Marking Words with Part-of-Speech (POS) Tags within Text Boundary of a Corpus: the Problems, the Process and the Outcomes

Niladri Sekhar Dash

Abstract

A natural language text stored in a corpus database in electronic version can be tagged at the part-of-speech (POS) level manually or automatically. In both cases, it has to be done carefully starting with the lowest level of hierarchy of tagset meticulously devised for a language or a language group. Once the lower level tag is selected and assigned to words, the higher level tags will be automatically identified and assigned. Although tagging of words may be done with a focus on the part-of-speech of words used in a piece of text, the long term goals should also be envisaged for developing a generic scheme that may be useful for incorporating various kinds of linguistic information easily at the later stages of text annotation. This paper argues for taking a judicious decision for tagging words with different types of information within a text following the universally accepted principles, maxims and rules adopted for part-of-speech tagging. It describes the strategies, rules and methods adopted for manual tagging of a Bengali written text corpus at the part-of-speech level following the guidelines and methods proposed in the Bureau of Indian Standard (BIS) suitable for the language.

1. Introduction

Part-of-speech (POS) tagging is a process of *grammatical annotation* of words used in a piece of text in which one aims at assigning – automatically or manually – part-of-speech tags to each and every word used in the text after the word has passed through

the stages of morphological analysis and lexico-grammatical interpretations (Garside 1995). Usually, a set of specially designed codes carrying grammatical information are assigned to the words to indicate their POS with regard to their usage in a text (Leech and Garside 1982). In usual cases, a well-defined set of rules and strategies are used to identify and assign the POS tags to the words to determine and fix their lexico-semantic identities as well as their syntactic and grammatical functions within a given text. We can perhaps visualize the advantages of POS tagging at three levels of a word in the following ways:

(a) Lexical level: It allows analysing morphological structure of words represented in their surface forms.

(b) Orthographic level: It makes some distinctions in semantic roles of homographic words used in the same text or similar other texts.

(c) Syntactic level: It tries to identify the syntactico-grammatical functions of words to assign their appropriate POS entities

In general, POS tagging is treated as a common form of text annotation, which is invoked to start more comprehensive text annotation tasks where multiword expressions such as *compound words, reduplicated forms, lexical collocations, idiomatic expressions, fixed phrases, proverbial expressions, etc.* are assigned with chunking markers leading to eventual assignment of phrase markers to each of the sentences used in a text (Sag *et al.* 2001).

Although the application of POS tagset on a piece of text makes the text difficult to read and comprehend for human beings, it becomes highly suitable for linguistic information and data asked by a computer system for differentiating words used in different part-of-speech (Leech and Eyes 1993). From application point of view, POS tagging is a highly useful method, which increases specificity in the work of data retrieval from language corpora and provides essential grammatical information of the

words required in sense tagging, discourse tagging, rhetoric tagging, parsing, dictionary compilation, grammar development, language teaching, language cognition, etc.

This paper is the outcome of the attempt made for manually tagging a Bengali written text corpus. While engaged in the task, it identifies the stages of POS tagging (Section 2) following which it tries to explicate the process of marking metadata in a text (Section 3), process of marking paragraphs, segment and sentence boundaries (Section 4), process of marking words within text (Section 5), discusses the outcomes of a tagged corpus (Section 6), and reveals problems and ambiguities found in POS tagged text (Section 7). Finally, it identifies utilities of a POS tagged text in various works of applied linguistics, language technology, and descriptive linguistics (Section 8). The data and information presented here may be considered as an attempt for designing a well-formed strategy to be followed for developing POS tagged corpus for Bengali and for other Indian languages to be utilized in different domains of linguistics and language technology.

2. Stages of POS Tagging

From hand-on experiences gathered in manual tagging, it has been understood that the process POS tagging on a piece of text, in a systematic manner, should be carried out through the following eight steps:

1. Identification of a word within a piece of text.
2. Identification of its orthographic appearance and form.
3. Analysis of its morphological structure and formation.
4. Identification of its syntactic (i.e., grammatical) function in a sentence.
5. Determination of its grammatical role as well as part-of-speech.

6. Identification of its semantic role within the sentence of its occurrence.

7. Assignment of POS tags – either manually or automatically.

8. Verification and validation by experts.

Following the steps stated above the process of POS tagging on the Bengali text corpus was carried out at the following three stages:

(a) Stage 1: Sanitation and pre-editing of the text.

