A Common Parts-of-Speech Tagset Framework for Indian Languages

Sankaran Baskaran[1] , Kalika Bali[1], Tanmoy Bhattacharya[2], Pushpak Bhattacharyya[3], Monojit Choudhury[1], Girish Nath Jha[4], Rajendran S.[5], Saravanan K.[1], Sobha L.[6], and KVS Subbarao

[1]Microsoft Research India, Bangalore [2] Delhi University, Delhi [3]IIT Bombay, Mumbai [4]Jawaharlal Nehru University, New Delhi [5]Tamil University,Thanjavur [6]AU-KBC, Chennai

E-mail: rudhbaskaran@yahoo.com, kalikab@microsoft.com, tanmoy1@gmail.com, pb@cse.iitb.ac.in, monojitc@microsoft.com, girishj@mail.jnu.ac.in, raj_ushush@yahoo.com, v-sarak@microsoft.com, sobha@au-kbc.org, kvs2811@yahoo.com

Abstract

We present a universal Parts-of-Speech (POS) tagset framework covering most of the Indian languages (ILs) following the hierarchical and decomposable tagset schema. In spite of significant number of speakers, there is no workable POS tagset and tagger for most ILs, which serve as fundamental building blocks for NLP research. Existing IL POS tagsets are often designed for a specific language; the few that have been designed for multiple languages cover only shallow linguistic features ignoring linguistic richness and the idiosyncrasies. The new framework that is proposed here addresses these deficiencies in an efficient and principled manner. We follow a hierarchical schema similar to that of EAGLES and this enables the framework to be flexible enough to capture rich features of a language/ language family, even while capturing the shared linguistic structures in a methodical way. The proposed common framework further facilitates the sharing and reusability of scarce resources in these languages and ensures cross-linguistic compatibility.

1. Introduction

Parts-of-Speech (POS) tagging – the task of assigning appropriate POS tag to every word in a given text – is an important process used as a building block for several Natural Language Processing (NLP) tasks. A POS tagset usually defines the list of morphosyntactic categories that are applicable at the word-level to a specific language, though in some cases they may also include pure morphological or syntactic information. The early efforts in POS tagset design in 1970s, that resulted in tagsets such as UPenn, Brown and C5, mainly focused on tagsets for English which were mostly simple lists of tags corresponding to the morphosyntactic features, and varied greatly in terms of their granularity (Hardie, 2004).

CLAWS2 tagset (Sartoni, 1987) marked an important change in the structure of POS tagsets from a flat structure with unitary tags to a hierarchical structure that allowed for decomposable tags. This enabled the provision of distinct encoding for all word classes with distinct grammatical behaviour and a systematic approach in building the final tag from its constituent. The publication of EAGLES recommendations for morphosyntactic annotation of corpora (Leech and Wilson, 1996) was an earliest attempt to develop a common tagset guideline for several European languages. The objective of EAGLES1 was to standardise the tagsets used in different projects and/or different languages to achieve *cross-linguistic compatibility*, *reusability* and *interchangeability*. According to Leech and Wilson (1999), at the cross-linguistic level annotations used for one language should as far as possible be "compatible" with annotations used for another. This means that any common descriptive categories between different languages should be encoded with the same string so as to be recoverable from the annotations applied to texts in different languages.

Though several tagsets have been developed for Indian Languages (IIIT-H, AU-KBC), a majority of these are designed for specific languages in a flat structure capturing only coarse-level categories. In this paper, we present a common POS-tagset framework for ILs, which has been designed to cover the morphosyntactic details of Indian Languages and offers advantages such as flexibility, cross-linguistic compatibility and reusability.

In the next section we will discuss briefly the nature of Indian languages leading to the discussion in Section 3 on the need for a Common POS tagset framework for ILs (IL-POSTS). Section 4 lays out the design methodology followed for designing IL-POSTS framework followed by a description of the framework in Section 5. Section 6 describes some efforts in designing language-specific tagsets within the IL-POSTS framework. In section 7 we conclude with current status and future plans.

2. Indian Languages

There are four main language families found in India, viz., Austro-Asiatic, Dravidian, Indo-Aryan and Tibeto-Burman, of which Dravidian and Indo-Aryan (IA) form the largest group of languages spoken in the sub-continent. This framework concentrates on Dravidian and IA language families for two main reasons: (i) practical issues of manageability, (ii) the fact that of the 22 official languages in India a large majority belonged to these two language families. However, the detailed linguistic analysis and discussions that led to the design of this framework leads us to believe that it is broad enough to cover Indian Languages from the other language families as well.

