) can be used as a finite verb, a noun, or a particle in a sentence. Similarly, the word করে (kare) can be used as a finite verb, a non-finite verb, a noun, or as an indeclinable; যে (ẏe) can be used as a relative pronoun, as a demonstrative, as a particle or as a conjunction; না (nā) can be used as a negative particle, an interjection, a conjunction, or an emphatic particle; and ছাড়া (chāṛā) can be used as a noun, as an adjective, as a postposition, as a particle and as a verb. Therefore, POS tagging rules should explicitly spell out the tagging conventions to be adopted for these ambiguous words.

Rule 12: Manual verification and validation is mandatory for POS tagged text

POS tagging should be done together by at least three experts who are well versed in morphology, grammar, morphosyntactic rules, semantics, and syntax of the language. This will provide the tagged corpus the much needed authenticity derived from 2:1 ratio of tag assignment for accuracy. Surely, for a large number of words, all three experts will agree with specific POS for words. Confusions will arise for the function words, such as, *demonstratives, adjectives, postpositions, pronouns, adverbs, particles, non-finite verbs, conjuncts, indeclinable, etc.* where experts will invariably differ in their opinions with regard to assigning specific POS tags. The rule of the thumb is: if two experts agree with the same POS for a disputed word, the game is over.

After a text is tagged at POS level, various other works of text processing, such as, *lexical sorting, frequency calculation of words in different part-of-speech, concordance, lemmatization, Local Word Grouping, Key-Word-In-Context,* etc. may be carried out on the POS tagged text corpus to retrieve information of various types and nature to be used in the works of both linguistics and language technology. Since these processes are more complex in computation and need special investigation, hence are not discussed here.

11. Conclusion: Value of a POS Tagged Corpus

The importance of a POS tagged corpus in simply enormous in language description, natural language processing, and language technology. The relevance of a POS tagged corpus within a wider canvas of computational linguistics, cognitive linguistics, and applied linguistics may be visualized in the following architecture (Fig. 6), which clearly shows how a POS tagged text becomes the primary source of data and information for linguistic works of various kinds.