(b) Stage 2: Tag assignment to the words.

(c) Stage 3: Post-editing of the tagged text.

At the pre-editing stage, the entire Bengali text database was converted into a suitable format in digital form for carrying out the tagging tasks. At this stage, the whole text database was meticulously checked to verify if there was typographical and/or orthographical error of any kind within the text, and if there was any, it was manually corrected in accordance with the physical source text before the digital text was made ready for POS tagging (Dash 2004). Also, the selected texts were passed through the processes of **normalization** and **tokenization** to make the text maximally suitable for error-free POS tagging.

The tag assignment stage was initiated with the assignment of just one and only one POS tag to each word used in the sentence after proper consideration of its morphological, syntactic, and semantic roles in the sentence (Leech, Garside and Bryant 1994). For achieving greater accuracy at this stage, we had to consult, for reference purposes, a separate lexical database where the words were previously assigned with possible parts-of-speech. This lexical database was an open-ended resource in the sense that, time-to-time, it was up-dated with addition of new words obtained from various new sources and are assigned with possible POS tags. To deal with the new words, which were not

available in the previously made lexical database, we had to adopt some new methods such as including lists of common affixes and case markers with their possible part-of-speech identities of words for achieving greater accuracy in POS tagging (Biber, Conrad & Reppen 1998: 258-259).

At the stage of manual post-editing, the entire tagged text database was post-edited manually to examine if words were rightly tagged, and if there was any error made in POS tag assignment. In case of large corpus, where manual verification of the text database is highly time-consuming, tedious, and error-prone, it is better to adopt **probability matrix,** which may be devised from a text already tagged at POS level to deal with the problems of ambiguous tagging and dubious tag assignment (Leech, Garside and Atwell 1983). This strategy can help to specify **transition probabilities** that underlie between the adjacent tags. For instance, in Bengali, if a particular word is tagged as a noun (W[N]), the probability of its immediately preceding word to be an adjective (W[JJ]) is very high.

Usually a human annotator, who is engaged in assigning tag to words manually, can do the work quite successfully, if (s) he is well-acquainted with the grammar of a language. Also, a computer can do this work automatically, if it is properly trained with adequate amount of linguistic information, data, and rules for POS tag assignment. However, it needs to be trained properly beforehand to do the work with less percentage of errors. What it implies is that a system designer who is engaged in designing a computer system for automatic POS tagging should be well-equipped with adequate morphological, syntactic, and semantic knowledge of a language so that (s)he can develop a robust and accurate system to assign correct POS tags to the words, terms, and other lexical items used in a piece of text (Kupiec 1992).

However, before POS tagging was executed on the written Bengali text corpus database, there was an urgent need for a hierarchically well-defined and standard POS tagset, which would be used in a uniform manner by human annotators engaged or to

be engaged in POS tagging of words.

3. Marking Metadata in a Text

Since the Bengali written corpus contained texts of various types, it was important on the part of the annotators to maintain and preserve some meta-level information for each text document included in the corpus. Thus, various extratextual meta-information regarding *title, author, language, source, domain, text type, creator of text document*, etc. was marked on each text within a **Header File** as **metadata**. At the initial stage, this had been done manually in the following manner (Table 1).

<Header File>	Information
<Language >	Bengali
<Genre >	Written Text
<Category>	Aesthetics
<Subcategory>	Literature-Novel
<Text Type>	Imaginative
<Source Type>	Book
<Title>	ভুত আর ভুতো
<Volume>	Single
<Issue>	NA
<Edition>	First
<Headline>	ভুত আর ভুতো
<Author>	শুধাংশু পাত্র
<Publisher>	Dey's Publishing
<Pub. Place>	Kolkata, India
<Pub. Date>	1993

<Index No.>	B0035
<Creator Code>	61802
<Date of Creation>	12. 09. 2006
<Data Collector>	Anami Sarkar
<Proof Reader>	Aprakash Gupta
<Proofreading>	16. 08. 2007
<Total Words>	5017

Table-1: Header File with Metadata

The information stored in the Header File was actually related to various kinds of extralinguistic information that are considered necessary and useful for maintaining records of the text documents as well as for dissolving issues of copyright of the text materials used in generation of the corpus. One can also visualize the functional utilities of information stored in the Header File for carrying out innovative research works in *sociolinguistics, ethnolinguistics, ecolinguistics, geolinguistics, discourse, stylistics, language education,* and *language planning*, since all these sub-domains of linguistics require not only words and terms tagged at the POS level but also ask for appropriate linguistic data and information related to various socio-cultural issues and aspects for investigating the nature and patterns of language use controlled by various demographic factors and sociolinguistic variables.