As distinct language families, it may be argued that Dravidian and IA would have very different morphosyntax that would be difficult to capture in a single framework. The issue of having two separate

frameworks corresponding to these two families was discussed at the earlier stages of this work. While it is true that the two language families have many differentiating features at every level of linguistic analyses, it is striking to note a number of typological similarities that allow for a common framework. For example, Dravidian languages are agglutinating in nature while IA languages are typologically defined as inflectional. However, we can find some level of agglutination in some of the major IA languages like Marathi and Bangla.

Tamil, one of the major Dravidian languages, has many instances of agglutinative word formation such as 'படிக்கமுடியாதவர்களுக்காகவே'(படி+க்க+முடி+ஆத+வர்+கள்+உக்காக+வே) [padikkakudiyaathavarkaLuk- kaagavee] 'for the sake of only those who cannot read'(read+infitive+accomplish+neg+nominalizer+plural +benifective case +emphatic).

Bangla also supports word formations such as ভেতরেরটাকেইতো (ভেতর+এর+টা+কে+ই+তো) [bhetorer-taakeito] 'to the one that is inside (inside+possessive +classifier+accusative+emphatic+clitic)' which is clearly agglutinative.

Both language families follow SOV structure and have a rich morphology for case. Some other similarities between the two include (a) distinctions made between adjectives, quantifiers and nouns, (b) nouns specifying location or time that can also act as adverbs and postpositions, (c) use of postpositions rather than prepositions, (d) elaborate verb morphology marking for tense, aspect, mood, gender, number, person etc.

Thus, despite strong differences there are sufficient similarities between the two language families to warrant a single framework for POS tagsets.

3. Need for a Common POS Tagset for ILs

Some of the earlier POS tagsets mentioned previously were designed for English (Greene and Rubin, 1981; Garside, 1987; Santorini, 1990) and remain in popular usage even today. However, even though they were designed for the same language, they differ significantly from each other so that a corpus tagged by one is totally incompatible with the other. Further, as these are English-specific they cannot be reused for any other language without substantial changes.

For tagsets to be reusable across languages and corpora, they require standardization. Leech and Wilson (1999) put forth a strong argument for the need to standardize POS tagset for *reusability* of annotated corpora and *interoperability* across corpora in different languages. EAGLES guidelines (Leech and Wilson, 1996) were a result of such an initiative to create standards that are common across languages that share morphosyntactic features.

Several POS tagsets have been designed by a number of research groups working on Indian Languages though very few are available publicly (IIIT-tagset, AU-KBC Tamil tagset). However, as each of these tagsets have been motivated by specific research agenda, they differ considerably in terms of morphosyntactic categories and features, tag definitions, level of granularity, annotation guidelines *etc*. Moreover, some of the tagsets (e.g., the AU-KBC Tamil tagset) are language specific and do not scale across other Indian languages. This has led to a situation where despite strong commonalities between the languages addressed, resources cannot be shared due to incompatibility of tasgets. This is detrimental to the development of language technology for Indian languages which already suffer from a lack of adequate resources in terms of data and tools.

In IL-POSTS an attempt is made to treat equivalent morphosyntactic phenomena consistently across all languages. The hierarchical design, discussed in detail in the next section, also allows for a systematic method to annotate language specific categories without disregarding the shared traits of the Indian languages.

4. Design Methodology

The design methodology of IL-POSTS is based on the EAGLES guidelines (Leech and Wilson, 1996). At the initial stages we analysed the suitability of adapting EAGLES guidelines for ILs with minor modifications as needed. This was primarily done to see whether a framework defined for a different set of languages (European languages) can be applicable to another set of typologically dissimilar languages like IA and Dravidian. However, it soon became apparent that EAGLES could only be used as a model and could not be extended to typologically different languages without revision.

The design process involved a series of workshops where a working group of linguists, computational linguists and computer scientists systematically analysed each of the major categories, using corpora extensively in the analysis process. Some of the principles described in the following sub-sections are discussed in more details elsewhere (Baskaran et al., 2007).

