On the Design of a Tag Set for Dravidian Languages

Kavi Narayana Murthy and **Badugu Srinivasu**
Department of Computer and Information Sciences
University of Hyderabad
knmuh@yahoo.com, srinivasucse@gmail.com

Abstract

Tagging is the process of assigning short labels to words in a text for the purpose of indicating lexical, morphological, syntactic, semantic or other pieces of information associated with those words. When the focus is mainly on syntactic categories (and sub-categories), this is also known as part-of-speech or POS tagging. It may be noted that the term tagging is broader than the term POS tagging. Lexical, morphological and syntactic levels are well recognized in linguistics and linguistic theories normally do not posit tagging or chunking levels at all. In computational implementations within the field of Natural Language Processing (NLP), however, it has generally been found that words are highly ambiguous and ambiguities multiply at an exponential rate, making syntactic parsing so much more challenging. Tagging is a level that has been introduced before syntactic parsing with the main intention of reducing these ambiguities. It may be noted that the lexicon and morphological analyzer typically look at words in isolation and consider all possible meanings, structures, or tags. A tagger, on the other hand, looks at the sentential context and using this knowledge, attempts to reduce the possible tags for a given word in context in which it appears. It is not absolutely necessary that all ambiguities should be removed before we go to syntactic analysis, it is important only to reduce the degree of ambiguity. Syntactic parsers are anyway capable of dealing with ambiguities. If the degree of tag ambiguities is very less, tagging may not even be necessary. Tagging is useful to the extent it simplifies syntactic parsing. There does not seem to be any evidence that the human mind carries out tagging or chunking as separate processes before it embarks upon syntactic analysis.

Critical issues connected with tagging are the design of the tag set and the approach to tagging and the exact methodology used. These issues are no doubt inseparably connected with the overall purpose and the design of the lexicon, morphology and syntactic modules. The beaten path is to develop a manually tagged database of sentences and then use this for training a machine learning algorithm. The machine learning algorithm is expected to generalize from these training examples so that it can then tag any new sentence. Manual tagging is difficult, time consuming and prone to human errors. Consistency is difficult to achieve especially if the tag set is elaborate and fine grained. Also, given the limited amount of training data that is practically possible to develop, a large and detailed tag set

will lead to sparsity of training data and machine learning algorithms will fail to learn effectively. From these considerations, NLP researchers tend to restrict themselves to small, shallow or flat tag sets which are least confusing to human annotators and easy for machine learning algorithms to model. When this idea is taken to the extreme, useful sub-categorizations can be lost. In this paper we propose an alternative view and a novel approach to tagging. We focus on Dravidian languages here and demonstrate our system for Telugu and Kannada languages. We believe the ideas presented in this paper are applicable to all languages.

Before we get into designing a tagger we must ask from where we get the information required to selecting a particular tag out of all the possible tags for a given word in a given sentence. The general assumption in most tagging work has been that this information comes from other words in the sentence. This is not always true. In the case of Dravidian, for example, we find that in most cases, information required for disambiguating tags comes from word internal structure, not from the other words in the sentential context. Morphology therefore does the major part of tagging and there is no need for Markov models and things of that kind. In fact we believe that the same approach can be effectively applied to all languages provided we change our view of what constitutes a word.

The lexicon deals with words and their meanings, morphology is all about the internal structure of words and Tags are assigned to words. Given all this, what exactly is a word is a fundamental question. We believe that a lot of avoidable confusion arises both in NLP and linguistics because of a heavy emphasis on the written form of a word. A sequence of characters separated by spaces is considered to be a word. We instead define words as meaningful sequences of phonemes and we try not to get influenced by the written form. Whether and where spaces appear are irrelevant to us. When viewed from this stand point, we find that the degree of lexical ambiguity is far less than normally seen in other approaches. A vast majority of words are not ambiguous at all. Most of the remaining cases of ambiguity at root word level get resolved automatically once we consider the morphology of the inflected forms in which these words usually occur in running texts. Therefore, we believe that the main task of tagging is not one of tag assignment but only of tag disambiguation. Tags can be assigned by the dictionary and morphology components quite effectively. This way, we can develop large scale annotated corpora with high quality of tagging, without any need for manual tagging or any machine learning algorithm. We demonstrate our approach for Telugu and Kannada and argue for the merits of our approach compared to other competing approaches.

Here we do not need to do any tagging manually. There is no need for any training data and there is no need for any machine learning algorithm. A major part of tagging work is done by the lexicon and morphology, automatically. Only a small part of the data will remain ambiguous after morphology. Most of these cases are easily resolved by syntax, as they are all fully rule governed. It is thus possible to develop large scale, high quality tagged data automatically once we have a proper morphology component.

Since the work is done automatically and there is no need to worry about the effectiveness of machine learning, we can afford to have a fairly large, elaborate, fine-grained, hierarchical tag set, which captures as much of lexical, morphological, syntactic and semantic information as necessary or useful. Here we shall present such a fine-grained hierarchical tag set designed especially for Dravidian languages but believed to be more universal than that.

One aspect that is often not given sufficient importance is the precise definition of each tag. When things are left to intuition and subjective interpretation, there will naturally be confusions and inconsistencies even in the manually prepared or manually checked tagged data. Our aim shall be to define the tags as precisely as possible so that such confusions can be minimised.

We shall conclude the paper with tagging experiments and quantitative results.

1 Background and Introduction

1.1 Language, Grammar and Computation

We human beings are capable of producing a number of different kinds of sounds. We are capable of making interesting patterns by stringing together these basic sound units, called phonemes, into larger structures. And, most importantly, we are capable of systematically associating meanings with these patterns of sounds. Further, we are capable of learning such associations, we are capable of effectively communicating these association rules to others. We are also capable of communicating our ideas, thoughts, feelings and emotions to others by expressing them in patterns of sounds according to these mutually agreed upon mapping rules. This faculty of the human mind is called language. Language is the capacity to map sounds to meanings and use this for speech and thought. Language is, by far, a unique gift of nature to mankind.

Computers are capable of storing and manipulating symbol structures. They are not capable of understanding meanings. Computers do not understand the meaning of any single word in any human language. How then can we put computers to good use in meaningful processing of language? The answer to this question comes from observing the fact that there is structure in language and there is a systematic relationship between structure and meaning. This relationship between structure and meaning can be observed, learned, taught and used. Therefore, if only we take care to store only meaningful structures and if only we take care to allow only meaningful manipulations of such structures, we can ensure that everything remains meaningful (to us) throughout, although the machine itself does not understand a word. This systematic relationship between structure and meaning is what we shall call grammar. Grammar is thus the key to language processing in machines.

Grammar is the key to language processing even in humans. We do not simply store all possible linguistic units and their corresponding meanings. The number of possible sentences, for example, is infinite and we are capable of understanding the meaning of sentences we have never heard before in our life. This is possible only because we have a grammar in our head and we use this grammar to construct new sentences or to understand sentences spoken by others. This is true not only of sentences but also of all levels of linguistic analysis.

To summarize, we must develop appropriate representations of basic linguistic units, appropriate representations of their structures and appropriate formalisms for manipulating these structures at all levels of linguistic analysis. What is appropriate and what is not is dictated mainly by meaning. The written form has no role at all in this. This is a brief summary of

the theory we have been working on. See [1] for more details.

1.2 Speech and Text

Language is a powerful means of communication. Of course we can also communicate certain ideas and feelings through body language, gestures etc. Simply getting up or walking out can also convey some message to others. We can even communicate at times through silence. Nonetheless, by and large the most effective and most widely used means of communication among humans is through speech. We have therefore defined language as a mental faculty of human beings that enables us to systematically map sound patterns to meanings. Language is speech, it has nothing to do with writing. We believe that an enormous amount of confusion has been created both within NLP and in Linguistics by giving too much of importance to the written form. We all learned our first language only by listening and speaking. Reading and writing are learned later, that too only upon being taught. Literacy is not as important as people think today, one can be a highly knowledgeable scholar without being literate. Many languages of the world do not have a script of their own, the need for writing was never felt all through the history of many cultures. Writing is a technology, that has been invented by us as an after-thought, while language is a natural gift of nature to mankind. Do not confuse language for writing or script. Language can exist without a script but not vice versa. It is unfortunate, therefore, that we have started defining everything based on the written form. Words are not sequences of letters or characters separated by spaces. It does not matter if there are zero, one or more spaces within or between words. In fact inserting spaces is also a newly developed idea - stone inscriptions do not have spaces between words, for example. There are no spaces between words in speech. A word cannot be defined in terms of written symbols separated by white space. This is just not right.