4. Marking Paragraph, Sentence and Segment

After the completion of metadata preservation in the Header File, the next stage started with the act of marking paragraphs, sentences, and segments used in the text. Paragraphs were manually marked with </p>, both at their beginning and their end in the following manner to indicate their unique linguistic identities (Fig. 1).

	ভূতো - আমাদের ভূতো বাবু!	
	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 খুব ভালবাসে, গপ্পো শুনতে আরও ভালবাসে, ইস্কুলে রিণা মিনা নান্টু-মিন্টুদের সাথে ছবি ও ছড়ার বই পড়ে, সেইই ভূতো - আমাদের ভূতোবাবু।	

Fig.-1: Paragraph Boundaries marked in a Text

The second part of this stage was the marking of sentences and segments with some special boundary markers, which was carried out in the following manner (Fig. 2). While complete and fully formed sentences were marked with a tag <sentence>, both at their beginning and at their end, the incomplete sentences as well as isolated phrases were marked with a tag , both at their beginning and their end, in the following manner (Fig. 2).

আপনার দাঁতের যত্ন <sentence>তাজা শ্বাস আর ঝকঝকে দাঁত আপনার ব্যাক্তিত্বকে আকর্ষণীয় করে ।</sentence>
দক্ষিণ ভারত ভ্রমণ <sentence> দক্ষিণ রেলওয়ের চেন্নাই স্টেশন থেকে ধনুষ্কোট যাওয়ার পথে প্রধান লাইনে চেন্নাই থেকে ৩৫ মাইল দূরে চঙ্গলপেট স্টেশন পড়ে। </sentence>

Fig.-2: Marking of Sentences and Segments

Marking fully well-formed and grammatically accepted sentences with boundaries within a piece of text is highly important, as it helps in automatic identification of sentences as well as counting the number of sentences used in a piece of text. Also, it simplifies the process of generating parsed sentences and tree-banks of different grammar formalisms.

5. Marking Words in Texts

After marking all paragraphs, sentences, and segments

within the Bengali text corpus, effort was made to mark POS tags to words used in the text. Following the standards of the *Bureau of Indian Standard* (BIS), the written Bengali text corpus has been POS tagged in the following manner (Fig. 3).

<sentence>যদি\CC_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_NN\করে\V_VM_VNF\দিই\V_VM_VF\তবে\CC_CCS\
আমরা\PR_PRP\নিজের\PR_PRF\চোখে\N_NN\একটা\QT_QTC\জ্যান্ত\JJ\
মস্তিষ্ককে\N_NN\দেখতে\V_VM_VINF\পাবো\V_VM_VF\। \RD_PUNC\
</sentence>

Fig.-3: POS tagging of Words within a Text

At the time of manual POS tagging it had been observed that there could be the cases where a piece of text had included words from a language other than the matrix language. For example, a Bengali text composed in the Bengali script contained many English words which were actually written in the Roman script. From tagging and processing point of view, it was highly necessary to mark these words at the level of **vocabulary tagging** with information related to the respective languages.

All the above information of POS tagging can also be marked automatically to a certain level of accuracy in a text and without error of any kind if a computer system assigned with the task is put to rigorous training with a corpus tagged manually. In spite of this, there is surely to have some errors and ambiguities in POS tag assignment, which have to be checked and corrected manually (Dash 2005a: 124-129). How these problems may arise and how these have to be solved are discussed in some details in the following section (Section 6).

6. Outcomes of a Tagged Corpus

After initial assignment of possible POS tags to words the

Bengali tagged corpus was available for manual verification for POS tag validation as well as disambiguation (Leech, Garside and Atwell 1983). We had to depend on the probability matrix for this purpose as it was capable to specify the transition probabilities underlying between the adjacent tags. For example, when a given word (W1) was tagged as adjective (W[JJ]), its immediately succeeding word (W2) was mostly tagged as a noun (W[N]). This kind of probability measurement was an open matrix that could be updated with data collected from corpora of different text types (Biber, Conrad and Reppen 1998: 258-259). After completion of open matrix of probability measurement we could carry out post-editing manually to examine if all correct outputs were obtained from the tagged database.