4.1 Hierarchical Structure

Flat tagsets are usually lists of mutually exclusive categories. Though they may be easier to process they cannot capture higher level of granularity without an extremely large list of independent labels. Further, they are difficult to modularise and scale across languages as there is no provision for feature reusability at the level of morphosyntax. This means that corpora annotated with different tagsets are incompatible with each other. Most of the popular English tagsets (including UPenn, Brown, C5 and C7) and the existing IL tagsets (IIIT-H, AU-KBC) fall under this type.

The categories in a *hierarchical*, on the other hand, are *structured* relative to one another (Hardie, 2004). This implies that instead of having a large number of independent categories, a hierarchical tagset contains a small number of categories at the top level, each of which has a number of sub-categories in a tree-structure. The morphosyntactic details are encoded in the separate layers of a hierarchy; beginning from the major categories in the top and gradually progressing down to cover morphosyntactic features. This hierarchical arrangement allows the selective inclusion and removal of features for a specific language/ project, thereby keeping the

framework *a common standard* across languages/ projects.

Decomposability is another desirable feature of a hierarchical tagset design as it allows different features to be encoded in a tag by separate sub-strings. A tag is considered *decomposable* if the string representing the tag contains one or more shorter sub-strings that are meaningful out of the context of the original tag. Decomposable tags help in better corpus analysis (Leech, 1997) by allowing to search with an underspecified search string.

IL-POSTS has a hierarchical layout of decomposable tags with three levels in the hierarchy viz., categories, types (subcategories) and attributes (features).

4.2 Derivation of Language-specific Tagsets

IL-POSTS is a framework for ILs that allows language specific tagsets to be derived from it. An important consideration for its hierarchical structure and decomposable tags is that it should allow users to specify the morphosyntactic information applicable at the desired granularity according to the specific language and task. Thus, IL-POSTS offers broad guidelines for users to define their own tagset for a particular language and/ or a specific application. While designing a tagset, a user will have liberty to choose only those types and attributes that are applicable to his/her requirements. Sibling types/attributes can be selectively included in the tagset, but not the dependent features. In other words, turning off (leaving out) a type/attribute will disallow other attributes/values listed under it. The user can also customize the tagset to their requirement by adding additional attributes as special extensions.

4.3 Morphosyntactic Encoding

As Parts-of-Speech annotation is the primary objective of the IL-POSTS design, the framework encodes only the morphosyntactic categories and not syntactic, semantic or discourse information. This implies that the required morphosyntactic information for tagging should be typically identifiable directly from the given word form including its morphemic composition and not require any *a priori* information about the syntactic structure or semantic "meaning". For example, the number and gender information for nouns in most languages can be identified by the explicit suffixes and hence are morphosyntactic in nature. However, whether or not a noun functions as an agent in a particular sentence would require parsing at a higher level and hence, not coded within the framework.

4.4 Form versus Function

The function of a word offers significant clues for the subsequent stages of processing (e.g. parsing) and hence cannot be missed. Here, we adopt a balance between the form and the function of a word in a systematic and consistent way. Based on our analysis, when a word is morphologically derived from other words then we propose to tag them by their *function.* In all other cases, the words are tagged by their *form*. Thus, for example, the infinitive form of the Hindi verb "रहना" [rahanaa] "to stay", is tagged as verbal noun in the example "रहने का कमरा चाहिए" [rahane kaa kamaraa chaahiye] "(subj) needs a room to stay" (Literally, "staying room is needed").

However, in "मुझे होटल में रहना है" [mujhe hotal me rahanaa hai] "I want to stay in a hotel" it is marked as a main verb with an infinitive attribute.

Similarly, in Tamil, the same form of a verb "பாடு" [paadu] 'sing' in "பாடும் பறவை" [paadum paRavai] 'singing bird', is tagged as relative participle, but in "பறவை பாடும்" [paRavai paadum] 'bird will sing', is tagged as a verb.

4.5 Morphological Granularity

Indian languages have a complex morphology with varying degrees of richness. Some of the languages, such as those belonging to the Dravidian family are also agglutinative and hence can have many instances where it is impossible to assign a single tag to a word. This implies that morphological analysis is a desirable pre-processing step for automatic POS tagging to achieve better results. We encode all possible morphosyntactic features in our framework assuming the existence of morphological analysers and leave the choice of granularity to the users. As pointed out by Leech (1997) some of the linguistically desirable distinctions may not be feasible computationally. Therefore, we ignore certain features that may not be computationally feasible at the level of POS tagging.

