Principles and Rules for POS Tagging of the Bengali Text Corpus

Niladri Sekhar Dash
Indian Statistical Institute, Kolkata, India
Email: ns_dash@yahoo.com

1. Introduction

Primary goal of part-of-speech (POS) tagging is to disambiguate words and assign them to particular parts-of-speech.

It involves the complex process of marking up words as corresponding to particular parts-of-speech, based on its form, function and contexts of usage (i.e., relationship of words with their adjacent and related words) within larger syntactic strings like phrases, sentences, paragraphs, and texts (Leech 1997).

Here I try to define **Principles** to be followed in designing POS tagset, and **Rules** to be used at actual POS tagging of words in corpus.

I take into consideration **intralinguistic** and **extralinguistic issues** related to POS tagset design and assignment of tags to words of a language or its sub-types.

Consider the examples of বদলে (badale):

(1) সময় কিন্তু এখন অনেক বদলে গেছে
(samay kintu ekhan anek **badale**$_{[NFV]}$ geche)
"Time has indeed changed a lot by now"

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

The word বদলে (badale) in 1st sentence (VNF) and in 2nd sentence is a PSP, because in two different sentences it performs two different semantico-grammatical roles.

Thus, taking into consideration syntactico-semantic roles of a word, we determine if a word is used as a non-finite verb or as a postposition in a particular sentential context (Leech 1997).

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

POS assigns part-of-speech tags to each word used in a text after morphological analysis and grammatical interpretations.

Advantages of POS tagging are realized at 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 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.

POS tagging is 1st stage for comprehensive process in which **multiword expressions (MWE)** are assigned with chunking markers leading to assignment of phrase markers to each sentences within a text.

Although assignment of POS tags to words makes a text difficult to read for humans, it is intelligible a computer to provide linguistic information needed by for differentiating between words used in different POS.

POS tagging is useful for increasing specificity in **data retrieval** from corpus and for providing basic grammatical information about words for semantic annotation, discourse annotation, parsing, dictionary making, language teaching, grammar development, and language planning.

POS tagging on a corpus is carried out following 10 basic steps:

(a) Normalization of a digital text corpus,

(b) Identification of words within a piece of text,
(c) Identification of orthographic form and appearance of words,
(d) Analysis of morphological structures & formation of words,
(e) Identification of syntactic (grammatical) functions of words in sentence,
(f) Determination of grammatical roles and parts-of-speech of words,
(g) Identification of semantic roles of words in sentences,
(h) Assignment of POS tags to words, and
(i) Verification of tagged texts with existing grammar of a language, and
(j) Final validation of tagged texts by experts.

3. Early POS Tagsets for Bengali

Arguably the 1st generic POS tagset for Bengali is made by Dash (2005a) to tag a text of nearly hundred thousand words of Bengali prose texts for academic and training purposes. [Table 1 and Fig. 1].

No.	POS Cats	Label	Example
1	Noun	[NN]	বালক (bālak) 'boy', শহর (śahar) 'city'
2	Pronoun	[PR]	আমি (āmi) 'I', তুমি (tumi) 'you'
3	Demonstrative	[DM]	যে (ýe) 'that', এই (ei) 'this',
4	Finite Verb	[FV]	করছি (karchi) 'I am doing',
5	Non-Finite Verb	[NF]	করলে (karle) 'doing',
6	Adjective	[AD]	ভাল (bhāla) 'good', মন্দ (manda) 'bad'
7	Adverb	[AV]	হঠাত্ (haṭhāt) 'by chance',
8	Postposition	[PP]	পরে (pare) 'after', কাছে (kāche) 'near',
9	Conjunction	[CN]	তবে (tabe) 'then', যদি (ýadi) 'if',
10	Indeclinable	[IN]	কিন্তু (kintu) 'but', অথবা (athabā) 'or'
11	Particle	[PT]	ই (i), ও (o), তো (to), না (nā), নে (ne),
12	Quantifier	[QT]	এক (ek) 'one', দুই (dui) 'two',
13	Reduplication	[RD]	বনে বনে (bane bane) 'in forest',
14	Punctuation	[PN]	., : ; - / …, !, ? ( ), [ ], {, etc.
15	Others	[OR]	Math signs, +, -, x, >, <, $, #, @, ^, etc.