At the time of manual verification of the tagged corpus database, we had found three broad types of words within the corpus:

(a) Rightly tagged words,

(b) Ambiguously tagged words, and

(c) Wrongly tagged words.

6.1 Rightly Tagged Words

Since most of the words used in the Bengali text corpus are inflected, most of the nouns, verbs (finite and non-finite), pronouns, adjectives, and adverbs, which are used in their inflected forms, are found rightly tagged. The basic reasons behind their right tag assignment are as follows.

a) Almost all the inflected nouns and pronouns are rightly tagged due to their inflectional elements. Various word-formative properties, such as, case markers, particles, and suffixes, etc. have worked here as distinctive marks for part-of-speech identification of the words. In fact, suffix elements of inflected words are vital clues for determining the grammatical roles of words in the sentence.

b) The majority of finite and non-finite verbs are also tagged rightly because of their inflectional elements. Based on the inflections we could easily identify if a word was used as a verb or not in the text, although in some cases it was difficult to determine if the word was used as a finite or non-finite verb, e.g., করে (kare), করতে (karte), করলে (karle), বলতে (balte), নিলে (nile), etc. Also, required information retrieved from the root and suffix lists used for this purpose helped to identify the right POS for words.

c) For adjectives and adverbs, the above process was followed with certain amount of accuracy, as most of the adjectives and adverbs are found to be used in the text in their inflected forms.

d) For indeclinables, there was high percentage of accuracy, since these forms were never found to be tagged with formative element, which is used with words of other parts-of-speech. Moreover, since these words are highly limited in number in Bengali, these are stored in a separate lexical database. At the time of POS tagging, once a perfect match was found in the lexical database, an indeclinable was identified and tagged. Thus, Bengali indeclinables like কিংবা (kiṃbā) 'or', এবং (ebaṃ) 'and', কিন্তু (kintu) 'but', বা (bā) 'or', তথাপি (tathāpi) 'yet', বরং (baraṃ) 'rather', আর (ār) 'and', যদি (ýadi) 'if', etc. are tagged rightly, because these words are usually fixed in their part-of-speech and they hardly use inflection or case markers.

6.2 Ambiguously Tagged Words

Ambiguity is bound to happen in POS tagging because ambiguity is a common feature in all natural languages and identification of actual POS of a word does not always depend of its form, but on its meaning and function it exerts in a piece of text. Moreover, contexts, discourse, intralinguistic and extralinguistic information that are embedded within a text also play crucial roles for making a word ambiguous. That means a single lexical item, based on the context of its use in a text, may convey more

than one meaning, event, or idea, vis-à-vis, part-of-speech (Dash 2005b). From experiences gathered in manual tagging of the Bengali text corpus, it has been understood that uncertainties in part-of-speech of words are quite frequent: not merely because of failures of human understanding, but because of the prototypical and/or fuzzy nature of most of the linguistic categories (Leech 1993: 280).

What is also understood from such hand-on experience is that efficiency and adequacy of a POS tagset comes from the way it succeeds in handling the feature of lexical ambiguity. In POS tagging, ambiguity arises at the lexical level, because most of the lexical items can allow more than one reading triggered from sense variation they generate. Thus, a word may be associated with a dozen different readings if all its idiomatic, figurative, proverbial, and contextual usages are taken into consideration.

At the time of initial POS tagging and manual verification, two types of ambiguity are found in the Bengali text corpus:

(a) Structural ambiguity, and

(b) Sequential ambiguity.

Structural ambiguity, which was noted at the lexical and sentence level, happens for inflected and non-inflected words where an inflected form or non-inflected root, stem or base, due to its homographic form appeared to belong to different parts-of speech. For instance, let us consider the underlined word of the example taken from the Bengali text corpus:

(1) আজ তোমাকে ছাড়া হবে না।

(āj tomāke chāṛā habe nā)

1st reading: "You will not be released today"

2nd reading: "Today it is not possible without you"

In the example given above we can have two different POS for the word ছাড়া (chāṛā). It can be a gerund if 1st reading is taken into consideration. On the other hand, it is a postposition if the 2nd reading is taken into consideration at the time of tagging of the word. Thus it became a problem for an annotator to decide in which part-of-speech this word should be tagged in the sentence.