4.6 Multi-word Expressions

Tagging of multi-word expressions (MWE) is an issue in ILswhere compounding is a very productive process not only for nominals, but for almost all the POS categories. Traditionally, multi-word compounds are clubbed together into a single category and almost all existing IL tagsets have a separate category for at least compound nouns and verbs. However, IL-POSTS treats constituents of MWEs, such as *Indian Space Research Organization*, as individual words and tags them separately rather than giving a single tag to the entire word sequence. This is done because we believe that grouping of MWE should be done at the level of chunking and cannot be handled efficiently during POS tagging. Also, it may not be possible to identify boundaries of the multi-word expressions in a given sentence without parsing it, at least partially, which in turn would require POS tagging.

4.7 Theoretical Neutrality

As Leech (1997) points out, an annotation scheme should be theoretically neutral to make it clearly understandable to a larger group and for its wide applicability. Further, as is the case of IL, terminological differences between the different grammatical traditions in different languages can also be a source of confusion. Hence, this framework is largely based on computational needs rather than any specific grammatical tradition.

4.8 Mapping with Existing Tagsets

One of the goals of the IL-POSTS framework is to allow the mapping of different tagsets to each other, thus, allowing different corpora tagged with disparate tagsets of the same language to be reused and also achieving cross-linguistic compatibility between different language corpora.

In general, any two different tagsets are most often incompatible with each other unless they are identical. While mapping it is possible that individual tags may be underspecified in either of the two tagsets. The IL-POSTS tags are designed such that whenever the source tagset is underspecified, the corresponding feature can be considered irrelevant and marked as "0" in the IL-POSTS derived tagset. Alternatively, in some cases where the IL-POSTS derived tagset is underspecified, the missing information can be added as an attribute under special extensions. For example, certain tagsets define transitive-intransitive distinctions for verbs, which is absent in ours and in this case it can be handled by adding *transitivity* as an attribute under special extensions category to achieve smooth mapping.

5. IL-POSTS Framework

The IL-POSTS framework is laid out in a hierarchy of three levels:

(i) Categories are the highest level part-of-speech classes. All categories are **Obligatory**, that is, are generally universal for all languages and hence, must be included in any morphosyntactic tagset derived from the framework.

(ii) Types are sub-classes of categories and are **Recommended,** that is, are recognised to be important sub-classes common to a majority of languages. Some types may also be **Optional** for certain languages.

(iii) Attributes are morphosyntactic features of Types. All attributes are optional, though in some cases they may be recommended.

Further, **Special extensions** to attributes provide for *features* to be specified for future use that are not covered in the currently defined list of attributes. These can be *generic attributes* that may be needed for a special purpose including those outside the scope of morphosyntax, and *language-specific attributes* that may be applicable to only a very small group of or even a single language(s).

All the tags were discussed and debated in detail by a group of linguists and computer scientists/NLP experts for eight Indian languages, viz. Bangla, Hindi, Kannada, Malayalam, Marathi, Sanskrit, Tamil and Telugu.

In the following section, we present the IL-POSTS tagset and its different levels of hierarchy

5.1 Categories, Types and Attributes

There are 11 categories (including the punctuations and residual categories) that are identified as universal categories for all ILs and hence, these are obligatory for any tagset derived from IL-POSTS.

All categories with the exception of Punctuations have sub-classes called Types which can have a number of attributes associated with each of them.

Categories	**Types**	**Attributes**
Noun (N)	Common (C)	1,2,5,6,11,12,13,14
	Proper (P)	1,2,5,6,11,12,13,14
	Verbal (V)	5,6,11,12
	Spatiotemporal (ST)	5,6,11,12,13
Verb (V)	Main (M)	1,2,3,4,7,8,9,12,13,16
	Auxiliary (A)	1,2,3,4,7,8,9,12,13,16
Pronoun (P)	Pronominal (PR)	1,2,3,5,6,10,11,12,13, 15,16
	Reflexive (RF)	1,5,6,11,12,13,15,16
	Reciprocal (RC)	1,5,6,11,12
	Relative (RL)	2,5,6,10,11,12,13,15,16
	Wh (WH)	2,5,6,10,11,12,13,15,16
Nominal Modifiers (J)	Adjectives (J)	1,2,5,13
	Quantifiers (Q)	1,2,5,11,12,13,17
Demon-Stratives (D)	Absolute (AB)	2,5,12,14
	Relative (RL)	2,5,12,14
	Wh (WH)	2,5,12,14
Adverb (A)	Manner (MN)	5,6,11,12,13
	Location (LC)	5,6,11,12,13
Participle (L)	Adjectival (RL)	1,2,4,12,13
	Adverbial (V)	12,13
	Nominal (N)	1,2,4,7,12,13
	Conditional (C)	12,13,18
Postposition (PP)		1,2,6,11,12
Particles (C)	Coordinating (CD)	
	Subordinating (SB)	
	Classifier (CL)	12
	Interjection (IN)	
	Others (X)	12
Punctuation (PU)		
Residual (RD)	Foreign word (F)	
	Symbol (S)	
	Other (X)	