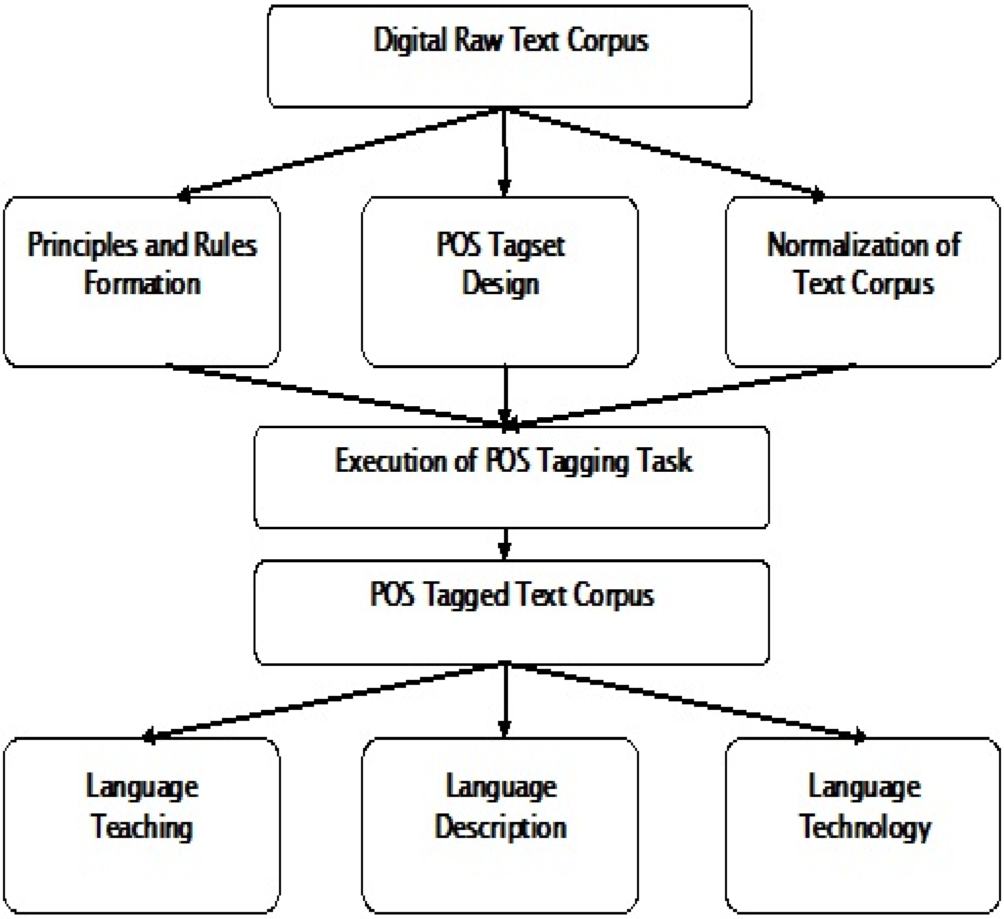

Fig. 6: Architecture of POS Tagged Corpus Generation

It is the first step for most of the complex Natural Language Processing applications like developing systems for grammar checkers, named entity recognition and extraction, word sense disambiguation, sentence parsing, text understanding, query addressing, information retrieval, machine translation, E-learning, and E-governance, etc. Besides major works of NLP, a tagged text corpus is also useful for machine learning, extraction of linguistic properties and grammatical elements, language modelling, and other works. In the area of mainstream linguistics and applied linguistics a POS tagged corpus may be useful for the works of frequency calculation of words, type-token analysis of words, lemmatization, lexical sorting, basic vocabulary list compilation, dictionary compilation, and language teaching, etc.

Although one can visualize many more application facilities of a POS tagged corpus in a natural language like Bengali, till date not much effort is initiated to develop this highly useful linguistic resource. However, in recent past an effort is initiated by the *Department of Information Technology, Govt. of India* (under the project entitled *Indian Languages Corpora*

Initiative-2009-2011) to develop parallel translation corpora in all major Indian languages in digital form; and a major component of this project is to develop tagged corpora for in the languages included in the project.

Whatever is done for Bengali or other Indian languages corpora, the rate of accuracy is far behind if compared with POS tagged corpora of English texts. For instance, in the one million word English text database of the *American National Corpus* (available at the Penn Treebank) the rate of accuracy is 97 to 98%, whereas for the in 10000 to 100000 words corpora of the Indian languages the rate of accuracy is 85 to 90% (Dandapat 2009). This clearly indicates that we sincerely need to take serious initiatives in this direction to develop POS tagged corpora for Bengali as well as for other Indian languages with two important goals: (a) we need to design maximally accurate tagset to increase the rate of accuracy of POS tagged data, and (b) we should develop POS tagged corpora in a large scale covering all text types for future linguistic works.

Notes and Comments

[1] Till date, there are many text corpora in English, German, Dutch, French, Italian, Japanese, Chinese, etc., which are tagged either manually or automatically using a set of grammatical rules. At present, these tagged corpora are used as valuable resources for designing tools for linguistic research as well as for devising techniques and systems for language technology (Atwell *at al.* 2000). For instance, *Lancaster-Oslo-Bergen* (LOB) *Corpus* is tagged by using *Constituent Likelihood Automatic Word-tagging System* (CLAWS) which is combined with post-editing by developers. It is now available both in its raw and annotated version for all kinds of linguistic works (Garside 1987).

[2] Although, this probability matrix gives us certain amount of assurance about the POS identity of a word by way of analysing the POS identify of its immediately preceding and succeeding words, it is not a full-proof strategy, since in a language like Bengali, a noun can easily be preceded by another noun or an adjective can easily be succeeded by two or more adjectives before a noun occurs. That means a probability matrix needs to be constantly updated and revised with new data collected from the corpus if we really want to understand the nature of word order exhibited in natural Bengali sentences.