A dictionary stores words and their meanings. Morphology deals with the internal structure of words. A tagger attaches tags to words. Sentences are built up from words. Words form very important and fundamental units of language. What exactly is a word then?

Look at the sentence: 'Arjuna went on dancing'. Here 'Arjuna' is the subject, the person about whom we are saying something. The predicate, that is what we are saying about the subject, is that he went on dancing. The predicate here indicates an action. A word that indicates action is called a verb. How many actions is Arjuna performing? Only one. There is only one action, therefore there is only one verb here. We classify words into word classes like nouns and verbs. Here there is one verb and hence only one word that is indicating the action. 'went on dancing' is thus one word, not three. This is the crux of our approach. We need to think in terms of meanings, not in terms of the written form.

This view is not new, linguists have always known very well that the written form is not to be taken too seriously. Despite this, we find that today even within linguistics, the written form has somehow come to have an unreasonably strong influence on our thinking and working.

This meaning based definition of a word will have far reaching implications. The so called auxiliary verbs do not indicate auxiliary activities. They do not stand for sub-actions or constituents of an action or supporting actions or incidental or related actions. Walking may imply lifting the legs one by one, moving forward, placing them back on the ground, changing the balance of weight on the two legs and so on but this is not what we mean when we talk of axillary verbs. We would therefore be compelled to reject the very idea of axillary verbs.

Taking 'has been coming' (English), or 'jaa rahaa tha' (Hindi) as single words has many merits. Features such as tense and aspect

apply to verbs, they cannot stand alone in isolation. What is the tense of 'jaa', what is the tense of 'rahaa' and what is the tense of 'tha'? Can there be several tenses for one verb? Items that indicate grammatical features are not words, they are morphemes, they form part of one word. We need to take instances such as 'has been coming' and 'jaa rahaa tha' as individual words, deal with their internal structure through morphology, assign tags to them as atomic units, look at sentences as sequences of such words. A drastic change in the way we think and work is called for. Once this happens, the differences we see between languages will melt away to a large extent and we will start seeing the underlying universals across human languages.

How far should one go along this line? 'A big tree' stands for a single object. A word that indicates an object, a thing, is a noun. Should we say 'a big tree' is therefore a single word? Of course we can. A word that is used in place of a noun is generally termed a pronoun. A pronoun stands not in place of a noun but in place of a whole noun phrase, in loose language. That is, if we wish to use the pronoun 'it' in this example, this 'it' would stand for 'a big tree', not just for 'tree'. Otherwise, we should be able to say 'a big it', right? Since pronouns stand in place of nouns, and since pronouns stand in place of items like 'a big tree', such items should be considered single words. Are we trying to say what we used to call as a phrase is now renamed as a word? What about the so called function words? Since their primary role is to indicate grammatical function rather than lexical content, should we say they are not words at all? What about items like 'in' or 'on'? Do they have any meaning at all? Don't they have some meaning? Where should we draw the dividing line for defining words and other linguistic units? How should we deal with the non-word items?

Semantics is a bit nebulous by nature and we must therefore be very careful. Since the broad goal is to discover the universal grammar that underlies all human languages, we must go by what is common across human languages. The notion of word classes leads us to good answers to all these questions we have raised here.

1.3 Words, Word Classes and Tagging

Words are minimal sequences of phonemes with meaning. Words that indicate things are called nouns. Words that indicate action or state of existence are called verbs. Nouns and verbs are examples of word classes. Nouns and verbs are found in almost all human languages, they are universal. Word classes provide broad, semantically motivated universal classification of words, and lead to further categorization at a grammatical level.

What about adjectives? Words that describe things are called adjectives. In 'a red shirt', shirt is a thing, it is a noun, the item 'red' is describing this thing, it is giving its colour, hence 'red' is an adjective. There is one view which says we can never mental visualize redness without some thing that is red. Attributes have no independent existence, they always need a substratum, some object on which they can be superimposed. As such, adjectives have no independent status at all. In contrast to this view, it is possible to argue that although properties of some object are unthinkable without that object, at a conceptual level, we can take adjectives as a legitimate word class in its own right. A word that indicates action is called a verb but the very word 'action' is a noun. Likewise, 'redness' is a noun, it is a conceptual object, but with the help of this noun, we can posit the adjective 'red'. Without getting too deep into these alternative views, let us observe that adjectives are widely recognized in human languages and it would not be wrong to consider adjectives as a class by themselves. Likewise, manner adverbs - items that indicate how an action is performed, can also be considered a universal word class. Further, pronouns which are used to avoid the boredom of repeating nouns, can also be considered a universal word class. By and large, noun, pronoun, adjective, adverb and verb are the only universal word

classes. They are semantically motivated and can be applied to all human languages.

Items that correspond to one of these word classes shall be called words, not the others. Conjunctions, articles, prepositions etc. are not universal word classes, they do not have a sufficiently clear meaning of their own. The so called function words are not words at all.

A sentence is not merely a sequence of words, a sentence may contain words, feature bundles and connectives. Items indicating feature bundles can be and should be connected up with the relevant words or other larger linguistic structures. Thus 'the' is not a word in English, it is an item which indicates a discourse level feature called definiteness of the noun phrase to which it is attached. It should be clear by now that our theory has far reaching implications at all levels of linguistics and NLP. Interested readers may see [1] for more details.

Some may argue that defining words in terms of meanings is not practicable since computers do not understand meaning. If the right thing is difficult to do that is not a valid excuse for doing the wrong thing. Do pre-processing, do post-processing, have manual intervention, do whatever you wish but do not stray away from what is right. We believe that it is practically possible to work with meaning-defined words in all languages of the world. It may prove difficult, but it should not be impossible.

Word classes such as noun, verb and adjective are also called 'Parts of Speech' (POS) by tradition. For the sake of convenience, we may use short labels, called tags, for these. For example, nouns may be indicated by N and verbs by V. POS Tagging is the process of attaching such short labels to indicate the Parts of Speech for words.

How should we deal with non-words in a sentence? Connectives and other such non-words are part of the grammar. They should ideally have no place in the lexicon or morphology but they have a definite role in syntax. We may assign tags to these non-words also if that simplifies the design of the syntactic module but we must never confuse non-words for words. Syntax is all about identifying the relations between words in a sentence, the non-words only facilitate this process.

Tags need not indicate purely syntactic categories. There is need for sub-categorization in syntax and a tagging scheme may include not only the major grammatical categories but also sub-categories. For example, one may talk of common nouns and proper nouns, of intransitive verbs and transitive verbs. One can actually go beyond purely syntactic properties and include lexical, morphological or even semantic information in the tags. It all depends upon what we need and what we can. In this paper we use the terms Tag and Tagging in this broader sense, not restricting ourselves to POS tags or POS tagging.

Tagging is only for convenience. Tagging is usually intended to reduce, if not eliminate, ambiguities at word level. It is well known that syntactic parsing is at least cubic in computational complexity and having to consider several alternative interpretations for each word can exponentially increase parsing complexity. Tagging has been invented in NLP as an independent layer of analysis, sitting between morphology and syntax, mainly to help the syntactic parser to do better in terms of speed and accuracy. However, if we take the definition of word we have given here, we will find that sentences are not as long as they appear to be (in terms of number of words) and words are not as ambiguous as they appear to be either. Therefore, syntactic parsing is actually orders of magnitude simpler than what we usually think it is. To this extent, the importance of tagging is reduced. It is worth reiterating that linguistic theories never posit tagging or chunking as separate layers of analysis sitting between morphology and syntax.

