Chapter II

Part of Speech Tagset

2.0 Introduction

Tagset development forms a foundation of any computational processing endeavor. It is generally accepted that, as a prelude to syntactic analysis of natural language by computers, a text must be annotated with tags indicating the POS. The first pre-requisite for automated POS tagging is a tagset that is a set of exhaustive categories into which any token of the language can be placed. While the nature of the language is that there will always be words that are hard to classify, or are ambiguous between two categories, the tagset categories should be designed in such a way so as to minimize such problems. The fundamental problems in POS tagging task stem from the fact that a word can take different lexical categories depending on its context. The tagger has to resolve this ambiguity and determine the best sequence for a sentence. Tags are also applied to punctuation markers, thus tagging for natural language is the same process as tokenization for computer languages, although tags for natural languages are much more ambiguous.

Given the prominence of the USA both in linguistics and in computing technologies, the earliest work on tagsets in the 1960s and early 1970s occurred in the US and focused on English. The most important tagsets of this earliest period are those of Klein and Simmons 1963 and Greene and Rubin 1971. Over the course of time, sequence of tagsets for English have been devised such as the Penn tagset and CLAWS tagset including the series C_1, C_2, C_5, C_7. The publication of Eagles recommendations for morphosyntactic annotation of corpora (Leech

and Wilson 1996) was the earliest attempt to develop common tagset guidelines for several European languages. The Eagles project is concerned with Natural Language Processing (NLP) and as such it has a very wide theme, and needs to cater to the large number of circumstances in which text is used**.** The objective of Eagles guidelines was to standardize the tagsets used in different languages to achieve cross-linguistic compatibility, reusability and interchangeability.

2.1 English Tagsets

Research on Part-of-speech tagging has been closely tied to corpus linguistics. Most of the work done in this field is in English. The first major corpus of English for computer analysis was the Brown Corpus developed at Brown University by Kucera and Francis in the mid 1960s. It consists of about 1,000,000 words of running English prose text, made up of 500 samples from randomly chosen publications. The Brown corpus was painstakingly "tagged" with part-of-speech markers over many years.

The most important tagsets of this early period are those of Klein and Simmons 1963 and Green and Rubin 1971. Klein and Simmons designed a CGC ("Computational Grammar Coder") as a component of a parser. The CGC had three components: a lexicon, a morphological analyzer and a context disambiguator. Their tagset contains thirty tags and the CGC program also outputs information, separate to the main tag. The CGC algorithm reported 90% accuracy on applying a thirty tag tagset to different articles having different domains.

By contrast TAGGIT program by Green and Rubin 1971 introduced the idea of providing a text corpus annotated with part-of-speech information as a useful tool for linguistic research. The program was based on context pattern rules. The basic idea in the TAGGIT program was to associate with each word a set of potential tags, and then use the context to choose the correct one. TAGGIT used a 71 item tagset a disambiguation grammar of 3,300 rules. Together these rules made about 77 percent of all words in the million word Brown University Corpus Unambiguous; the remaining 23 percent of words remained to be resolved by human post editors.

These two early Tagsets display some consistent design features. Both Green and Rubin, and Klein and Simmons incorporate tags for punctuation marks, which are treated as words, a practice which has continued to the present day.

2.2 CLAWS1 Tagset

The next major advances took place in the late 1970s. The first effort in this new wave of data-driven statistical taggers was carried out as part of the annotation of Lancaster-Oslo-Bergen Corpus a one-million word corpus of British English designed to match the Brown Corpus in size, scope and structure. For annotating this corpus, a system called CLAWS1 was developed at the University of Lancaster (Marshall 1983; Garside; Leech and Sampson 1987).

