Designing a Common POS-Tagset Framework for Indian Languages

Sankaran Baskaran, Microsoft Research India. Bangalore. baskaran@microsoft.com
Kalika Bali, Microsoft Research India. Bangalore. kalikab@microsoft.com
Tanmoy Bhattacharya, Delhi University, Delhi. tanmoy1@gmail.com
Pushpak Bhattacharyya, IIT-Bombay, Mumbai. pb@cse.iitb.ac.in
Girish Nath Jha, Jawaharlal Nehru University, Delhi. girishj@mail.jnu.ac.in
Rajendran S, Tamil University, Thanjavur. raj_ushush@yahoo.com
Saravanan K, Microsoft Research India, Bangalore. v-sarak@microsoft.com
Sobha L, AU-KBC Research Centre, Chennai. sobha@au-kbc.org
Subbarao K V. Delhi. kvs2811@yahoo.com

Abstract

Research in Parts-of-Speech (POS) tagset design for European and East Asian languages started with a mere listing of important morphosyntactic features in one language and has matured in later years towards hierarchical tagsets, decomposable tags, common framework for multiple languages (EAGLES) etc. Several tagsets have been developed in these languages along with large amount of annotated data for furthering research. Indian Languages (ILs) present a contrasting picture with very little research in tagset design issues. We present our work in designing a common POS-tagset framework for ILs, which is the result of in-depth analysis of eight languages from two major families, viz. Indo-Aryan and Dravidian. Our framework follows hierarchical tagset layout similar to the EAGLES guidelines, but with significant changes as needed for the ILs.

1 Introduction

A POS tagset design should take into consideration all possible morphosyntactic categories that can occur in a particular language or group of languages (Hardie, 2004). Some effort has been made in the past, including the EAGLES guidelines for morphosyntactic annotation (Leech and Wilson, 1996) to define guidelines for a common tagset across multiple languages with an aim to capture more detailed morphosyntactic features of these languages.

However, most of the tagsets for ILs are language specific and cannot be used for tagging data in other language. This disparity in tagsets hinders interoperability and reusability of annotated corpora. This further affects NLP research in resource poor ILs where non-availability of data, especially tagged data, remains a critical issue for researchers. Moreover, these tagsets capture the morphosyntactic features only at a shallow level and miss out the richer information that is characteristic of these languages.

The work presented in this paper focuses on designing a common tagset framework for Indian languages using the EAGLES guidelines as a model. Though Indian languages belong to (mainly) four distinct families, the two largest being Indo-Aryan and Dravidian, as languages that have been in contact for a long period of time, they share significant similarities in morphology and syntax. This makes it desirable to design a common tagset framework that can exploit this similarity to facilitate the mapping of different tagsets to each other. This would not only allow corpora tagged with different tagsets for the same language to be reused but also achieve cross-linguistic compatibility between different language corpora. Most importantly, it will ensure that common categories of different languages are annotated in the same way.

In the next section we will discuss the importance of a common standard vis-à-vis the currently available tagsets for Indian languages. Section 3 will provide the details of the design principles behind the framework presented in this paper. Examples of tag categories in the common framework will be presented in Section 4. Section 5 will discuss the current status of the paper and future steps envisaged.

2 Common Standard for POS Tagsets

Some of the earlier POS tagsets were designed for English (Greene and Rubin, 1981; Garside, 1987; Santorini, 1990) in the broader context of automatic parsing of English text. These tagsets popular even today, though designed for the same language differ significantly from each other making the corpora tagged by one incompatible with the other. Moreover, as these are highly language specific tagsets they cannot be reused for any other language without substantial changes this requires standardization of POS tagsets (Hardie 2004). 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, 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 (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 this paper, we present a common framework for all Indian languages where 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 particular categories without disregarding the shared traits of the Indian languages.

3 Design Principles

Whilst several large projects have been concerned with tagset development very few have touched upon the design principles behind them. Leech (1997), Cloeren (1999) and Hardie (2004) are some important examples presenting universal principles for tagset design.

In this section we restrict the discussion to the principles behind our tagset framework. Importantly, we diverge from some of the universal principles but broadly follow them in a consistent way.