1.4 Grammar

Words are finite, the mapping from words (that is phoneme sequences) to meanings can be stored in our brain. This is the mental lexicon. But there are infinitely many possible sentences and we can understand all of them. Our mental capacity is finite and so we must necessarily be using a finite device to handle the infinitely many sentences. This mental device we have that enables us to construct and analyze infinitely many valid sentences using the finite vocabulary we have is called grammar.

Consider the sentences 'Rama saw the running deer' and 'Rama saw the deer running'. Sentences having the same set of words can thus vary in meaning and the difference can only be accounted for by the structure. Grammar is the finite device that maps an infinite variety of structures to their corresponding meanings. Grammar is the central core of the human language faculty - without grammar, language is impossible.

The lexicon maps phoneme sequences to meanings. Sequence is also one kind of structure, although a simple one. Therefore, a lexicon is also a grammar. The lexicon is not an adjunct to grammar or an independent module, it is itself very much a part of grammar.

Words of a language are finite and hence list-able. If all words of a language can be simply listed in the lexicon, there is no need for morphology. However, there is structure inside words too, there are systematic relationships between these structures and meanings and these systematic relations can be observed, learned, taught and used by the human mind. The human mind has a natural tendency to observe systematic relationships and make generalizations. Thus, morphology, which deals with the internal structure of words in relation to their meanings, also become a component of grammar. The lexicon, the morphology and the syntax constitute the three main components of grammar up to the sentence level. Grammar is not arbitrary string manipulation. Grammar must help us understand meanings in terms of structures.

1.5 Computational Grammar

A complete grammar of any given language must include the complete lexicon, the complete morphology and the complete syntax. Samples will not do, we need to be comprehensive and exhaustive. Only a complete grammar can define the language fully. The lexicon, morphology and syntax should be mutually exclusive and complementary. For example, what is handled by morphology need not be, in fact, should not be listed in the lexicon. Because of such interactions, it is not possible to have a dictionary without morphology and syntax, nor can we have a syntactic grammar without a dictionary. The dictionary, morphology and syntactic grammar always go together and must always be viewed as a whole.

We cannot open up the human brain and see what kind of grammar is sitting there but a good grammar needs to be psychologically plausible as also simple, neat and elegant. It must capture generalizations adequately and satisfactorily. It must have predictive and explanatory power. It must be universal. A child born in any language community anywhere in the world picks up its mother tongue with equal ease and in more or less the same amount of time. Hence there must be universal principles underlying and governing all human languages. The main goal of modern linguistics is to discover such a universal grammar.

The main goal of computational linguistics is also to discover such a comprehensive yet simple, neat, elegant, and universal grammar. The computer is only a powerful tool in our hand in this grand project. A computational grammar is not a different kind of grammar, it is only a comprehensive, universal, yet simple and elegant grammar, the only difference is that it has been implemented, tested and validated on real data by actually building computational

models.

A complete grammar must be capable of handling each and every valid structure and map it to appropriate meanings. How can we be sure? The only way to test and ascertain this is to test on large scale real life data. A computational grammar is simply a grammar that has actually been implemented as a computer program and subjected to extensive testing and validation on real life linguistic data. Even to build such an exhaustive grammar, we invariably need large scale data. A computational grammar is designed, developed, tested, validated on large scale real data. In order to do this, the grammar itself needs to be defined at a very minute level and in a very precise way. A whole lot of definitions and treatments you will find in grammar books are very superficial, cursory, exemplary and merely illustrative, not sufficiently detailed and precise. These ideas have perhaps never been tested and validated.

There is no evidence to show that the human mind uses different grammars for analysing utterances heard and for generating utterances. That would be wasteful and very unintelligent. Therefore, grammars must be designed in such a way that they can used both for analysis and generation with equal ease.

There is also another common confusion about generation. Generation does not mean generating each and every possible form. That would be a completely unnatural process. Human beings never generate all possible forms of a word or all possible sentences in all their lives. They never even try. Asking a computer to generate all forms is therefore not right, except purely as a means of testing and validating a computational grammar. What linguists really mean is that a grammar should be capable of generating all valid forms, this is only an abstract specification, not a practical requirement. Even here, we must realize that human beings are also capable of generating invalid forms, if asked to. Therefore, if you ask a computer to generate an invalid or meaningless form from a given grammar and it does, there is no fault. Computational grammars must be designed to correctly generate and analyze all valid forms but it is not essential to ensure that such a system will never generate invalid forms even if asked to. Over-generation is perfectly fine, both theoretically and practically.

We will either need to generate particular forms based purely on given grammatical properties or we will need to generate linguistic utterances to convey an intended meaning. Computers do not understand meanings and so the second case never really arises. However, we may be working with some symbolic representation of meanings, in which case a computer program may be expected to generate valid sentences to convey the intended meanings. Generating sentences as part of automatic translation is a much more common requirement.

Linguists often fail to understand the importance of the size of data used for building and testing grammatical systems. They think there are only a few types of structures and there is no point in looking at a thousand examples of each kind. There is more to it than meets the eye. There are a large number of significant linguistic phenomena and not all of them occur equally frequently. Those who are used to looking at only a few important phenomena will not understand the importance of large scale data. Rare phenomena are more likely to occur in a large corpus than in a small corpus. Therefore, when the aim to develop wide coverage, comprehensive, if not exhaustive grammars, the importance of a large and representative corpus cannot be undermined.

We believe that we can go much closer to the dream of a universal grammar if we take the definition of word we have given here seriously. Meaning is central to language and linguistics and any process that causes loss or distortion of meaning is simply not acceptable.

2 A Novel Approach to Tagging

There is only one critical question that we need to ask when it comes to tagging - where can we find the crucial bits of information required to assign the correct tag to a given word in a given sentence? Statistical approaches assume that the necessary information comes from the other words in the sentence. In many cases, only the words that come before the current word are taken into direct consideration. We believe, in sharp contrast, that the crucial information required for assigning the correct tag comes from within the word. It is the internal structure of a word that determines its grammatical category as also sub-categorization and other features. True, there will be instances where the internal structure alone is not sufficient. Firstly we find that such cases are not as frequent as you may be thinking. A vast majority of the words can be tagged correctly by looking at the internal structure of the word. The crux of tagging lies in morphology. This is clearly true in the case of so called morphologically rich languages but this is actually true of all human languages if only we define words in terms of meanings rather than in terms of the written form and spacing considerations. Secondly, in those cases where morphology assigns more than one possible tag, information required for disambiguation comes mainly from syntax. Syntax implies complex inter-relationships between words and this cannot be reduced to a mere statistics of sequences. In Kannada, for example, the plural/honorific imperative form of a verb and a past conjunctive verbal participle form are the same. Hence morphology cannot resolve this ambiguity. This ambiguity can only be resolved by looking at the sentence structure. If this word is functioning as a finite verb, it must be the imperative. If it is followed by another finite verb later in the sentence, this could a conjunctive participle. Statistical techniques are perhaps not the best means to capture and utilize such complex functional dependencies. Instead, syntactic parsing will automatically remove most of the tag ambiguities. Given this observation, we use a simple pipeline architecture as depicted in the figure below. We keep going forward and we do not need to come back again and again to preceding modules. We carry with us all the necessary/useful information in the form of tags, each module adding or refining the information as we move on. The lexicon assigns tags to words that appear without any overt morphological inflection. Morphology handles all the derived and inflected words, including many forms of sandhi. The bridge module combines the tags given by the dictionary and the additional information given by the morph, making suitable changes to reflect the correct structure and meaning. The chunker takes these tag sequences to produce chunks. A chunker may or may not be required, it is included in the architecture for generality. In fact for our work on Dravidian, we do not need a chunker at all. The parser analyzes these chunk sequences and produces a dependency structure. The overall tag structure remains the same throughout, making it so much simpler and easier to build, test and use.