The earliest CLAWS1 tagset also known as LOB[1] tagset was used in the tagging of the LOB corpus. Since this Corpus was designed to parallel the structure of Brown Corpus, the tags were also parallel, and CLAWS1 or the LOB tagset having 135 tags is very similar to the later version of the Brown Tagset (Francis and Kucera 1982). The second version of CLAWS (CLAWS2) was developed over the period 1983-86. The development of the CLAWS2 Tagset was motivated by two requirements: "Providing distinct coding for all classes of words having distinct grammatical behavior", and making the tagset more "systematic in the way that tags are built up from individual characters" (i.e. more decomposable and hierarchical) (Sampson 1987:167). As a result this tagset contains 166 tags. The major subsequent development in the CLAWS tagset were the C5 and C7 tagsets, developed for the tagging of the BNC[2] and the BNC Sampler (Leech, Garside and Bryant 1994; Leech 1997; Garside and Smith 1997; Smith 1997). The C7 tagset (146 tags) is more fine grained of the two and was used for the million word Sampler. The C7 tagset can be regarded as a further refinement of the CLAWS2 tagset. The C5 tagset is somewhat different from the

[1] The Lancaster-Oslo-Bergen Corpus (often abbreviated as LOB Corpus) was compiled in 1980s in collaboration between the University of Lancaster, the University of Oslo, and the Norwegian Computing Centre for the Humanities, Bergen, to provide a British counterpart to the Brown Corpus compiled by Kucera and Francis for American English in 1960s.

[2] The British National Corpus (BNC) is a 100 million word corpus collection of samples of written and spoken language from a wide range of sources, of modern British English for use in linguistic research. It is a collaborative, pre-competitive initiative carried out by Oxford University Press (OUP), Longman Group UK Ltd., Chambers, Lancaster University's Unit for Computer Research in the English Language (UCREL), Oxford University Computing Services (OUCS), and the British Library.

others, since it has far fewer tags (61). This was in order to make it useful to the greatest number of end users.

Cloren (1999:50) characterizes the C5 tagset as flat, i.e. non- hierarchical. In fact, although none of the CLAWS tagsets are laid out in the hierarchical fashion described by Cloren, the C7 tagset is hierarchical in conceptional terms (Leech 1997:27-28).

2.3 The TOSCA Scheme

The TOSCA (Tools for Syntactic Corpus Analysis) is an annotation project developed at the Katholieke Universiteit at Nijmegen, the Netherlands. The main aim of the project is the production of resources for linguistic research in the areas of syntax and language use. The TOSCA annotation scheme has been used in the analysis of the Nijmegen corpus, and of the TOSCA corpus, both of which consist of mainly written language. It is also being used for some parts of the ICE corpus, in which spoken language is also included. This scheme is described by Van Halteren and Oostdijk 1993 and includes a POS tagset. This Tagset differs considerably from CLAWS tagsets, firstly in its form i.e. it is made up of only 32 word class tags. However, most word classes allow sub classification to be annotated in a feature list following the tag, meaning that the actual number of combinations is much higher. The TOSCA tagset is also notable in that it makes many more distinctions relating to the syntactic function of the word than the CLAWS tagsets. E.g., there are three major word class tags for the

word ‘it’ depending upon whether it is a Pronoun, a formal ‘it’, a cleft ‘it’, or a provisional ‘it’ (Van Halteren and Oostdijk 1993:160).

2.4 The ICE Tagset

An important development from the TOSCA tagset is the ICE (International Corpus of English) tagset, described by Greenbaum and Yibin in 1996. It distinguishes 19 word classes (a substantial reduction) but, like the TOSCA tagset, gives most words a feature list as well as a major word class tag. This means that the tagset contains, in effect, around 260 tags. This tagset as well contains significant differences of classification from the CLAWS tagset: for example, the Verb ‘be’ is tagged as both an auxiliary and a Verb depending on its function (AUX and V being different major categories in this system of description).