Table 1: Categories, Types and Attributes defined in IL-POSTS framework

There are 18 attributes currently defined in the IL-POSTS framework. The attributes can be either binary or multi-valued.

Table 1 shows the different categories and their types along with the attributes that are relevant for them. It may be noted that this is the maximal set of attributes that are applicable to a particular type across languages. In Table 2 we present the values that the attributes can take.

5.2 Decomposability of Tags

The IL-POSTS framework recommends the use of decomposable tags, such as "*NC.sg.loc.n.n*", where *N* stands for the category "noun", *C* stands for the type "common" and the attribute values are specified in order separated by dots. In this specific example, *sg* implies that the number is singular, *loc* implies that the case-marker is locative, and the two 'n's imply that classifier and emphatics are not present (i.e., their values are "No"). While designing the tags, the following principles have been adopted.

- Each of the Categories and Types is represented by a unique single letter or two-letter combination. These tags are in uppercase.
- We also make sure that after the concatenation of a Category and its Type, the resultant string never exceeds three characters.
- The values of the Attributes are also assigned 1 to 4 character unique strings of letters or numbers.

Note that these letter based tags are for the ease of human annotation, editing and manual inspection phases. Nevertheless, for the purpose of machine readability and compact storage, the tags could be a simple string of numbers and characters (e.g., "*N11500*" instead of "*NC.sg.loc.n.n*").

6. Deriving Language Specific Tagsets

Currently, the IL-POSTS framework has been used to derive language specific tagsets for three Indian languages, viz. Bangla and Hindi (IA), and Tamil (Dravidian). The tagsets were designed after discussions with linguists and deep analyses of the morphosyntax of the languages concerned. We followed an iterative process where each draft version was followed by manual tagging exercises. As Hardie (2004) points out, manual tagging of data is crucial to language-specific tagset design to validate whether or not the tags reflect the morphosyntax of the language. Applying the tags to natural language data helps to uncover ambiguities and other issues that might not have been covered in the tags and helps establish the annotation principles for that language. Another issue with hierarchical tagsets is that of cognitive load on the annotator and this process helps in the optimal design of an annotating tool. Manual tagging of data also aids in deciding which attributes can be automatically derived from the text, thereby decreasing the load on the human annotator.

Our initial experience with manual tagging indicates that while hierarchical tagsets do contain a large number of tags for the annotator to choose from, an annotation tool can be designed taking into account the hierarchy associated with each top-level category to minimize the load on the annotator. We are currently in the process of studying this in a systematic manner.

A number of language-specific issues were also uncovered in this process, which required some modification of the tagset and some instances and certain open issues are discussed in the following sub-sections. A thorough discussion of all the issues is beyond the scope of this paper.

No.	Attribute	Values
1	Gender (Gen)	Masculine (mas), Feminine (fem), Neuter (neu)
2	Number (Num)	Singular (sg), Plural (pl), Dual (du)
3	Person (Per)	First (1), Second (2), Third (3)
4	Tense (Tns)	Present (prs), Past (pst), Future (fut)
5	Case (Cs)	Direct (dir), Oblique (obl)
6.	Case-marker (Csm)	Ergative (erg), Accusative (acc), Instrumental (ins), Dative (dat), Genitive (gen), Sociative (soc), Locative (loc), Ablative (abl), Benefective (bnf), Vocative (voc), Purposive (pur)
7	Aspect (Asp)	Simple (sim), Progressive (prg), Purposive (prf)
8	Mood (Mood)	Declarative (dcl), Subjunctive (sbj), Conditional (cnd), Imperative (imp), Presumptive (psm), Abilitative (abt), Habitual (hab)
9	Finiteness (Fin)	Finite (fin), Non-finite (nfn), Infinite (inf)
10	Distributive (Dstb)	Yes (y), No (n)
12	Emphatic (Emph)	Yes (y), No (n)
13	Negative (Neg)	Yes (y), No (n)
14	Distances (Dist)	Proximal (prx), Distal (dst), Sequel (seq)
15	Incl/Excl (Set)	Inclusive (inl), Exclusive (exl)
16	Honorificity (Hon)	Yes (y), No (n)
17	Numeral (Nml)	Ordinal (ord), Cardinal (crd), Non-numeral (nnm)
18	Realis (Rls)	Yes (y), No (n)