[3] Viterbi Algorithm was conceived by Andrew Viterbi in 1967 as a decoding algorithm for convolutional codes over noisy digital communication links. The algorithm has found universal application in decoding the convolutional codes used in both CDMA and GSM digital cellular, dial-up modems, satellite, deep-space communications, and wireless LANs. It is now also commonly used in speech recognition, keyword spotting, computational linguistics, and bioinformatics. For example, in speech-to-text (speech recognition), the acoustic signal is treated as the observed sequence of events, and a string of text is considered to be the 'hidden cause' of the acoustic signal. The Viterbi algorithm finds the most likely string of text given the acoustic signal.

[4] Indian Language Part-of-Speech Tag set: Bengali [Linguistic Data Consortium (LDC) catalog number LDC2010T16 and ISBN 1-58563-561-8] is a corpus developed by *Microsoft Research Labs*, India to support the task of part-of-speech tagging and other data-driven linguistic research on Indian Languages in general. It is created as a part of the *Indian Language Part-of-Speech Tagset* (IL-POST) project, a collaborative effort

among linguists and computer scientists from *Microsoft Research Labs India; AU-KBC, Anna University, Chennai; Delhi University; IIT Bombay, Jawaharlal Nehru University, Delhi;* and *Tamil University, Tamil Nadu*. The goal of the IL-POST project is to provide a common tagset framework for Indian languages that offers flexibility, cross-linguistic compatibility, and reusability across those languages. It supports a three-level hierarchy of categories, types, and attributes. The corpus mainly consists, therefore, of two different levels of information for each lexical token: (a) lexical category and types, and (b) set of morphological attributes and their associated values in the context. The Bengali corpus contains 7168 sentences (102933 words) of manually annotated text from modern standard Bengali sources including blogs, *Wikipedia, MultiKulti* and a portion of the *EMILLE/CIIL* corpus. The annotated data is structured into two folders, Bangla-1 (3684 sentences, 51091 words) and Bangla-2 (3484 sentences, 51842 words), which represent the two stages in which the data was annotated. All annotated data is provided in both XML and text files. Each data file contains between 3,000-5,000 words. The XML file contains metadata about the material, such as language, encoding, and data size. The annotation guidelines for Bengali included in this release contain a detailed description of the annotation methodology. Additional information, updates, bug fixes may be available in the LDC catalog entry for this corpus at LDC2010T16. © 2010 *Microsoft Research Labs India,* Pvt. Ltd., © 2010 Trustees of the *University of Pennsylvania, USA.*

[5] In reality however, it is not conducive to the basic criteria assigned for text annotation. The amount or degree of accuracy in annotation has been a long-distant dream even for the advanced languages, because each new text is potential to throw up many unprecedented decision-making challenges to annotators. Therefore, the basic task of an annotator is to meticulously maintain consistency in the practice of annotation even though his or her decisions may show some degrees of arbitrariness with the tagset defined beforehand for his or her use.

[6] The definition of a multiword expression refers to the “idiosyncratic interpretations that cross word boundaries (or spaces)” (Sag *et al.* 2001). In other words, any expression that is composed of two or more words (lexemes) and is not predictable by way of the individual words that are used to construct it, can be called a multi word expression. Some examples of it include **compound words**, such as, বেদনা প্রসূত (bedanā prasūta) ‘generated through pain’, জীবন কল্প (jīban kalpa) ‘like life’, ভ্রমর কৃষ্ণ (bhramar kṛṣṇa) ‘black as bumble bee’, ভাব গম্ভীর (bhāb gambhīr) ‘serene with dignity’, রৌদ্র দগ্ধ (raudra dagdha) ‘burnt with sun rays’, সরকার নিযুক্ত (sarkār niẏukta) ‘appointed by government’, etc.; **idiomatic expressions,** such as, চোখের মণি (cokher maṇi) ‘apple of one’s eye’, আষাঢ়ে গল্প (āṣāṛhe galpa) ‘cock and bull story’, দেওয়াল লিখন (deoyāl likhan) ‘writing on the wall’, উভয় সঙ্কট (ubhay saṅkaṭ) ‘horns of a dilemma’, etc.; **complex verb forms**, such as, উঠে পড়া (uṭhe paṛā) ‘rise’, শুয়ে পড়া (śuye paṛā) ‘lie’, চলে যাওয়া (cale ẏāoyā) ‘going’, ফেলে আসা (phele āsā) ‘leaving’, দেখে নেওয়া (dekhe neoyā) ‘seeing’, গিলে ফেলা (gile phelā) ‘swallowing’, etc.; and **proverbial expressions,** such as, কাটা ঘায়ে নুনের ছিটে দেওয়া (kāṭā ghāye nuner chiṭe deoyā) ‘to add insult to injury’, বিড়ালের গলায় ঘণ্টা বাঁধা (biṛāler galāy ghanṭā b⍰dhā) ‘to bell a cat’, তেলা মাথায় তেল দেওয়া (telā māthāy tel deoyā) ‘to carry coal to New Castle’, etc.