Table 1: Generic POS for Bengali (Dash 2005a)

তাই tāi_[AV] মানুষ mānuṣ_[NN] তাহার tāhār_[PR] সংগৃহীত saṃgṛhīta_[AD] দ্রব্যের drabyer_[NN] মধ্যে madhye_[PP] যেইটুকু ýeiṭuku_[PR] প্রয়োজনের prayojaner_[NN] অতিরিক্ত atirikta_[AD] সেইটি seiṭi_[PR] অন্যের anyer_[PR] সংগৃহীত saṃgṛhīta_[AD] পৃথক pṛthak_[AD] ধরনের dharaner_[NN] দ্রব্যের drabyer_[NN] সহিত sahit_[PP] বিনিময় binimay_[NN] করিয়া kariyā_[NF] নিজের nijer_[PR] প্রয়োজনীয় prayojanīiya_[AD] দ্রব্য drabya_[NN] সংগ্রহ saṃgraha_[NN] করিত karita_[FV] ._[PN] মানুষ mānuṣ_[NN] যেই ýei_[DM] যুগে ýuge_[NN] গুহার guhār_[NN] অভ্যন্তরে abhyantare_[PP] বসবাস basabās_[NN] শুরু śuru_[NN] করিয়াছিল kariyāchila_[FV] ,_[PN] সেই sei_[DM] যুগ ýug_[NN] হইতে haitei_[PP] মানব mānab_[NN] সভ্যতার sabhyatār_[NN] ক্রমবর্ধমান kramabardhaman_[AD] অগ্রগতির agragatir_[NN] সূচনার sūcanar_[NN] সঙ্গে saṅge_[RD] সঙ্গে saṅge_[RD] হিসাব hisāb_[NN] শাস্ত্রের śāstrer_[NN] ভিত্তি bhitti_[NN] প্রস্তর prastar_[NN] স্থাপিত sthāpita_[AD] হইয়াছিল haiyāchila_[FV] ._[PN] আদিম ādim_[AD] যুগে ýuge_[NN] মানুষ mānuṣ_[NN] পশু paśu_[NN] শিকার śikār_[NN] ,_[PN] মতস্য matsya_[NN] শিকার śikār_[NN] ও o_[IN] বন্য banya_[AD] ফলমূল phalmūl_[NN] সংগ্রহ saṃgraha_[NN] করিয়া kariyā_[NF] জীবিকা jībikā_[NN] নির্বাহ nirbaha_[NN] করিত karita_[FV] ._[PN]

Fig. 1: A POS tagged Bengali text (Dash 2005b)

Recently enterprises: Microsoft Research Labs, India, LDC-IL, CIIL Mysore), and DIT, MICT, Govt. of India, (BIS Tagset).

4. Principles of POS Tagging

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

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

Principle 3: Layered Approach should be used for POS tagging

Principle 4: Hierarchical Formation of tagset is desired

Principle 5: Extensibility of Tagset is required for specific languages

Principle 6: Metadata Format should be used for tagset

Principle 7: Dispensability of tagset from a text is mandatory

Principle 8: Standardization of tagset is a pre-requisition

Principle 9: User Friendly tagset should be designed

Principle 10: Non-ambiguity is must for POS tagset

Example (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 dā̃tke\N_NN\ পরিষ্কার pariṣkār\JJ\ ও o\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 POS tagging comes to an end, tagged corpus becomes ready for verification and validation.

Eventual tagged corpus may be used for chunking as well as for extracting suitable patterns, rules, and features to be used for training a system for automatic tagging of other corpora of the language.

5. Rules of POS Tagging

It is not easy to identify 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.

Taking these factors into consideration I propose here a few Rules, which we should follow when we try to tag words in a text corpus.

Rule 1: Tagging should be done at sentence level

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

(4) 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".

Rule 2: Words should be normalized before POS tagging

(5)

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.

(6)

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.