Table 2: Attributes and their values

6.1 Language Specific Issues

In this section we describe a specific issue with Verbs that was brought up during the annotation and its resolution that required modification of the earlier version of IL-POSTS (Baskaran, 2008)

In Bangla, a combination of a non-finite followed by a finite verb can have several different morphosyntactic functions. For example, "মেরে ফেলল" [mere phello] 'kill+non-finite throw+finite' can mean 'threw after killing' (in which case "মেরে" is a sequential participle) or just 'killed' with a completive sense (where "মেরে" is the polar verb and "ফেলল" the vector verb of a finite verb group). On the other hand, constructs like "হেসে বলল" [henshe bollo] 'smile+non-finite say+finite' might mean 'said while smiling' (where "হেসে" is functioning as an adverbial participle). Similarly, it is hard to distinguish between the adjectival participle and verbal nouns. For

instance, "খাওয়া" [khaoyaa] might mean 'eaten' (adjectival participle) as well as 'eating' (verbal noun). Thus, the form of a verb does not contain the information regarding its function, that is, whether it is a sequential, adjectival or adverbial participle, or simply a polar-vector combination or verbal noun.

In Tamil, a combination of a non-finite followed by a finite verb can also play several different morphosyntactic roles as in Bangla. For example "படித்து போ" [padiththu poo] 'read+non-finite go+finite' can mean 'read and go' (in which case "படித்து" [padiththu] 'read+non-finite' is a sequential participle) or just 'read' with a completive sense (where "படித்து" [padiththu] 'read+non-finite' is the polar verb and "போ" [poo] 'go+finite' the vector verb of a finite verb group). On the other hand, constructs like "பார்க்கையில் கூறினான்" [paarkkaiyil kooRinaan] 'see+non-finite say+finite' might mean 'said while seeing' (where "பார்க்கையில்" [paarkkaiyil] is functioning as an adverbial participle.

Further, in Hindi, the finite imperfect or simple verb form – Verb + *ta* can also function as a participle. For example, in "राम सेब खाता" [raam seb khaataa hai] 'Ram eats an apple', "खाता" [khaataa] 'eat+finite+imperfect' is an imperfective verb whereas in "सेब खाता लड़का" it functions as a participle. The problem is further complicated when we take into account the Verb + (y)a form as in "आया" [aayaa] "come + finite+perfect, which in "राम आया है" [raam aayaa hai] 'Ram has come' is a perfective verb but in "आया हुआ लड़का राम है" [aayaa huaa laDkaa raam hai] "the boy who has come is Ram" is a participle. Also, as in the examples above, in constructs like "ले आया" [le aayaa] 'get+non-finite come+finite', can be a sequential verb construction meaning "(subj.) got and came" or a polar+vector combination meaning "got" where "आया" [aayaa] acts as a vector giving a sense of finality. In such instances, it is almost impossible to disambiguate the different functions of the verb.

As many instances of such ambiguity were discovered from the data, it became evident that disambiguation between these various cases often requires semantic information. This led us to decide in favor of form based tagging of verbs for IL, where we mark only the types – main verb and auxiliary verb, and the finiteness is marked as an attribute.

6.1 Some Open Issues

The word "বলে" [bole] in Bangla has several different syntactic functions. It can act as a finite ('say') or non-finite ('having said') verb, conjunction ('so') or a quotative particle. However, it is difficult to disambiguate among these different functions, especially between the finite verb and quotative usages, from the local context of the word. For instance, in the fragment "রাম বলে একটা ছেলে…"[raam bole ektaa chhele…] 'A boy called Ram …' or 'Ram says the boy …', "বলে" could be a quotative as well as a finite verb. The disambiguation is not possible unless the whole sentence is considered. Thus, it is not clear whether to distinguish between these two functions, which otherwise seems to be absolutely necessary for the purpose of parsing.