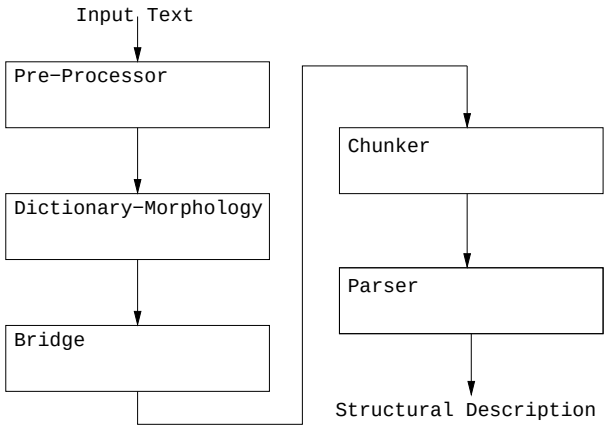

FIG 1 The System Architecture

In this paper we shall focus on the design of a tag set given the above theory and background. Details of the lexicon, morphology, syntax etc. are being published elsewhere. We shall include here transcripts from the actual implemented system to give a feel for the readers.

2.1 Designing a Tag Set

Tagging approaches based on machine learning require manually tagged training data. Manual tagging becomes difficult and error prone as

the tag set becomes large and elaborate and so there is a strong tendency to go for small, flat tag sets. Such tag sets may not capture all the required and/or useful bits of information for carrying out various tasks in NLP. Flat tag sets are also rigid and resist changes. Hierarchical tag sets are more flexible. In our case, we do not depend upon statistical or machine learning techniques and we do not need any training data. No manual tagging work is involved and so we can afford to have large, fine-grained, elaborate, hierarchical tag set that carries as much of lexical, morphological, syntactic and semantic fields as we wish.

One of the biggest difficulties that researchers face while designing tag sets, while performing manual tagging and while building and evaluating tagging systems is the very definition of tags. Tags involve lexical, morphological, syntactic and semantic considerations and often there are conflicts. One cannot go purely by intuitive definitions such as 'nouns are things, pronouns stand in place of nouns and adjectives modify nouns'. We will need to give precise definitions and criteria to decide which tag label should be given to which word. We shall illustrate the kind of reasoning we need to go through with just a few examples below. More will come later.

Let us take a clear case. Since both nouns and pronouns can take nominal inflections and both can function as subjects and objects of sentences, why should we even make a distinction between them? After all, there is no real difference in morphology either. How do we answer such a question? Upon careful examination, we find that pronouns are neither modified by adjectives nor do they function like adjectives, modifying other nouns. Nouns can. A noun can be modified by a demonstrative adjective, a pronoun cannot be. Nouns can be modified by quantifying adjectives, pronouns cannot be. As such, there are significant differences at the level of syntax and that is why we must distinguish between these two categories. Similarly, common nouns differ from proper nouns in significant ways. Demonstrative adjectives, quantifying adjectives, ordinals, and descriptive adjectives need to be treated differently since they have different roles in chunking. Not making such distinctions will lead to unnecessary explosion of possible parses, making the parsing process slow and less accurate. Even within proper names, names of persons, places etc. vary in their grammatical properties. Place names appear more in nominative, dative and locative cases, and not so much in accusative case. A place name can modify a person name but not vice versa. Place names can modify common nouns, person names usually will not. Grammar includes hard constraints as also softer restrictions. We need to design a tag set keeping all the relevant linguistic phenomena in mind. It is not possible to have a general purpose tag set suitable for all applications. Here our goal is syntactic parsing and this goal guides us at every step.

A tag set is a classification system. the moment we introduce sub-categories, it becomes a hierarchical classification system. There are certain general principles that must be followed in the design of any hierarchical classification system. If set X is divided into sets A, B and C, a) all the properties of X must be valid for A, B and C b) A, B and C must be mutually exclusive c) X must not include anything other than A, B and C. Further, we believe that the major categories must correspond one-to-one with the universal word classes. We find that such basic requirements are often violated in the tag sets being proposed by other groups and we shall demonstrate this after we define our own tag set.

Keeping these scientific considerations in mind, we have developed a fairly elaborate hierarchical tag set starting from Kannada and Telugu data. Tagging is an intermediate level and appropriateness of tag set and tag assignments can only be verified by placing it in between morphology and syntax and building and testing actual systems. We are in the process of developing an end-to-end system and this gives us strong basis to argue why our design is superior and why exactly other possible alternatives

are not good enough.

2.2 The Tag Set

We have seen that noun, pronoun, verb, adjective, adverb are the major and universal word classes or categories. We shall denote these with the widely used and highly readable mnemonics N, PRO, V, ADJ and ADV respectively. We shall try to define these major categories and their sub-categories. Sub-categorization will naturally be a bit more specific to Dravidian languages. In each case, we shall start with a broad semantic description and then look at the specific morphological and syntactic properties.

2.2.1 Noun

Words that are used to name a person, animal, place or a thing, including abstract concepts or ideas, are called nouns. This is the layman's definition found in elementary grammar books, it is not precise enough. We need to look at the morphological and syntactic properties of nouns.

Morphologically, nouns can take number and case. Most commonly, number can be singular or plural. In Sanskrit, there is also a dual number. Usually, the singular form is the default. Not all nouns show number distinction. Only countable nouns take plural forms. Uncountable nouns, including mass nouns (water, for example) and abstract nouns (beauty, for example) usually do not take plural. Some nouns may not appear in plural form simply because situations where we can put several of them will be rare. We can talk of eyes or ears but rarely do we need to talk of noses. Some nouns are actually plural but often construed as singular and vice versa. Some appear only in plural form. Pronouns also take number but other categories do not. Given all this, if a word takes plural form, we can say it is a noun or a pronoun but we cannot say a word is not a noun just because it does not show number feature. The number feature is often helpful in clearing confusion between nouns and adjectives. In many languages, verbs also show singular-plural distinction. This is not number, a plural marker on a verb never indicates plurality of actions. Verbs may be marked for number purely for showing agreement, say, with the subject of the sentence.

At a broad level, we may look at case as simply direct or oblique. At another level, we may consider cases as indicators of the thematic roles played by various constituents in a sentence in relation to the verb. These indicators may manifest in various forms. In some languages, they appear as overt affixes called case markers. In other languages, these case indicators may take the form of prepositions or post-positions. In some languages, case distinctions may be expressed indirectly by the position of the word in the sentence. In any case, it is important to realize that prepositions and postpositions are not words, they are only features of other words. Sometimes it is argued that prepositions also show relation between two nouns. For example, in the phrase 'the book on the table', the preposition 'on' is said to relate the tow nouns 'book' and 'table'. This analysis is not entirely right. The phrase 'the book on the table' actually means 'the book which is/was on the table' and the preposition 'on' indicates the place where the book exists, existence itself being indicated by the verb 'is/was'. Upon sufficiently deep and detailed analysis, we can assure ourselves that prepositions and post-positions are very much like case markers, with the same primary purpose. If a word shows case distinctions, it can only be a noun or a pronoun, not any other category. From a morphological point of view, words which are usually considered adverbs of place or time may have to considered as nouns if they take number and/or case distinctions. Words like 'here' and 'there' in English may be treated as adverbs but their counterparts in Dravidian ('illi', 'alli' for example in Kannada) will show case distinctions. Only a few case distinctions are used simply because these are the only meaningful possibilities. Thus, 'illi' (here) is already a place indicator and we cannot think of adding a locative case marker

to this word. We can think of 'from here', or 'to here', but not ' in here'. Given this, one alternative is to consider them as nouns and depend upon the standard noun morphology to take care of all the case distinctions that occur, without worrying about over-generation. An alternative view is to consider each of these few forms as separate words in their own right, call them adverbs and take them out of the realm of morphology. Generalizations are lost in this latter view and so we take the first view here.

In terms of syntax, nouns can take up major thematic roles including subject and object roles in a sentence. Pronouns can also do but adjectives and adverbs cannot. At a more local level, nouns can be modified by various kinds of nominal modifiers including demonstratives, quantifiers and other kinds of adjectives, including participles. Nouns can also function as nominal modifiers and modify other nouns. This is a general feature of human languages whether the resulting noun-noun structure is a compound or a phrase, permitting nouns to be used as adjectives in syntax eliminates the need to distinguish between the nominal and adjectival use. There will no longer be any need to distinguish between lexical and syntactic categories as far as this particular case is concerned, tagging will become so much more easier and less confusing and less ambiguous.