Moreover, some inflected words, due to close structural similarities in their roots and suffixes, could also become ambiguous and these were very difficult to be sidelined to one or the other part-of-speech. For instance, the word করে (kare), at the time of POS tagging could be tagged as a non-finite verb, a finite verb, an indeclinable, or a noun. In the same manner, the word ছাড়া (chāṛā), based on the context of its use and its semantic function in the sentence, could be tagged as an adjective (chāṛā[JJ]) 'freed', a postposition (chāṛā[RB]) 'without', a noun (chāṛā[N]) 'a female calf', or as a gerund (chāṛā[V]) 'releasing'. What all these examples mean is that identification of actual part-of-speech of a word is not a trivial task; it seriously asks for information from various levels before it is fixed in its proper semantico-syntactic role and is tagged accordingly.

The indeclinables, due to their one-dimensional linguistic entities, are usually assigned with single POS tag, but postpositions and adjectives are highly ambiguous and are often prone to double POS tags. For instance, the word সুন্দর (sundar) can be tagged as adjective (sundar[JJ]) 'beautiful' as well as noun (sundar[N]) 'beauty' based on its use in sentence.

On the other hand, sequential ambiguity was usually caused due to the presence of immediately following word, which if tagged together with the word under investigation, would produce a part-of-speech, which differed from the individual parts-of-speech of words. For instance, when বিশেষ (biśeṣ) and ভাবে (bhābe) were POS tagged separately, বিশেষ (biśeṣ) was tagged as an adjective (biśeṣ[JJ]) while ভাবে (bhābe) was tagged as a finite verb (bhābe[V]) or a postposition (bhābe[PSP]).

But when they were treated as a single word unit বিশেষভাবে (biśeṣbhābe), they were combined together to be POS tagged as an adverb (biśeṣbhābe[RB]), which was different from respective independent parts-of-speech of the words.

Similar ambiguities arose for detached compound words, reduplicated forms, collocations, idioms, set phrases, and proverbs, such as, বেদনা প্রসূত (bedanā prasūta) 'generated through pain', জীবন কল্প (jīban kalpa) 'like life', ভ্রমর কৃষ্ণ (bhramar kṛṣṇa) 'black as a bumble bee', ভাব গম্ভীর (bhāb gambhīr) 'serene with dignity', রৌদ্র দগ্ধ (raudra dagdha) 'burnt with sun rays', সরকার নিযুক্ত (sarkār niẏukta) 'appointed by government', 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', উঠে পড়া (uṭhe paṛā) 'rise', শুয়ে পড়া (śuye paṛā) 'lie', চলে যাওয়া (cale ýāoyā) 'going', ফেলে আসা (phele āsā) 'leaving', দেখে নেওয়া (dekhe neoyā) 'seeing', গিলে ফেলা (gile phelā) 'swallow', etc. Since the BIS tagging scheme designed for Bengali and other Indian languages works for only single word unit using information of words at the lexical level, such ambiguities are bound to take place.

6.3. Wrongly Tagged Words

In the Bengali tagged text corpus, we have come across some words, which were assigned wrongly with inappropriate tags. The basic reasons behind this phenomenon are possibly the followings:

a) Erroneous identification of POS for a word. It often happens for those words, which belong to more than one part-of-speech. For instance, the Bengali word কি (ki) can be a pronoun or an emphatic particle. If the syntactic function of the word in a piece of text is miss-read, it can be tagged as a pronoun in place of an emphatic particle or vice versa.

b) In case of non-inflected verbs, wrong tagging has happened

due to non-availability of roots in lexical database or due to non-acquaintance with the forms by the annotators.

c) For some nouns, pronouns, and adjectives it is noted that wrong tagging is mostly caused due to non-acquaintance with the forms or due to wrong identification of POS of the words.

d) For some adverbs, wrong tagging is caused due to a different reason. In most cases, space given between the two formative parts of an adverb puts a barrier for its proper analysis and tagging (discussed in 6.2). The other reason may be the same problems faced for the words of other POS categories.

e) Some nouns, which are used as verbs within a piece of text, are not tagged rightly.

f) Proper nouns (e.g., person names, place names, item names, etc.), transliterated foreign words, dialectal vocabularies, etc., due to their unique lexical entities are usually undefined in their part-of-speech and are tagged wrongly.