Rule 3: Words should be tokenized before POS tagging

(7) 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) "whole life".

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

Rule 4: Exact POS tag should be assigned to words

Example: 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.

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

There should be a support system for identification of MWEs.

(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 ýāoyā) "leave", ফেলে আসা (phele āsā) "leaving", দেখে নেওয়া (dekhe neoyā) "seeing", গিলে ফেলা (gile phelā) "swallow", etc.;

(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 bãdhā) "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

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. For instance, the conjugated Bengali verb বলেইছিলাম (baleichilām) "I had indeed said"

(10) 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.

Information extracted from morphological analysis may be used in POS tagging of word and 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 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, as these formative elements are actually the part of the base form.

(11) 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 words, 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 hyphen as separate entities, because here hyphen is just a functional connector between the words.

(12) (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 (ýe) 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 3 experts well versed in morphology, grammar, morphosyntactic rules, semantics, and syntax of the language.

It 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.

6. Discussion on Outputs

Following Principles and Rules, POS tagging is carried out on a Bengali corpus database at 3 separate stages:

Stage 1: Pre-editing of texts stored in corpus,
Stage 2: Tag assignment to words in the text, and
Stage 3: Post-editing of the tagged text.

At pre-editing stage, text is converted into a suitable format for carrying out POS tagging task. Entire text is manually checked to find out if there is any error (typographical or orthographical) within the text, and if there is any, it is corrected in accordance with the physical source text before the text is put to POS tagging.

Moreover, when required, selected text passed through *normalization* and *tokenization* to make the text maximally suitable for POS tagging.

Tag assignment stage begins with assignment of just one and only one POS tag to each word used in sentences after proper consideration of its syntactico-grammatical function in the sentences.

For achieving greater accuracy, we also refer to a lexical database where words are previously assigned with possible POS for reference purposes. Such a lexical database is open-ended in the sense that it is updated time-to-time with addition of new words obtained from various sources of language use.

To deal with newly found words in corpus that are not available in previous lexical databases, we adopt various methods such as the lists of common affixes and case markers with possible parts-of-speech for achieving higher accuracy in POS tagging.

At post-editing stage, tagged corpus is post-edited to verify if words are rightly tagged and if any error is made in POS tag assignment.

In case of larger texts databases, where verification of text database is tedious, time-consuming, and error-prone, we use **probability matrix** to deal with problems of ambiguous tagging and dubious tag assignment.

This matrix helps to specify **transition probabilities** underlying between the adjacent tags. For example, if a given word is tagged as noun (W_{N}), then the probability of its immediately preceding word to be an adjective (W_{JJ}) is quite high in a language like Bengali.

The entire method of POS tagging with assistance of Principles and Rules is explicated in the following diagram (Fig. 3).