In Bangla, the classifiers (e.g., "টা" [taa], "গুলি" [guli]) can be added to the adjectives, adverbs, quantifiers, pronouns and all types of nouns. When added to adjectives or adverbs, the resultant word usually functions as a noun (e.g., "লালটা" [laaltaa] 'the red one', "এখানটা" [ekhaantaa], 'this place'). However, if added to quantifiers the resulting words can function as both noun/pronoun and quantifiers. For instance, in "এতগুলো বই ছিল, এখন একটাও নেই" [etagulo boi chhilo, ekhan ektaao nei] 'there were so many books, but none now', the first quantifier+classifier combination (এত+গুলো) is acting as a quantifier ('so many'), whereas the second one (এক+টা+ও) is acting as a noun/pronoun ('none'). Therefore, it is not clear whether in such cases, we should annotate by form or function.

In the Tamil tagset, some of the clitic markars "உம்" (um) – coordinator, "ஓ" (o) – or (also coordinator), "ஆ" (aa) – interrogative are missing as they are not available in the IL-POSTS framework. If we were to add these to the framework, we could have a separate attribute called clitic with values coordinator, emphatic and interrogatives. However, a word can have more than one clitic marker for example "அவனையுமா" [avanaiyumaa] 'he+accusative case+inclusive or coordinator clitic+interrogative clitic' which might mean 'is him also' and takes two clitics. But as per the IL-POSTS framework a word can have an attribute with a value only once. This creates a problem for tagging these clitics. Also, as the clitic information is needed for higher level analysis, for example during parsing, one cannot tag the first clitic and ignore the other clitics.

Currently, we are discussing as a part of the annotation guidelines for the language specific tagsets, how best to deal with such issues.

7. Conclusion

In this paper we have presented IL-POSTS - a common parts-of-speech framework for Indian Languages. This hierarchical framework allows designing language-specific tagsets that are interoperable and flexible. The framework has been used for deriving three language specific tagsets and we are currently involved in manual annotation of a significant amount of data using these tagsets. These datasets along with the annotation guidelines for specific languages with recommendations for dealing with ambiguities will be released in the public domain in the near future. Mapping of the IL-POSTS to some of the existing IL-tagsets is also being undertaken to enhance the interoperability of the framework and reuse of existing corpora. IL-POSTS framework hopes to provide researchers working with Indian Languages with a systematic and flexible way to not only share data but also tools across languages and applications.

8. References

Baskaran, S et al. (2008). Designing a Common POS-Tagset Framework for Indian Languages. Paper presented in *The 6th Workshop on Asian Language Resources.* Jauary, 2008, Hyderabad.

Baskaran, S. et al. (2007). Framework for a Common Parts-of-Speech Tagset for Indic Languages. Available in http://research.microsoft.com/~baskaran/POSTagset

Cloeren, J. (1999) Tagsets. In *Syntactic Wordclass Tagging*, ed. Hans van Halteren, Dordrecht: Kluwer Academic.

Hardie, A. (2004). The Computational Analysis of Morphosyntactic Categories in Urdu. PhD Thesis submitted to Lancaster University.

Greene, B.B. and Rubin, G.M. (1981). Automatic grammatical tagging of English. Providence, R.I.:Department of Linguistics, Brown University.

Garside, R. (1987) The CLAWS word-tagging system. In *The Computational Analysis of English*, ed. Garside, Leech and Sampson, London: Longman.

Leech, G. (1997). Grammatical Tagging. In Corpus Annotation: Linguistic Information for Computer Text Corpora, ed: Garsire, Leech, and McEnery, London: Longman.

Leech, G and Wilson, A. (1996), Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Report EAG-TCWG-MAC/R.

Leech, G and Wilson, A. (1999). Standards for Tag-sets. In *Syntactic Wordclass Tagging*, ed. Hans van Halteren, Dordrecht: Kluwer Academic.

Sampson G. 1987. Alternative Grammatical Coding Systems. In *The Computational Analysis of English*, R.G. Garside et al. (eds.), London: Longman.

Santorini, B. (1990). Part-of-speech tagging guidelines for the Penn Treebank Project. Technical report MS-CIS-90-47, Dept. Of Computer and Information Science, University of Pennsylvania.

IIIT-Tagset. A Parts-of-Speech tagset for Indian Languages. http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf

AU-KBC tagset. AU-KBC POS tagset for Tamil. http://nrcfosshelpline.in/smedia/images/downloads/Tamil_Tagset-opensource.odt