[7] For example, Hindi language has several such forms like আগে (āge) ‘front’, পিছে (piche) ‘behind’, উপর (upar) ‘above/upstairs’, নিচে (nice) ‘below/down’, বাদ (bād) ‘after’, পহলে

(pahle) 'before', অন্দর (andar) 'inside', বাহার (bāhār) 'outside', etc. It is important to note that not every noun denoting time or space has to be marked as NST. It is only this sub-class of locational nouns, which also act as postpositions, have to be marked as NST.

[8] The term 'deictic' is used here in generic sense. It can be further discussed and, if required, one can replace this term with some other terms such as 'default'.

[9] A despicable tendency is observed among some of the POS taggers who either try to introduce and accept English tagset to define part-of-speech of Bengali words. The extreme example is the use of grammatical category preposition (PP) for Bengali, because it has been there in English, although it is universally known that Bengali does not have any preposition; it has only postpositions and case markers to denote almost the same functions carried out by English prepositions.

[10] This is however, not the final verdict that has to be obeyed by one and all engaged in the task of POS tagging, since it is inevitable that disputes are bound to arise in POS tagging of a natural language text. However, it is always better to come to a common agreement, for the time being, for achieving common goal in vision, rather than drawing daggers and fighting revengefully against each other with a mission for parading one's morphological knowledge in the pretext of hiding one's grammatical ignorance.

Acknowledgement

I humbly appreciate the critical constructive comments of the anonymous reviewers and inform that all their valuable comments and suggestions are meticulously implemented in the revised version of the paper.

References

Abney, Steven (1997) "Part-of-speech tagging and partial parsing". In, Susan Schreibman, Raymond George Siemens, and John M. Unsworth (Eds) *Corpus-Based Methods in Language and Speech: A Companion to Digital Humanities.* London: Blackwell. Pp. 118-136.

Atwell, Eric, G. Demetriou, J. Hughes, A. Schiffrin, C. Souter and S. Wilcock (2000) "A comparative evaluation of modern English corpus grammatical annotation schemes." *International Computer Archive of Modern English Journal.* 24(1): 7-23.

Biber, Douglas, Susan Conrad and Randy Reppen (1998) *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Brants, Thorsten (2000) "TnT- A Statistical Part-of-Speech Tagger". In the *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP-2000)*, Seattle, WA, USA, Pp. 37-42.

Brill, Eric (1992) "A simple rule-based part of speech tagger". In the *Proceedings of the Workshop on Speech and Natural Language (HLT-91)*, Morristown, NJ, USA: Association for Computational Linguistics. Pp. 112-116.

Brill, Eric (1995) "Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging". *Computational Linguistics.* 21(4): 543-565.

Charniak, Eugene (1997) "Statistical Techniques for Natural Language Parsing". *Artificial Intelligence Magazine*.18(4): 33-44.

Chattopadhyay, Suniti Kumar (1995) *Bhasa-prakash Bangala Byakaran (Bengali Language Grammar)*. Kolkata: Rupa Publications.

Church, Ken (1983) "A Finite-State parser for use in speech recognition". *Journal of the Acoustical Society of America*, Supplement 1, Vol. 74. P. S16.

Dandapat, Sandipan (2009) *Part-of-Speech tagging for Bengali*. MS Thesis, Dept. of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, India (MS).

Dash, Niladri Sekhar (2004) "Text annotation: a prologue to corpus processing". *Indian Journal of Linguistics.* 23(1): 71-82.