Tagset structure: *Flat tagsets* just list down the categories applicable for a particular language without any provision for modularity or feature reusability. *Hierarchical tagsets* on the other hand are structured relative to one another and offer a well-defined mechanism for creating a common tagset framework for multiple languages while providing flexibility for customization according to the language and/ or application.

Decomposability in a tagset allows different features to be encoded in a tag by separate sub-stings. Decomposable tags help in better corpus analysis (Leech 1997) by allowing to search with an underspecified search string.

In our present framework, we have adopted the hierarchical layout as well as decomposable tags for designing the tagset. The framework will have three levels in the hierarchy with categories, types (subcategories) and features occupying the top, medium and the bottom layers.

What to encode? One thumb rule for the POS tagging is to consider only the aspects of morphosyntax for annotation and not that of syntax, semantics or discourse. We follow this throughout and focus only on the morphosyntactic aspects of the ILs for encoding in the framework.

Morphology and Granularity: Indian languages have complex morphology with varying degree of richness. Some of the languages such as those of the Dravidian family also display agglutination as an important characteristic. This entails that morphological analysis is a desirable pre-process for the POS tagging to achieve better results in automatic tagging. We encode all possible morphosyntactic features in our framework assuming the existence of morphological analysers and leave the choice of granularity to 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 POS tagging level.

Multi-words: We treat the constituents of Multi-word expressions (MWEs) like *Indian Space Research Organization* as individual words and tag them separately rather than giving a single tag to the entire word sequence. This is done because: Firstly, this is in accordance with the standard practice followed in earlier tagsets. Secondly, grouping MWEs into a single unit should ideally be handled in chunking.

Form vs. function: We try to adopt a balance between form and function in a systematic and consistent way through deep analysis. Based on our analysis we propose to consider the *form* in normal circumstances and the *function* for words that are derived from other words. More details on this will be provided in the framework document (Baskaran et al 2007)

Theoretical neutrality: As Leech (1997) points out the annotation scheme should be theoretically neutral to make it clearly understandable to a larger group and for wider applicability.

Diverse Language families: As mentioned earlier, we consider eight languages coming from two major language families of India, viz. Indo-Aryan and Dravidian. Despite the distinct characteristics of these two families, it is however striking to note the typological parallels between them, especially in syntax. For example, both families follow SOV pattern. Also, several Indo-Aryan languages such as Marathi, Bangla etc. exhibit some agglutination, though not to the same extent of Dravidian. Given the strong commonalities between the two families we decided to use a single framework for them

4 POS Tagset Framework for Indian languages

The tagset framework is laid out at the following four levels similar to EAGLES.

I. **Obligatory** attributes or values are generally universal for all languages and hence must be included in any morphosyntactic tagset. The major POS categories are included here.

II. **Recommended** attributes or values are recognised to be important sub-categories and features common to a majority of languages.

III. **Special extensions**[1]

a. *Generic attributes* or values

b. *Language-specific* attributes or values are the attributes that are relevant only for few languages and do not apply to most languages.

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

Now, because of space constraints we present only the partial tagset framework. This is just to illustrate the nature of the framework and the complete version as well as the rationale for different categories/features in the framework can be found in Baskaran et al. (2007).[2]

In the top level the following 12 categories are identified as universal categories for all ILs and hence these are obligatory for any tagset.

1. [N] Nouns
2. [V] Verbs
3. [PR] Pronouns
4. [JJ] Adjectives
5. [RB] Adverbs
6. [PL] Participles
7. [PP] Postpositions
8. [DM] Demonstratives
9. [QT] Quantifiers
10. [RP] Particles
11. [PU] Punctuations
12. [RD] Residual[3]

The partial tagset illustrated in Figure 1 highlights entries in *recommended* and *optional* categories for verbs and participles marked for three levels.[4] The features take the form of attribute-value pairs with values in italics and in some cases (such as case-markers for participles) not all the values are fully listed in the figure.

5 Current Status and Future Work

In the preceding sections we presented a common framework being designed for POS tagsets for Indian Languages. This hierarchical framework has

[1] We do not have many features defined under the special extensions and this is mainly retained for any future needs.