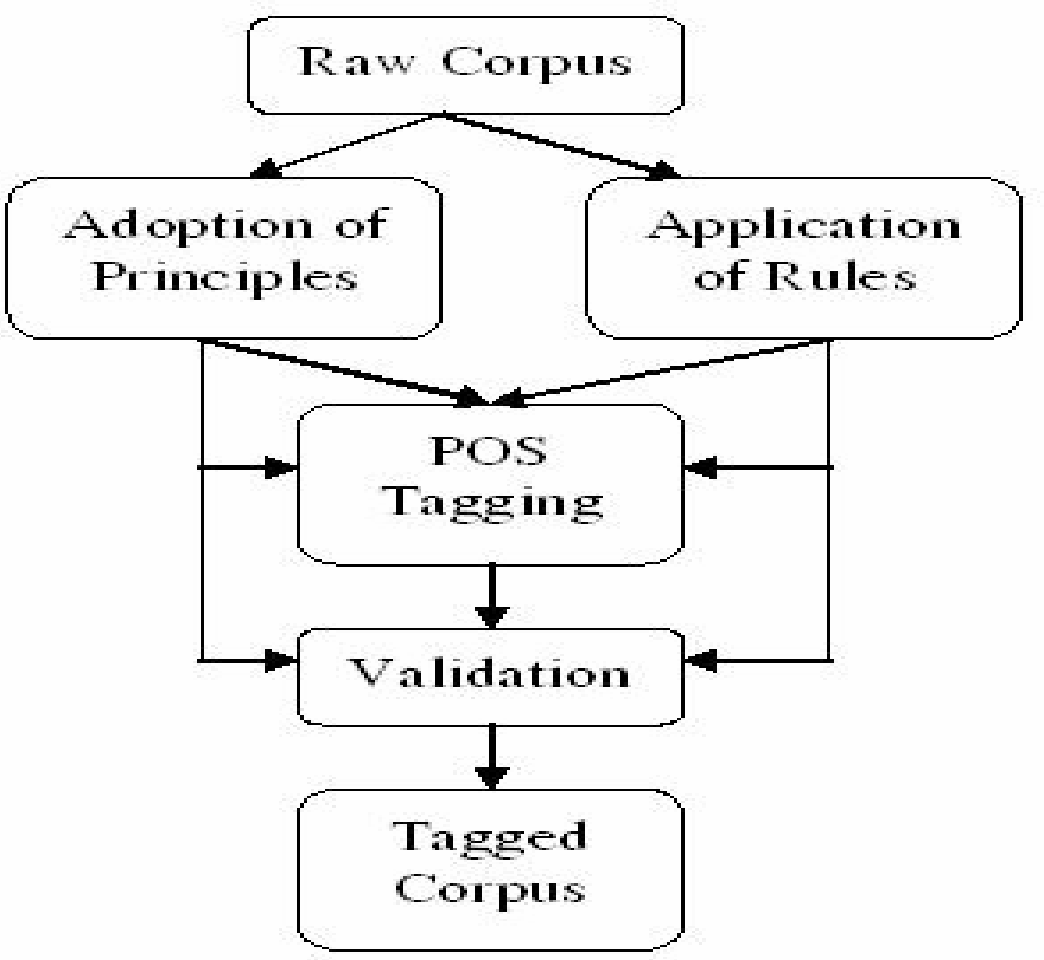

Fig. 3: Generic framework used in POS tagging

A native person engaged in assigning POS to words manually, can do the work successfully if she is well versed in morphology and grammar of the language.

A computer program can do this task successfully if it is supplied with adequate amount of linguistic information, data and rules for POS tag assignment as well as it is trained properly to do the work with minimum errors.

A system designer engaged in the task of designing a tool or a system for automatic POS tag assignment to words should be well-equipped with adequate amount of linguistic as well as grammatical knowledge of the language so that he can design exhaustive rules as well as develop a robust and accurate system to assign correct POS to words, terms, and other lexical items used in the corpus.

7. Conclusion: Value of POS Tagged Corpus

Importance of a POS tagged corpus in enormous in language description, natural language processing, and language technology.

In NLP it is used in grammar checking, named entity recognition and extraction, sentence parsing, word sense disambiguation, query addressing, machine learning, machine translation, language modelling, text understanding, information retrieval, extraction of grammatical properties and elements, E-learning, E-governance, and other works.

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

Although we visualize many more applications of a POS tagged corpus in Indian languages, till date, not much effort has been initiated to develop such a highly useful linguistic resource.

In recent past an effort is initiate by *DIT, Govt. of India* (under *Indian Languages Corpora Initiative:* 2009-2011) to develop parallel translation corpora in some Indian languages; and a major part of the project is to develop tagged text corpora for the languages included in the project.

Whatever have been done so far for the Indian languages, the rate of accuracy is far below if compared with POS tagged corpora of English (Bharathi and Mannem 2007).

For instance, in one million words English text database of *American National Corpus* the rate of accuracy is 97% to 98%, whereas for the 10000 to 100000 words corpus of Bengali the rate of accuracy is 85% to 90% (Dandapat 2009).

It clearly indicates that we sincerely need to take initiatives in this direction to develop POS tagged corpora for the Indian languages with two immediate goals: design maximally accurate tagset to increase rate of accuracy of the POS tagged data, and develop the POS tagged corpora in a large scale covering all text types for future linguistic works.

Thank you