Dash, Niladri Sekhar (2005a) "Introduction to Corpus Linguistics and Language Technology". Delivered as ten-hours' lecture at the *Centre for Computational Linguistics, Linguistics Studies Unit, University of Madras*, Chennai. 17th – 22nd April 2005, Pp. 1-90.

Dash, Niladri Sekhar (2005b) *Corpus Linguistics and Language Technology: With Reference to Indian Languages*. New Delhi: Mittal Publications.

Dash, Niladri Sekhar (2009) *Corpus-based Analysis of the Bengali Language*. Saarbrucken, Germany: VDM Publications.

Dash, Niladri Sekhar (2010) "Polysemy and homonymy: a conceptual labyrinth". Presented in the *IndoWorrdnet-2010, Dept. of Computer Science and Engineering, IIT, Kharagpur*, 8th December 2010.

Dash, Niladri Sekhar (2011a) "Principles of Part-Of-Speech (POS) Tagging in Indian Language Corpora". In Vetulani, Zygmunt (Ed.) *Proceedings of 5th Language Technology Conference (LTC-2011): Human Language Technologies as a challenge for computer science and linguistics.* Poznan, Poland, 25th - 27th November 2011, Pp. 101-105. [Publisher: Fundacja Uniwersytetu im. A. Mickiewicza, ul. Rubiez 46 61-612 Pozana, Poland].

Dash, Niladri Sekhar (2011b) "Principles and Rules for POS Tagging of the Bengali Text Corpus". Presented in the *National Seminar On POS Annotation for Indian Languages: Issues & Perspectives (POSANIl-2011),* LDC-IL, CIIL, Mysore, 12-13th December 2011. (The paper is available online at http://www.ldcil.org/AnnPOSANILPresentations.aspx)

Dash, Niladri Sekhar (2011c) "Extratextual (Documentative) Annotation in Written Text Corpora". In, Sharma, Dipti Misra, Rajeev Sangal and Sobha L. (Eds.) *Proceedings of the 9th International Conference on Natural Language Processing (ICON-2011)* Pp. 168-176, Anna University, Chennai, India, 16th – 19th December 2011.

Dash, Niladri Sekhar (2011d) *A Descriptive Study of the Modern Bengali Script*. Saarland, Germany: Lambert Academic Publishing.

DeRose, Steven J. (1988) "Grammatical category disambiguation by statistical optimization". *Computational Linguistics.* 14(1): 31–39.

DeRose, Steven J. (1990) *Stochastic Methods for Resolution of Grammatical Category Ambiguity in Inflected and Uninflected Languages*. Doctoral Dissertation, Department of Cognitive and Linguistic Sciences, Providence, RI: Brown University, USA.

Fligelstone, Steve, Mike Pacey, and Paul Rayson (1997) "How to generalise the task of annotation". In, Roger Garside, Geoffrey Leech and Anthony McEnery (Eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman. Pp 122-136.

Fligelstone, Steve, Paul Rayson, and Nicholas Smith (1996) “Template analysis: bridging the gap between grammar and the lexicon”. In, John Thomas and Michael Short (Eds.) *Using Corpora for Language Research.* Harlow: Longman. Pp 181-207.

Francis, W. Nelson and Henry Kuchera (1964) “Manual of information to accompany ‘A standard sample of present-day edited American English, for use with digital computers’”. Providence, RI: Department of Linguistics, Brown University.

Garside, Roger (1987) “The CLAWS Word-tagging System”. In, Garside, Roger, Geoffrey Leech, and Geoffrey Sampson (Eds.) *Computational Analysis of English: A Corpus-based Approach*, London: Longman. Pp. 30-41.

Garside, Roger (1995) “Grammatical tagging of the spoken part of the British National Corpus: a progress report”. In, Leech, Geoffrey, George Myers, and John Thomas (Eds.) *Spoken English on Computer: Transcription, Mark-up and Application.* London: Longman. Pp. 161-167.

Garside, Roger (1996) “The robust tagging of unrestricted text: the BNC experience”. In, John Thomas and Michael Short (Eds.) *Using corpora for language research: Studies in the Honour of Geoffrey Leech*. London: Longman. Pp. 167-180.