[2] Currently this is just the draft version and the final version will be made available soon

[3] For words or segments in the text occurring outside the gambit of grammatical categories like foreign words, symbols,etc.

[4] These are not finalised as yet and there might be some changes in the final version of the framework.

three levels to permit flexibility and interoperability between languages. We are currently involved in a thorough review of the present framework by using it to design the tagset for specific Indian languages. The issues that come up during this process will help refine and consolidate the framework further. In the future, annotation guidelines with some recommendations for handling ambiguous categories will also be defined. With the common framework in place, it is hoped that researchers working with Indian Languages would be able to not only reuse data annotated by each other but also share tools across projects and languages.

References

Baskaran S. et al. 2007. Framework for a Common Parts-of-Speech Tagset for Indic Languages. (Draft) 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 and Wilson, A. 1996. Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Report EAG-TCWG-MAC/R.

Leech, G. 1997. Grammatical Tagging. In *Corpus Annotation: Linguistic Information from Computer Text Corpora*, ed: Garside, Leech and McEnery, London: Longman

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

Santorini, B. 1990. Part-of-speech tagging guidelines for the Penn Treebank Project. Technical report MS-CIS-90-47, Department 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

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

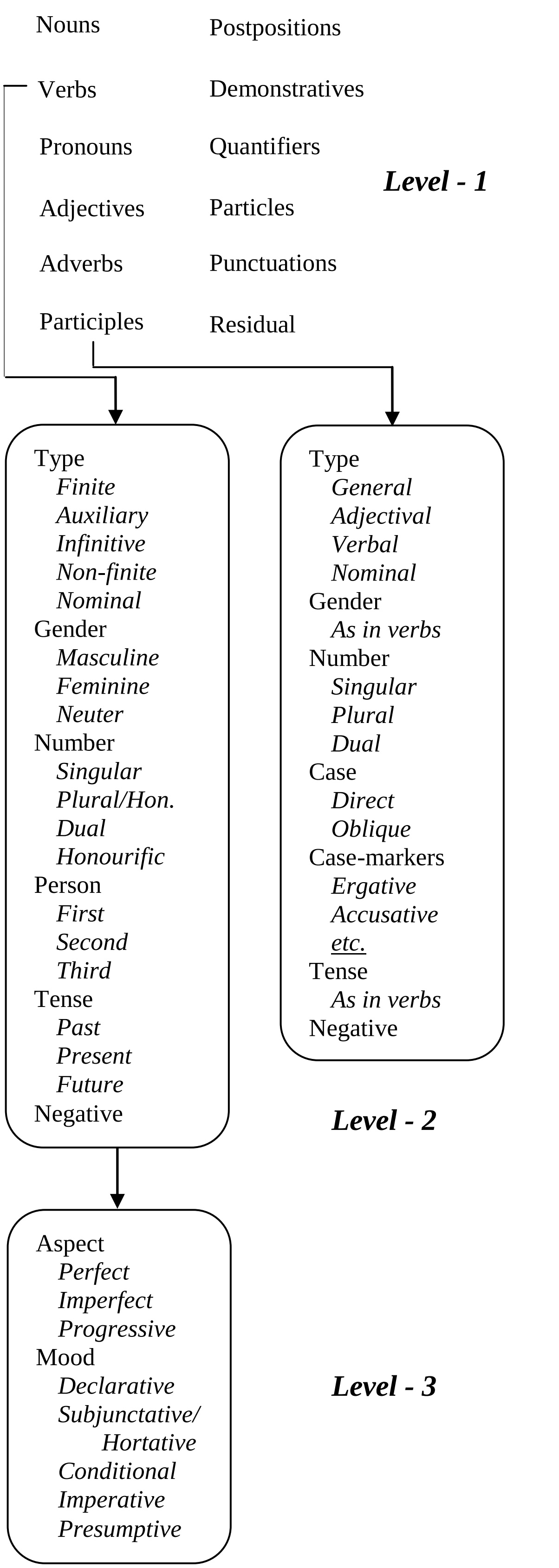

Fig-1. Tagset framework - *partial representation*