It must be noted that possessive forms of nouns and pronouns always function as adjectives. From a purely morphological point of view, we usually treat them under noun or pronoun categories. Once this point is clearly understood, there should not be any more confusions. Whenever we talk of nouns or pronouns at a syntactic level, we exclude possessive forms. A possessive noun or a pronoun can modify other nouns but can never take thematic roles such as subject or object.

In fact there are many situations where the morphological and syntactic view points differ. We need to have a specified policy to deal with all such situations. In our work, we always favour the morphological view point in deciding tags. The flow of information in our architecture is from lexical and morphological levels to syntactic level and tag assignments are first done at the morphological level. If and where needed, suitable refinements can be done at later levels of analysis. On the other hand, if we get stuck up right in the initial levels, we cannot progress at all.

At a discourse level, nouns can be referred to and they can also refer. Pronouns can refer to nouns or noun phrases. We can say 'the big tree' and then refer to it using the pronoun 'it'. Definite noun phrases can also refer to other nouns or noun phrases. 'The cap' can refer to a part of 'the pen'. Here by reference we mean non-syntactic relations at inter and intra-sentential levels, not exactly what linguists do in binding theory. Here the important questions are what refers to what and what is the nature of the semantic relation between the two. reference sets nouns and pronouns apart and sets all other categories aside too.

Subcategories of Noun

Nouns are subclassified into common nouns, proper nouns, locative nouns and cardinals, denoted by the tags NCOM, NPRP, NLOC and NCARD. These distinctions are required mainly because these subcategories occur in differing positions inside noun phrases and restrict the occurrence of other constituents within noun phrases. Proper nouns, locative nouns and cardinals have special properties and all other nouns are grouped under the default subcategory called common noun.

Proper nouns usually do not take plural forms, have a somewhat different statistical distribution of case marker occurrences, are not modified by various kinds of modifiers (including demonstratives, quantifying and qualifying adjectives, possessive nouns and pronouns, relative participles etc.), and they act as modifiers under restricted situations.

Proper nouns are also not found in the lexicon unless the words have some other common noun meanings also. Proper nouns can be further subclassified into person, Location, organization and others, denoted by PER, LOC, ORG and OTH respectively. There are significant differences amongst these. Agreement features vary. For example, N-PRP-PER words have implicit masculine or feminine gender, which agrees with the verb. Locations can be part of other proper names including person names and names of organizations. Person names can be part of an organization name but they rarely modify location names. Location names occur more frequently in nominative, dative, genitive and locative cases and relatively less in, say, accusative case. Proper nouns cannot be relativized, common nouns can be - 'the boy I know' is OK but 'the Rama I know' is odd. Proper nouns usually signify specific, single objects and so annotators tend to get confused whenever they come across fairly specific objects which may not occur in plural form frequently. Criteria such as the above should help to resolve the common noun - proper noun confusions.

Nouns indicating space or time, also called spatio-temporal nouns, are subcategorized under nouns because they show nominal morphology, although they are usually adverbial in function. They can also function like nouns and become subjects or objects although rarely. Here again we find restrictions on case, modification, etc.

Cardinals behave like nouns in terms of morphology and are therefore grouped under nouns. Their morphology is, however, a bit irregular. They usually act as nominal modifiers but can also act as nouns and take on subject and object roles. This happens when we are talking about these numbers themselves, as in mathematics. Singularity and plurality are implicit but overt case marking is seen.

Proto-Nouns

There are certain words whose classification requires a more careful look. Consider the Kannada word 'obba' meaning one person. By default the gender is masculine. One may argue that 'obba' is a pronoun because it stands for some one person. Somebody else may argue that it should be treated as an adjective since it indicates number, (apart from indicating that the following noun should be human) as in 'obba huDuga' (cf. 'oMdu mara'). A third person may counter this by saying it cannot be an adjective since it can take nominal inflections such as number and case. The word 'obba' takes nominal inflections and can be subject of a sentence and so it can only be a noun or a pronoun. But pronouns cannot modify nouns, they can only stand in place of a noun, and so this word can only be taken as a noun. But 'obba' indicates some man in general and not any specific person. How do we resolve this?

We need to look at related word forms such as 'obbanu' (one man), 'obbaLu' (one woman) and obbaru (one man/woman, plural indicating honorificity). Effecting a change in number is very much a standard aspect of morphology but change of gender cannot be considered a grammatical process. How can grammar change gender, that too in languages where the grammatical and biological genders are closely related? Therefore, although 'obbaru' can be obtained from 'obbanu' or 'obbaLu', 'obbanu' cannot be obtained from 'obbaLu' or vice versa. We will have to consider 'obbanu' and 'obbaLu' as separate words, related in meaning though they are. More importantly, we need to realize that 'obbanu', 'obbaLu' and 'obbaru' are clearly pronouns. They have all the properties of pronouns, they can be subjects of sentences, for example. In contrast, 'obba' is adjectival in nature, it is not a noun or a pronoun. The source of the confusion is the fact that nominative suffixes can be optionally dropped in Kannada and so we confuse 'obba' to be the same as 'obbanu'. 'obba baMda' is the same as 'obbanu baMdanu' (one man came) but in 'obba huDuga' we cannot replace 'obba' with 'obbanu'. Therefore, 'obba' should

be considered as an adjective indicating the modified noun to be one person, whereas the other forms are clearly pronouns.

In fact a large number of adjectives can become pronominals by a similar change in morphological structure. Thus the adjective 'praamaaNika' can become 'praamaaNikanu', 'praamaaNikaLu', 'praamaaNikaru' and then take other cases thereafter. Here the addition of suitable nominative suffixes is deriving pronouns from adjectives by systematic and highly productive processes. Instead of storing all of 'praamaaNika', 'praamaaNikanu' and 'praamaaNikaLu' directly in the lexicon, we can store only 'praamaaNika' and mark it as an adjective that can become a pronoun by the addition of suitable nominative case suffixes. Thus words like 'praamaaNika' behave like proto-nouns, they are actually adjectives but ready to become nouns/pronouns by the mere addition of gender information. We have chosen to call these words proto-nouns and indicated this with the PTN label.

Words such as 'kelavaru', 'mattobbaru', innobbaru' are best treated as pronouns, they have all the properties of pronouns. They indicate an indefinite number of persons. This is better than treating these words as cardinal nouns, subclassified as human, treating other non-human cardinals as a separate subclass of cardinal nouns. They can replace noun phrases, they indicate generic groups and are thus pronouns, not nouns.

2.2.2 Pronoun

Pronouns are abbreviated forms of nouns or noun phrases, abbreviated in the sense of information they carry. The proper noun 'Ram' stands for a particular person and the pronoun 'he' stands for any one masculine person, other than you and I. It has most of the information that the word 'Ram' carries but not quite every bit. We use pronouns to avoid repeating the same nouns time and again and the partial information that the pronouns carry are sufficient for us to understand the discourse properly.

Pronouns can act as subjects and objects. They also take number and case like nouns. However, pronoun morphology is somewhat irregular in many languages. Also, pronouns show overt marking for gender and person, which are usually implicit in nouns. Pronouns may also show other finer distinctions such as proximate and distal, inclusive or exclusive, and so on. Pronouns are usually not modified nor do they act as modifiers. Pronouns are usually not relativized although rare usages such as 'he who would climb the ladder must begin at the bottom' are found. Reference is a very important property of pronouns.

Pronouns are subcategorized into personal pronouns, interrogative pronouns, reflexive pronouns and indefinite pronouns, labelled PER, INTG, REF and INDF respectively. Pronouns other than personal pronouns are usually in third person only. These subcategories are motivated not by morphology but by the various kinds of syntactic constructions in which these pronouns participate.

2.2.3 Verb

Verbs are words that indicate actions or events. Existence or state of existence is considered to be a very special kind of event. Existence is the most fundamental event, without which we cannot even talk of other kinds of events or actions. Therefore, words that indicate existence or state of existence are also called verbs.