Garside, Roger and Nicholas Smith (1997) “A hybrid grammatical tagger: CLAWS4”. In, Roger Garside, Geoffrey Leech, and Anthony McEnery (Eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora.* London: Longman. Pp. 102-121.

Garside, Roger, Geoffrey Leech and Anthony McEnery (Eds.) (1997) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.

Greene, Barabara B. and Gerald M. Rubin (1971) *Automatic Grammatical Tagging of English.* Technical Report. Providence RI: Department of Linguistics, Brown University.

Hans van Halteren, Jakub Zavrel, Walter Daelemans (2001) “Improving accuracy in NLP through combination of machine learning systems”. *Computational Linguistics.* 27(2): 199–229.

Kennedy, Graeme (1998) *An Introduction to Corpus Linguistics*. London, New York: Longman.

Kupiec, John (1992) “Robust part-of-speech tagging using a hidden Markov model”. *Computer Speech and Language*. 6(1): 3-15.

Leech, Geoffrey (1993) “Corpus annotation schemes”. *Literary and Linguistic Computing.* 8(4): 275-281.

Leech, Geoffrey (1997) “Introducing corpus annotation”. In, Garside, Roger, Geoffrey Leech and Anthony McEnery (Eds.) *Corpus annotation: Linguistic information from computer text corpora*. London: Longman. Pp. 1-18.

Leech, Geoffrey and Elizabeth Eyes (1993) “Syntactic annotation: linguistic aspects of grammatical tagging and skeleton parsing”. In, Black, Ezra,; Roger Garside, and Geoffrey Leech (Eds.) *Statistically-driven Computer Grammars of English: the IBM/Lancaster Approach*. Amsterdam: Rodopi. Pp. 36-61.

Leech, Geoffrey and Nicholas Smith (1999) “The use of tagging”. In, Halteren, Hans V. (Ed.) *Syntactic Word Class Tagging*. Dordrecht: Kluwer Academic Press. Pp. 23-36.

Leech, Geoffrey and Roger Garside (1982) “Grammatical tagging of the LOB Corpus: general survey”. In, Johansson, Stig and Knut Hofland (Eds.) *Computer Corpora in English Language Research*. Bergen: NAVF. Pp. 110-117.

Leech, Geoffrey, Roger Garside, and Eric Atwell (1983) “The automatic tagging of the LOB Corpus”. *International Computer Archive of Modern English News.* 7(1): 110-117.

Leech, Geoffrey, Roger Garside, and Mike Bryant (1994a) "CLAWS4: The tagging of the British National Corpus". In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)* Kyoto, Japan. Pp. 622-628.

Leech, Geoffrey, Roger Garside, and Mike Bryant (1994b) "The large-scale grammatical tagging of text: experience with the British National Corpus". In, Oostdijk, Nicholas and Peter deHaan (Eds.) *Corpus Based Research into Language*. Amsterdam: Rodopi. Pp. 47-63.

Majumdar, Paresh Chandra (1993) *Bangla Bhasa Parikrama (Survey of the Bengali Language)*. Vol-II. Kolkata: Dey's Publishing.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Aann Copestake and Dan Flickinger (2001) "Multiword Expressions: A Pain in the Neck for NLP". In, Gelbukh, Alexander (Ed.) *Proceedings of CICLING2002*. Verlag: Springer. Pp. 35-41.

Sarkar, Pabita and Ganesh Basu (1994) *Bhasa-jijnasa (Language Queries)*. Kolkata: Vidyasagar Pustak Mandir.

Shahidullaha, Muhammad (1967) *Bangala Byakaran (Bengali Grammar)*. 13th Edition. Dhaka: Maola Brothers.

Thompson, Hanne-Ruth (2010) *Bengali: A Comprehensive Grammar.* London and New York: Routledge (Taylor and Francis Group).

Toutanova, Kristina and Christopher D. Manning (2000) "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger". In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora* (EMNLP/VLC-2000). Pp. 63-70.

Toutanova, Kristina, Dan Klein, Christopher D. Manning, and Yoram Singer (2003) "Feature-rich part-of-speech tagging with a cyclic dependency network". In *Proceedings of HLT-NAACL 2003*. Pp. 252-259.