These issues are, however, mostly related to linguistics and these may be resolved with proper training to the annotators. In case of a computer system for automatic tagging such problems may be dissolved by regular up-gradation of lexical database and by modification of the algorithms used for the purpose.

7. Dissolving Problems of POS Tagging

To resolve lexical ambiguities in POS tagging in Bengali corpus, we propose to adopt the approach of delayed tagging, which in principle, is based on information extracted from the immediate context of the local contextual environment of a word under consideration. For instance, consider the Bengali word ভাল (bhāla), which can be tagged as a noun or an adjective in a sentence. When we encounter the word in a particular position in a given sentence, we shall not try to tag the word until and unless we finish reading the entire sentence and take into account the

actual role of the word with regard to its semantico-syntactic function in the sentence. What we argue is that a full and complete reading of the sentence is indispensable as it will supply necessary information to understand in which part-of-speech the word is actually used in sentence. After knowing the role of a word in a sentence, we can tag the word accordingly – as noun or adjective. In our view, the implementation of this approach will invariably minimise the problems of wrong and ambiguous tagging. In case of automatic tagging, however, we have to think of a method, which can act in the same manner.

The problem of ambiguity in POS tagging is also related to some higher level ambiguities, such as, attachment ambiguity, assignment ambiguity, referential ambiguity, etc. In all these cases, sense disambiguation and POS tagging have to be done after understanding the nature of association of the lexical items, analysing internal structure of words, investigating contextual occurrence of words, and understanding intralinguistic and extralinguistic information embedded in a sentence or a piece of text. Only then we can think of adopting a principled way of disambiguation (Dagan and Itai 1994).

Since POS tagging relied heavily on various kinds of information to different extents, we had to put together different information to identify rightly the POS tag of a word (for the ambiguous lexical items) or to make the best guess based on the information available to us. In all these cases, we had to wait till information from the syntactic and semantic levels were acquired and combined with extralinguistic knowledge (Justeson and Kats 1995).

Moreover, we had to access a dictionary, which had listed up the part-of-speech to which a particular word could belong. This dictionary helped us to identify the POS of some of the fixed expressions, e.g., যাচ্ছেতাই (ýācchetāi) 'simply worthless', নাহলে (nāhale) 'if not', etc. In case of automatic tagging, one can use machine-readable dictionary that also lists up the words, which usually exercise certain grammatical constraints in conditional

statements, such as, যদি (ýadi) and তবে (tabe) 'if…then', etc.

Also, we had to use probabilistic information acquired from the previously tagged text. It had guided us about how likely it is that a given word can belong to one part-of-speech or the other. For instance, although the Bengali word কর (kar) is used as verb or noun, information from previously tagged corpora showed that it had much higher probability of occurring as a verb than as a noun.

Furthermore, we had used another innovative method developed with information of grammatical uniformity of words used in Bengali. For instance, it was noted that indeclinables were tagged easily and accurately as these were mostly non-inflected in form and less in number. Similarly, the conjugated finite and non-finite verbs were tagged with limited errors, because with a fixed number of suffixes we could easily identify these forms and tag these forms accordingly with least knowledge of grammatical agreement between roots and suffixes. On the other hand, in case of nouns, pronouns, adjectives, and adverbs, we had to toil very hard as most of these forms were used with or without inflection markers. Besides, these forms tended to change their lexical roles and identities within a text based on contexts of their occurrence. Therefore, their ambiguity had to be dissolved first with reference to their usages in the sentence before these were tagged to any part-of-speech.

Finally, it can be argued here that each word used in a piece of text can be sidelined to a fixed part-of-speech or lexical class if we can analyse its role and meaning in the context of its use. For example all proper, common, material, collective, abstract, human, and non-human nouns can be brought under the single head category: noun (N) while all relative, reciprocal, definite, indefinite, reflexive, emphatic, interrogative, others may be brought under the head category: pronoun (PR). Similarly, all adjectives (JJ), adverbs (RB), finite and non-fine verbs (V), postpositions (PSP), etc. can be grouped and put under single

head. It can reduce multiple parts-of-speech of words into one as well as can simplify the process of POS tagging.

Although there is every possibility for tagging additional information regarding the grammatical properties and semantic sub-classes of words to a POS tagset, it requires detailed investigation into the formal and functional aspects of the words. If it becomes possible, then a POS tagging scheme can be robust and useful for next levels of tagging and processing.