Morphologically, verbs may show tense, aspect and mood. they may also show a variety of agreement markers. Verb morphology can be extremely rich and complex. Indeed it is in Dravidian. Some verbs are defective, they show little, no or morphology, and are depicted as such in the tag set.

Syntactically, a verb has expectations about thematic roles such as subject and object which are to be filled by noun phrases or clauses as appropriate. A verb, along with its complements

and optional adjuncts, constitutes a clause. there can be one or more clauses in a sentence and if there are two or more clauses, the verbs in various clauses are inter-related. Syntax is all about finding the inter-relationships between words in a sentence and the verb plays a central role in this whole process.

Verbs have been classified along various dimensions but from the point of view of syntax, the sub-categorization frames form the most important characteristic. Sub-categorization frames specify what roles are essential, optional and prohibited and what syntactic kinds of constituents can fill up these roles. As such these are extensions of the traditional notion of transitivity. Here we classify verbs into transitive, intransitive and bitransitive (denoted TR, IN and BI) and mark further restrictions by using numerical indices as in TR1 or TR12 or TR13.

2.2.4 Adjective

Adjectives are words that modify nouns by specifying their attributes or properties. Adjectives can be modified only by intensifiers. As a general rules adjectives do not show any inflectional morphology. Inflections if any are only agreement markers. Thus, if any word takes a plural form or case markers, it is not an adjective. Adjectives used attributively are part of noun phrases occurring in modifier positions. When used predicatively, we may consider adjectives as heads of adjectival phrases. Possessive nouns and pronouns, relative participles, cardinals etc. are all adjectival in function although they may be classified elsewhere from the point of view of morphology.

Adjectives (other than those that have already been considered elsewhere) are subcategorized into demonstratives, ordinals, quantifiers, question words, and all other by default as absolute, labelled DEM, ORD, QNTF, QW and ABS. Demonstrative adjectives occur zero or one times. Question words occur zero or one times and question the noun being modified. Ordinals and Quantifiers are mutually exclusive. Absolute adjectives may occur zero, one or more times. There are also positional restrictions. In head final languages, demonstratives or question words occur first, then come ordinals or quantifiers, after which appear absolute adjectives. Subcategorization of adjectives is justified by these observations.

2.2.5 Adverb

Words that modify a verb are called adverbs. However, we find a number of other kinds of words, which modify or add more information to adjectives, other adverbs, or a whole sentence and all these words are traditionally called adverbs. They usually answer questions such as where, when, how or how much. Such words may provide more information in terms of time, frequency, manner etc. Adverbs appear in various positions in a sentence and add quite a bit of complexity in syntactic and semantic analysis. Adverbs usually have no morphology. Adverbs form the default category, if a word does not fit anywhere else, there is a tendency to push into the adverb basket.

In keeping with the tradition, we include all these various types of words under the adverb category and subcategorize adverbs into adverbs of manner (MAN), place (PLA), time (TIM), question words (QW), intensifiers (INTF), negation (NEG), conjunctions (CONJ), post-nominal modifiers (POSN), and by default, all the others into absolute (ABS). Intensifiers modify adjectives. Adverbs of negation add a negative aspect to verbs. Post-nominal modifiers add clitic-like information to nouns. Certain conjunctions occur in the sentence initial position, indicating conjunction with preceding sentences at discourse level, and have no role within the sentence as far as syntax is concerned. Although they are conjunctions, they do not join any two items within the sentence. We have chosen to treat them under adverbs keeping syntax in mind. Note that time and place indicators are classified as locative nouns if they show any nominal inflections. When no inflections occur, they have been grouped under adverbs.

2.2.6 Other Tags

Tag sets proposed by various groups include tags for interjections, post-positions, particles, punctuation marks, symbols, foreign words, echo formations, reduplication, etc. and perhaps one unknown tag to take care of situations when no tag seems to fit. Since these do not correspond to universal word classes and since these tags do not map on to words which are defined based on a specific theory of meaning, we shall avoid getting into all these here. How about function words like conjunctions? Such items do have a very definite role in syntax. They cannot be ignored or wished away. At the same time, we should not group them with words of the language. These items have less of a lexical role and more of a grammatical role and so they should be excluded from the lexicon, morphology and tagging and included within the grammar at the syntactic level. Having said this much, it may still be felt practically convenient to assign tags to them. This way, syntax would see a sequence of tags as input, not a mixture of tags and other meta symbols. Sometimes such items may even be listed within the lexicon but we must understand that this is purely for the sake of practical convenience.

2.3 Our Tag Set

Firstly, we believe the same overall structure can be and should be retained at all levels of processing, starting from the lexicon, through morphology, tagging and syntactic parsing. Each module may add or refine the relevant parts but the overall tag structure should not change.

Secondly, although every word must ideally be passed through morphology, we can avoid some work and save time by not passing word forms that are directly found in the lexicon through the morphological analyzer. In that case, the lexicon should give the same tag that the morphology would have given. Defaults such as singular number and nominative case for nouns should therefore be shown in the lexicon itself. The tags assigned by the dictionary and morphology should be directly suited for syntactic analysis, without going back to any other module. Note that syntax works only with tags, not with the words themselves. The whole idea of defining word classes and the tagging scheme is to reduce the set of word forms into a set of tags. As far as syntax is concerned, a sentence is simply a sequence of tags. All the lexical, morphological, syntactic and semantic pieces of information necessary or useful for syntactic parsing should therefore be directly reflected in the tag structure.

The lexicon is truely a list of exceptions and all idiosyncratic word forms must be stored directly in the lexicon. Thus idiosyncratic cases of word forms including plurals, various cases, clitics etc. will be listed in the dictionary with appropiate tags.

Certain categories such as pronouns and cardinals show partly irregular morphology. One extreme solution would be to store all forms of these words directly in the lexicon, avoiding morphology altogether. Another extreme would be to somehow try and handle all the irregularities within morphology. A good intermediate solution would be to store only the irregular forms in the lexicon and let morphology handle all the regular forms. In Kannada, for example, nominative, dative and genitive forms of pronouns are stored in the lexicon and other forms are derived by the morphology starting from the genitive form. For example, 'nannalli' can be obtained from 'nanna'. This would require suitable adjustments in tags after morphology.

The list of tags used in the Kannada lexicon is given below. All the major categories and sub-categories are included but only samples of feature combinations are shown here. For pronouns the exhaustive set is included to help the reader get a feel for the complete tag set. Each item here is a tag. It has a hierarchical structure shown via dashes. The parts separated by dashes are called tag elements. Tag elements may in turn be made up of tag atoms. Thus, 'PROREFP23ṀFNṖLNOM' is one single

tag, PRO, REF, etc. are the tag elements, and P23ṀFNṖL is a single tag element containing the person, number and gender information. P23 is a short hand for second or third person and MFN is a short hand for Masculine or feminine or Neuter. Short hands such as these encapsulate tag ambiguities at the level of features and reduce the overt ambiguity in tagging. The labels used here are generally quite well known and need no explanation.