2.5 The Penn Tagset

The Penn Treebank tagset (Marcus et al 1993:314-318) has been applied to the Brown Corpus and a number of other corpora. The Penn Tree bank tagset was culled from the original 87-tag tagset for the Brown corpus. The reduced set leaves out information that can be recovered from the identity of the lexical item. For example, the original Brown tagset and other large tagsets like C5 include a separate tag for each of the different forms of the Verbs ‘do’ (e.g. C5 tag ‘VDD’ for ‘did’ and ‘VDG’ for ‘doing’), ‘be’, and ‘have’. These were omitted from the Penn set.

Certain syntactic distinctions were not marked in the Penn Treebank tagset because Treebank sentences were parsed, not merely tagged, and so some syntactic information is represented in the phrase structure.

2.6 Eagles Guidelines

The Eagles guidelines were written for the languages of European Union. Eagles guidelines (Leech and Wilson 1999) outline a set of features for tagsets, in which some features are obligatory, some are recommended and some optional. Only one feature of an Eagles-compatible tagset is considered obligatory that of the major word categories, or parts-of-speech. Eagles suggest thirteen major categories like Nouns, Pronouns, Verbs, etc simultaneously, a scheme of encoding all these features into an "intermediate tagset" is given. This is encoding uses numerical values for the assorted Eagles attributes. The choice of how the features are encoded within a given Eagles-compliant tagset is left to the user, as long as the categories thus created can also be expressed using the intermediate tagset. The purpose of the intermediate encoding is to allow mapping between any two tagsets created in compliance with the Eagles guidelines, thus ensuring their compatibility. Eagles tags are defined as sets of morphosyntactic attribute-value pairs (e.g. Gender is an attribute that can have the values Masculine, Feminine or Neuter). The recommended and optional attributes are then organized according to these word categories, and do not necessarily correspond across word classes. For example, the first recommended attribute is Type (Common/Proper), Gender (Masculine/Feminine) for Nouns; Person (First/Second/Third) for Pronouns; Person (First/Second/Third), Gender (Masculine/Feminine), Tense

(Present/past/future) for Verbs. The recommended attributes also cover number, case, finiteness, voice, and other important features which are relevant to a range of languages. The optional part of the recommendations consists of similar attributes of more narrow applicability, and some additional values – mainly specific to one language or a small group of languages. Attributes can be important for a particular language or may be for two or three languages at the most, but do not apply to the majority of European languages. In practice, generic and language-specific features cannot be clearly distinguished.

The Eagles guidelines provide a flexible framework that in theory encompasses all the things which one would wish to mark up, without restricting the freedom of the tagset designer. It promotes consistency and reusability of linguistic resources for different languages and discourages "reinvention of the wheel".

2.7 Indian Tagsets

For Indian languages, several tagsets have been developed. The most prominent among those is that developed under ILMT (Indian Language Machine Translation) guidelines, which is designed for specific languages in a flat structure capturing only coarse-level categories. Another tagset which is designed for Indian languages is that of IL-POSTS (Indian Language Part of Speech Tagsets) hierarchical framework. IL-POSTS is a framework for Indian languages 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.

2.7.1 IIIT-H Tagset

IIIT-H Tagset is based on ILMT (Indian Language Machine Translation) guidelines. ILMT is a project in which a number of institutes have come together to form a consortium and their work focuses on developing (Machine Translation) MT systems for various Indian language pairs. The guidelines provided by ILMT are designed in such a way so that they can be easily used for any Indian language. The tagset provided by them is based on three main assumptions viz:

i. The tags should be common for all Indian languages.

ii. It should be comprehensive/complete.

iii. It should be simple.

Maintaining simplicity is important for the following two reasons:

a. Ease of Learning

b. Consistency in annotation

Another important point which was discussed by various scholars and experts was that POS tagging is NOT a replacement for morph analyser. A 'word' in a text carries the following linguistic knowledge:

a. Grammatical category and

b. Grammatical features such as gender, number, person etc.

The POS tag should be based on the ‘category’ of the word and the features can be acquired from the morph analyser.