Even then the completely error-free tagged text was not possible to generate due to difference in opinions of the human annotators. However, the problem of lexical ambiguity and its solution in the domain of POS tagging is separate area of research, which is just hinted here.

8. Utilities of POS Tagged Corpus

A text corpus tagged at the POS level is a useful resource for research and development in language description, language processing, and language technology. It is the most common resource, which has established it functional relevance in chunking and sense tagging (Leech and Smith 1999). After the generation of an error-free POS tagged corpus, it can be used for chunking as well as for extracting suitable patterns, rules, and features to be used for various NLP activities.

In the area of natural language processing a POS tagged corpus may be used for developing systems for grammar checking, recognition of the named entities, text understanding, parsing, word sense disambiguation, query addressing, lexical mapping, machine translation, and machine learning. A POS tagged corpus is also useful for extracting linguistic items and terms, in information retrieval, language modelling, and other works.

In descriptive and applied linguistics, a POS tagged corpus is useful for frequency calculation of words, type-token analysis, lemmatization, lexical sorting, primary vocabulary compilation, dictionary compilation, language teaching, etc.

Although one can visualize many applications of a POS tagged corpus for the Indian languages, till date not much effort is initiated to develop this highly useful linguistic resource. So far whatever tagging is done for Indian languages corpora, the rate of accuracy is far behind than expected if compared with POS tagged corpora made in English (Dandapat 2009). It implies that we sincerely need to take serious initiatives in this direction to develop POS tagged corpora for Indian languages with two basic goals: design maximally accurate tagset to increase the rate of accuracy of POS tagged corpora, and develop POS tagged corpora in a large scale covering all text types. If accurately POS tagged corpora of different types of text are made available for the Indian languages, may unaccomplished goals of language processing and language technology can be accomplished within a short span of time.

REFERENCES

Biber, D., S. Conrad, and R. Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*, Cambridge: Cambridge University Press.

Dagan, I. and A. Itai. 1994. Word sense disambiguation using a second language mono-lingual corpus, *Computational Linguistics*, 20(4): 563-596.

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

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

Dash, N.S. 2005a. *Corpus Linguistics and Language Technology: With Reference to Indian Languages.* New Delhi: Mittal Publications.

Dash, N.S. 2005b. Role of context in word sense disambiguation, *Indian Linguistics*, 66(1-4): 159-176.

Garside, R. 1995. Grammatical tagging of the spoken part of the British National Corpus: a progress report, in, Leech, G., G. Myers,

and J. Thomas, eds., *Spoken English on Computer: Transcription, Mark-up & Application*, London: Longman, pp, 161-167.

Justeson, J.S. and S.M. Katz. 1995. Principled disambiguation: discriminating adjective senses with modified nouns, *Computational Linguistics*, 21(1): 01-27.

Kupiec, J. 1992. Robust part-of-speech tagging using Hidden Markov Model, *Computer Speech and Language*, 6(1): 3-15.

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

Leech, G. and E. Eyes. 1993. Syntactic annotation: linguistic aspects of grammatical tagging and skeleton parsing, in, E. Black, R. Garside and G. Leech, eds. *Statistically-driven Computer Grammars of English: the IBM/Lancaster Approach*, Amsterdam: Rodopi, pp, 36-61.

Leech, G. and N. Smith. 1999. The use of tagging, in, Halteren, V.H. ed. *Syntactic Word Class Tagging*, Dordrecht: Kluwer Academic Press, pp, 23-36.

Leech, G. and R. Garside. 1982. Grammatical tagging of the LOB Corpus: general survey, in, Johansson, S. and K. Hofland, eds., *Computer Corpora in English Language Research*, Bergen: NAVF, pp, 110-117.

Leech, G., R. Garside, and E. Atwell. 1983. The automatic tagging of the LOB corpus. *International Computer Archive of Modern English News*, 7(1): 110-117.

Leech, G., R. Garside, and M. Bryant. 1994. The large-scale grammatical tagging of text: experience with the British National *Corpus, in, Oostdijk, N. and P. deHaan eds. Corpus Based Research into Language*, Amsterdam: Rodopi, pp, 47-63.

Sag, I.A., T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. 2001. Multiword expressions: a pain in the neck for NLP, in, Gelbukh, A. ed. *Proceedings of the CICLING-2002*, Verlag: Springer, pp, 35-41.