1. ADJ-ABS
2. ADJ-DEM
3. ADJ-ORD
4. ADJ-QNTF
5. ADJ-QW
6. ADV-ABS
7. ADV-CONJ
8. ADV-INTF
9. ADV-MAN
10. ADV-PLA
11. ADV-POSN
12. ADV-QW
13. ADV-TIM
14. N-CARD-N.SL-NOM
15. N-COM-COU-M.SL-NOM
16. N-COM-UNC-N.SL-NOM
17. N-LOC-TIM-COU-ABS-N.SL-NOM
18. N-LOC-TIM-UNC-ABS-N.SL-NOM
19. N-LOC-TIM-UNC-DIST-N.SL-NOM
20. N-LOC-TIM-UNC-PROX-N.SL-NOM
21. N-LOC-TIM-UNC-QW-N.SL-NOM
22. N-LOC-PLA-COU-ABS-N.SL-NOM
23. N-PRP-LOC
24. N-PRP-ORG
25. N-PRP-OTH
26. N-PRP-PER-M.SL-NOM
27. PRO-INTG-P3.F.SL-NOM
28. PRO-INTG-P3.MF.PL-NOM
29. PRO-INTG-P3.M.SL-NOM
30. PRO-INTG-P3.N.PL-NOM
31. PRO-INTG-P3.N.SL-COMP
32. PRO-INTG-P3.N.SL-DAT
33. PRO-INTG-P3.N.SL-GEN
34. PRO-INTG-P3.N.SL-NOM
35. PRO-INTG-P3.N.SL-PURP1
36. PRO-INTG-P3.N.SL-PURP2
37. PRO-PER-P1.MFN.PL-ABS-COMP
38. PRO-PER-P1.MFN.PL-ABS-DAT
39. PRO-PER-P1.MFN.PL-ABS-GEN
40. PRO-PER-P1.MFN.PL-ABS-NOM
41. PRO-PER-P1.MFN.PL-ABS-PURP1
42. PRO-PER-P1.MFN.PL-ABS-PURP2
43. PRO-PER-P1.MFN.SL-ABS-COMP
44. PRO-PER-P1.MFN.SL-ABS-DAT
45. PRO-PER-P1.MFN.SL-ABS-GEN
46. PRO-PER-P1.MFN.SL-ABS-NOM
47. PRO-PER-P1.MFN.SL-ABS-PURP1
48. PRO-PER-P1.MFN.SL-ABS-PURP2
49. PRO-PER-P2.MFN.PL-ABS-COMP
50. PRO-PER-P2.MFN.PL-ABS-DAT
51. PRO-PER-P2.MFN.PL-ABS-GEN

52. PRO-PER-P2.MFN.PL-ABS-NOM
53. PRO-PER-P2.MFN.PL-ABS-PURP1
54. PRO-PER-P2.MFN.PL-ABS-PURP2
55. PRO-PER-P2.MFN.SL-ABS-COMP
56. PRO-PER-P2.MFN.SL-ABS-DAT
57. PRO-PER-P2.MFN.SL-ABS-GEN
58. PRO-PER-P2.MFN.SL-ABS-NOM
59. PRO-PER-P2.MFN.SL-ABS-PURP1
60. PRO-PER-P2.MFN.SL-ABS-PURP2
61. PRO-PER-P3.F.SL-DIST-NOM
62. PRO-PER-P3.F.SL-PROX-NOM
63. PRO-PER-P3.MF.PL-DIST-NOM
64. PRO-PER-P3.MF.PL-PROX-NOM
65. PRO-PER-P3.M.SL-DIST-NOM
66. PRO-PER-P3.M.SL-PROX-NOM
67. PRO-PER-P3.N.PL-DIST-COMP
68. PRO-PER-P3.N.PL-DIST-DAT
69. PRO-PER-P3.N.PL-DIST-NOM
70. PRO-PER-P3.N.PL-DIST-PURP1
71. PRO-PER-P3.N.PL-DIST-PURP2
72. PRO-PER-P3.N.PL-PROX-COMP
73. PRO-PER-P3.N.PL-PROX-DAT
74. PRO-PER-P3.N.PL-PROX-NOM
75. PRO-PER-P3.N.PL-PROX-PURP1
76. PRO-PER-P3.N.PL-PROX-PURP2
77. PRO-PER-P3.N.SL-DIST-COMP
78. PRO-PER-P3.N.SL-DIST-DAT
79. PRO-PER-P3.N.SL-DIST-GEN
80. PRO-PER-P3.N.SL-DIST-NOM
81. PRO-PER-P3.N.SL-DIST-PURP1
82. PRO-PER-P3.N.SL-DIST-PURP2
83. PRO-PER-P3.N.SL-PROX-COMP
84. PRO-PER-P3.N.SL-PROX-DAT
85. PRO-PER-P3.N.SL-PROX-GEN
86. PRO-PER-P3.N.SL-PROX-NOM
87. PRO-PER-P3.N.SL-PROX-PURP1
88. PRO-PER-P3.N.SL-PROX-PURP2
89. PRO-INDF-P1.MF.PL-NOM
90. PRO-INDF-P3.F.SL-NOM
91. PRO-INDF-P3.MF.PL-NOM
92. PRO-INDF-P3.M.SL-NOM
93. PRO-INDF-P3.N.PL-NOM
94. PRO-INDF-P3.N.SL-DAT
95. PRO-INDF-P3.N.SL-NOM
96. PRO-REF-P23.MFN.PL-COMP
97. PRO-REF-P23.MFN.PL-DAT
98. PRO-REF-P23.MFN.PL-GEN
99. PRO-REF-P23.MFN.PL-NOM
100. PRO-REF-P23.MFN.PL-PURP1
101. PRO-REF-P23.MFN.PL-PURP2
102. PRO-REF-P3.MFN.PL-GEN
103. PRO-REF-P3.MFN.SL-COMP
104. PRO-REF-P3.MFN.SL-DAT
105. PRO-REF-P3.MFN.SL-GEN
106. PRO-REF-P3.MFN.SL-NOM
107. PRO-REF-P3.MFN.SL-PURP1
108. PRO-REF-P3.MFN.SL-PURP2
109. PTN

110. V-BI1

111. V-DEFE

112. V-IN

113. V-TR1

Tags shown in the lexicon get enriched with the information coming from morphology. Look at the examples of morph output below:

$annavannu = anna(N - COM - UNC - N.SL - NOM) + epsilon(Singular) + annu(Accusative)$
$koTTanu = koDu : (V - BI1) + epsilon + id(Past - Tense) + anu(Third - Person.Masculine.Singular)$

The theory and implementation details of the morphology component is being published elsewhere. The tagger combines the elements obtained from the lexicon and morph and assigns the following tags:
$annavannu||anna||N - COM - UNC - N.SL - ACC$

$koTTanu||koDu||V - BI1 - ABS - PAST - P3.M.SL$

More than 10,000 word forms can be generated and analyzed for a given nominal stem including all the number, case, clitic and vocative combinations. A mind boggling number of word forms can be analyzed and generated for verbs. The complete set of tags is therefore extremely large. This does not cause any problems since in our approach tagging is done automatically and there is neither any need for manual tagging nor for any machine learning algorithm. The current system is more or less complete with respect to inflection, derivation and to some extent external sandhi.

2.3.1 Derivation

Derivational morphology can introduce changes in grammatical category. A gerund, for example, is a noun derived from a verb. Derived words can retain some properties of their initial category and obtain new properties in their new category. Both are important and classifying them either under the initial category or under the final category would be incorrect. A gerund, for example, may retain its transitivity property and ability to take an object, while at the same time take nominal suffixes and behave like a subject or object of another verb. Look at this example: $heccuvudariMda||heccu||V - TR1 > v - ABS - FUT - GRND- > n - ABL||$ In our tagging system, we retain the complete history and all the relevant properties at each stage. This is essential for syntax.

2.3.2 Comparison with Other Tag Sets

Instead of making an exhaustive comparison with other tag sets, we shall only highlight a few important aspects of divergence. There have been several Initial efforts to develop tag sets by various groups in India over the last few years. Most of these have been motivated by, if not based on, the tag sets and tagging approaches followed by others in the world, especially for English. These tag sets tend to be coarse, quite shallow if not completely flat, intended for manual tagging and machine learning, under the assumption that tag assignment and disambiguation are largely a matter of arrangement of tokens in a given sentence. Morphology does not have a central role. Most importantly, all these tagging schemes are orthography oriented and consider words to be sequence of written symbols separated by white spaces. Neither meanings nor pronunciation are taken as the basis for defining words. The specific purpose for which these tag sets have been designed is not clear. No attempt seems to have been made to precisely define each of the tags and publish the same. Presumably, all the efforts so far have been driven by the needs of machine translation. Over time, all these various proposed schemes are converging into a proposed draft BIS standard tag set. We shall therefore take up some issues with this proposed draft BIS tag set for comparison.

The proposed draft BIS tag sets show division

of verbs into main and auxiliary verbs, which we have already questioned. Also, there are attempts to divide verbs into finite and non-finite. What is the basis for finiteness? Is this division based on tense? Is it based on agreement marking? What exactly does this mean to syntax? Precise definitions have not been provided, infinitive is sometimes listed separately, and the fact that it is possible to derive finite verb forms from non-finite forms through legitimate processes of morphology seems to have been overlooked. Gerunds and the so-called verbal nouns add to the confusion relating to verbs. It is not clear why demonstratives and quantifiers have been considered as top level categories. The notion of a post-position as a grammatical category is questionable.