Some of the issues which were handled and resolved by ILMT guidelines were:

I. Fineness Vs coarseness in linguistic analysis

II. Syntactic function Vs lexical category

III. New tags Vs tags from a standard tagger

I. Fineness Vs coarseness in linguistic analysis

It was decided to come up with a set of tags which avoids 'finer' distinctions. The motivation behind this is to have less number of tags since less number of tags lead to efficient machine learning. Further, accuracy of manual tagging is higher when the number of tags is less. The analysis should not be so fine as to hamper machine learning and also should not be so coarse as to miss out important information. It is also felt that fine distinctions are not relevant for many of the applications (like sentence level parsing, dependency marking, etc.) for which the tagger may be used in future.

II. Syntactic function Vs lexical category

In AnnCorra, (annotation of corpus) the syntactic function of a word is not considered for POS tagging. Since the word is always tagged according to its lexical category there is consistency in tagging. This reduces confusion involved in manual tagging. Also, the machine is able to establish a word-tag relation

which leads to efficient machine learning. In short, it was decided that syntactic and semantic/pragmatic functions were not to be the basis of deciding a POS tag.

III. New tags Vs tags from a standard tagger

The Penn tags have been used as a benchmark for ILMT guidelines. Since the Penn tagset is an established tagset for English, ILMT have used the same tags as the Penn tags for common lexical types. However, new tags have been introduced wherever Penn tags have been found inadequate for Indian language descriptions.

The annotation standards for POS tagging and chunking for Indian languages include 26 tags. The tags are decided on coarse linguistic information with an idea to expand it to finer knowledge if required.

2.7.2 IL-POSTS

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. IL-POSTS (Indian language part-of-speech tagsets) is a common POS-tagset framework for Indian languages which has been designed to cover the morphosyntactic details of Indian Languages and offers advantages such as flexibility, cross-linguistic compatibility and reusability. 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 tagsets. This is detrimental to the development of language technology for Indian languages which already suffers 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, also allows for a systematic method to annotate language specific categories without disregarding the shared traits of the Indian languages. The design methodology of IL-POSTS is based on the EAGLES guidelines (Leech and Wilson 1996). IL-POSTS has a hierarchical layout of decomposable tags with three levels in the hierarchy viz., categories, types (subcategories) and attributes (features). IL-POSTS is a framework for ILs (Indian Languages) 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.

2.7.2.1 IL-POSTS Framework

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

I. Categories

II. Types

III. Attributes

I. Categories are the highest level part-of-speech classes. All categories are obligatory, that is, they 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 recognized 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 are provided for features. 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 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.

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

2.7.2.2 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.

I. Each of the categories and types is represented by a unique single letter or two-letter combination. These tags are in uppercase.

II. We also make sure that after the concatenation of a category and its type, the resultant string never exceeds three characters.

III. The values of the attributes are also assigned 1 to 4 character unique strings of letters or numbers.

These letter based tags are for the ease of humans during the 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”).

2.7.3 Other Indian Language Tagsets

In addition to the tagsets discussed above several other tagsets are also developed like Sanskrit Tagset, AU-KBC Tagset for Tamil and Urdu Tagset. These tagsets are developed taking into consideration all possible grammatical features and lexical constituents of the language in question. Sanskrit Tagset at JNU, Delhi is highly exhaustive and is classified according to morphological classification of Sanskrit words, similarly other tagsets like AU-KBC (AU-KBC NLP team 2001) and Urdu Tagset (Hardie 2004) are also highly exhaustive. However, the use of ILMT guidelines is also observed.

2.8 Conclusion

Developing a Tagset is a prerequisite for natural language processing of any language. Thus we can say prior to the development of an automatic tagger it is necessary to build suitable tagging guidelines for any language, so that tagged

corpus for any language can be built. The tagged corpus can be useful for other processes like Chunking, Morph Analyzer, Parser etc. In this way the ultimate goal of Machine translation can be fulfilled.