2.4 The Bridge Module and Tag Disambiguation

Ideally, the input to a grammatical system should be a normalized sentence, wherein orthographic tokens have already been preprocessed into meaningful words. Ideally, the morphological analyzer should depict the complete and correct structure of the given words, helping the readers to obtain the correct meanings. Due to a variety of design and implementation strategies followed while building a computational system, there could be some deviations. The purpose of the bridge module is to iron out all such deviations so that the subsequent syntactic module gets to see a completely normalized and tagged sentence.

Here in the bridge module, the bits of information obtained from the dictionary and morphology are combined to generate final tags. For example,

```
manege||mane||N-COM-COU-N.SL-DAT

maaDuttaane||maaDu||V-TR1-PRES-P3.M.SL

maaDidare||maaDu||V-TR1-PAST-COND

maaDabeekaagibaMdaaga||maaDu||
V-TR1-INF-CMPL-AUX.aagu-CJP.PAST
-AUX.baru-PAST-RP-adj-CLIT.aaga

maaDibiTTaraMtaa||maaDu||V-TR1
-CJP.PAST-AUX.biDu-PAST-P3.MF.PL
-CLIT.aMte-CLIT.INTG
```

It may be noted that noun verb ambiguities have been resolved by morphology. Note that Dravidian morphology is very rich and standard linguistic terms may not always be readily available. We have chosen to tag such cases with the morphemes themselves. Further, certain morphemes have several possible interpretations and since the dictionary and morph work with words in isolation, there is no way to disambiguate. In such cases, it is best to retain the ambiguity in implicit form, rather than provide several different tags. The clitic 'aMte' is an example of this.

Since we have decided to give complete grammatical information for all entries in the dictionary and since we have decided to store only the irregular forms of pronouns in the dictionary, we would store, for example, 'nanna' in Kannada as a personal pronoun in first person, singular, genitive case. We would be deriving 'nannannu' through rules of nominal morphology. This is the accusative form, but dictionary would be showing 'nanna' as the root in genitive case. One pronoun cannot have two cases, here the final case should be accusative, not genitive. The bridge module clears up and gives the correct analysis as the final output.

Look at the Kannada word 'maaDalilla'. Structurally, this is 'maaDu (verb) (Do) + alu (infinitive) + illa (existential negative)'. This is how grammar books generally analyze this word. If you look at the meaning, there is no sense of the infinitive, this is simply the past tense form. The infinite suffix is used as a glue in Kannada and Telugu, so that a variety of other suffixes can be added, there is really no infinitive sense. More interestingly, here it actually signifies past tense. Grammar

is not all simple, neat and round, like school text books tend to suggest. The mapping from structure to meaning is not so straight forward in all situations. Our strategy here is to work with structure first and correct aberrations if any in the bridge module. This makes sense since computers can only work with structure, not with meanings directly.

The bridge module can handle ambiguities, unknown words, proper names or named entities, spelling errors, colloquial forms, etc. All this cleaning up would make the job of the subsequent modules so much simpler.

3 Experiments and Results

Here we shall give details of some experiments we have carried out for Telugu. The status of Kannada is similar.

We have performed POS tagging experiments on various corpora. F1 is a randomly selected file from TDIL corpus. F2 is set of sentences extracted from the TDIL corpus containing about 15,000 most frequent words from this corpus. These most frequent words have a great significance not just because they account for more than 60% of the whole corpus but also because they include the most confusing items from the point of view of lexicon, morphology and tag assignment. The rest of the words forming the bulk of the corpus are the simplest as far as tagging is concerned. In order to be sure that there is no over-fitting for any particular data set, we have next attempted tagging files F3 and F4 from the Eenadu Telugu daily newspaper. The table below shows the performance of the system as on date. Here D indicates the number of words directly found in the dictionary, M indicates the number of words analyzed by the morph, D-AMB is the number of words found in the dictionary and having more than one tag, M-AMB is the number of words analyzed by morph and having more than one tag. UNK indicates the number of words that remain untagged.

It may be observed that about 40% of the word forms are directly found in the dictionary and 50-60% of the words are analyzed by morph. Unless the texts contain a large percentage of proper nouns (ex. F4), only some 5-10% of the words remain untagged. Most of the unknown words are loan words, proper nouns, or words involving external sandhi or compounds. Further work on dictionary and morphology can reduce the number of untagged words in future. Of the tagged words, only about 10% of the words have more than one tag assigned. The maximum number of tags that can get assigned is 4 but this is a very rare case. Even getting 3 analysis is a rare phenomenon. Only a few typical kinds of ambiguities occur most of the times and preliminary studies have shown that a majority of them will automatically get resolved through chunking and parsing. We may not need a statistical approach to resolve these but if one wishes, we can easily tag the whole corpus using our system and create training data from that. One can also try a variety of heuristic rules to resolve the remaining ambiguities.

Upon careful observation of the tagged samples, we find that most words are tagged correctly. In order to be doubly sure, we again tagged 252 sentences selected from various grammar books including a wide variety of sentence structures. These sentences include 1278 tokens. All of these get tagged. Only two words were tagged incorrectly. Only 140 words are assigned more than one tag.

4 Conclusions

In this paper we have presented a new approach to tagging based on our theory of language, grammar and computation. We have described in detail a large, hierarchical tag set being designed by us for Dravidian. We have demonstrated the viability and merits of our ideas through actually developed systems for Kannada and Telugu. More work is on.

File	#Sent	#Tok	Dict	Morph	M-AMB	D-AMB	UNK	Ambiguity
F1	365	4910	2186 (45%)	2389 (49%)	158 (3%)	170 (4%)	313 (6%)	328 (7%)
F2	15100	76004	45225 (59%)	31103 (41%)	4917 (6%)	3194 (4%)	20 (0%)	8111 (10%)
F3	33	282	107 (38%)	173 (61%)	15 (5%)	10 (3%)	2 (1%)	25 (9%)
F4	27	237	88 (37%)	99 (42%)	8 (3%)	7 (3%)	50 (21%)	15 (6%)

Table 1: Tagging Performance

5 Bibliography

1 Kavi Narayana Murthy,"Computational Grammars for Indian Languages", Central Institute of Indian Languages, Mysore, forthcoming.

2 Kavi Narayana Murthy and Badugu Srinivasu, "A New Approach to Tagging in Indian Languages", CIIL, Mysore

3 Kavi Narayana Murthy, "A Network and Process Model for Morphological Analysis/Generation", ICOSAL-2, The Second International Conference on South Asian Languages, 9-11 January, 1999, Punjabi University, Patiala, India

4 G Bharadwaja Kumar, Kavi Narayana Murthy and B B Chaudhuri, "Statistical Analysis of Telugu Text Corpora", International Journal of Dravidian Languages, 36:2, June 2007, pp 71-99

5 CH. Narsinga Rao and Kavi Narayana Murthy, "On the Design of a Hierarchical POS Tagset for Telugu ", MTech thesis, Department of Computer and Information Sciences, University of Hyserabad, 2008

6 K. Anil Kumar, "Morphological Analysis of Telugu Words", MTech thesis, Department of Computer and Information Sciences, University of Hyderabad, 2003

7 Sankaran Baskaran, "Hindi Part of Speech Tagging and Chunking",Proceedings of NLPAI Machine Learning Workshop on Part of Speech Tagging and Chunking for Indian languages, IIIT Hyderabad, Hyderabad, India, 2006

8 Rama Sree R.J, Umamaheshwar Rao G and Madhu Murthy K.V, "Assessment and Development of POS Tagset for Telugu", The 6th Workshop on Asian Language Resources, IJCNLP, IIIT Hyderabad, Hyderabad, India, 11-12 January, 2008

9 Umamaheshwar Rao G, "Compound Verb Formation in Telugu", National Workshop-cum-Seminar on Lexical Typololgy, Telugu University, Hyderabad, 1996