A New Approach to Tagging in Indian Languages

Kavi Narayana Murthy and Srinivasu Badugu

School of Computer and Information Sciences,
University of Hyderabad, India
knmuh@yahoo.com,srinivasucse@gmail.com

Abstract. In this paper, we present a new approach to automatic tagging without requiring any machine learning algorithm or training data. We argue that the critical information required for tagging comes more from word internal structure than from the context and we show how a well designed morphological analyzer can assign correct tags and disambiguate many cases of tag ambiguities too. The crux of the approach is in the very definition of words. While others simply tokenize a given sentence based on spaces and take these tokens to be words, we argue that words need to be motivated from semantic and syntactic considerations, not orthographic conventions. We have worked on Telugu and Kannada languages and in this paper, we take the example of Telugu language and show how high quality tagging can be achieved with a fine grained, hierarchical tag set, carrying not only morpho-syntactic information but also some aspects of lexical and semantic information that is necessary or useful for syntactic parsing. In fact entire corpora can be tagged very fast and with a good degree of guarantee of quality. We give details of our experiments and results obtained. We believe our approach can also be applied to other languages.

Keywords: Tagging, Morphology, Part-Of-Speech, Lexicon, Hierarchical Tag Set, Telugu

1 Introduction

Word classes such as noun, verb, adjective and adverb are called 'Parts of Speech' (POS) by tradition. For the sake of convenience, we may use short labels such as N and V, called tags. Tagging is the process of attaching such short labels to indicate the Parts of Speech for words. One can actually go beyond syntactic categories and/or sub-categories and include lexical, morphological or even semantic information in the tags depending upon the need. In this paper we use the terms Tag and Tagging in this slightly broader sense.

Lexical, morphological and syntactic levels are well recognized in linguistics. Linguistic theories normally do not posit separate tagging or chunking levels at all. 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.

However, in practice it has generally been found that tagging can significantly reduce lexical ambiguities and thereby speed up syntactic parsing. Tagging is thus useful only to the extent it reduces ambiguities. Of course tagging can also help in other tasks such as word sense disambiguation, text categorization and text summarization.

There are mainly two broad approaches for POS tagging: 1) Linguistic, Knowledge Based or Rule Based approaches 2) Machine Learning or Stochastic or Statistical approaches (HMM and Viterbi decoding, for example). Combinations of the two are also used. We may either do a purely statistical tagging first and then rule out linguistically impossible assignments, or, we may start with linguistically possible tag assignments and then use statistics to the choose the 'best' assignments. Stochastic tagging techniques can be either supervised / unsupervised / hybrid. One may think of tagging as assignment of tags to words or as disambiguation of possible tags. It may be noted that a dictionary or a morphological analyzer typically looks at words in isolation while a tagger looks at the sentential context and attempts to reduce the possible tags for a given word in context in which it appears. Statistical approaches may assign a tag sequence to a word sequence, instead of assigning tags to individual words. Each method has its own merits and demerits.

Machine learning approaches require training data. Generating training data is not an easy task and the quality and quantity may both be important considerations. Training data needs to be large and representative. Labeled training data can be either generated completely manually or tagged data generated by an existing tagger can be manually checked and refined to create high quality training data and both of these methods have their obvious limitations. In practice, we will have to live with sparse data and smoothing techniques used may introduce their own artifacts.

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[1]. Manual tagging and checking also become difficult and error prone as the tag set becomes large and fine-grained and so there is a strong tendency to go for small, flat tag sets in machine learning approaches [2–6]. Such small tag sets may not capture all the required and/or useful bits of information for carrying out syntactic parsing and other relevant tasks in NLP. Morphological features are essential for syntactic analysis in many cases. These have also been the conclusions of a practical experiment of using fine grained morphological tag set reported by Schmid and Laws[7]. Their experiments were carried out using German and Czech as examples of highly inflectional languages. Fine-grained distinctions may actually help to disambiguate other words in the local context. Flat tag sets are also rigid and resist changes. Hierarchical tag sets are more flexible. Thus the design of the tag-set is strongly influenced by the approach taken for tagging. Further,

it is also influenced by the particular purpose for which tagging is taken up. A dependency parser of a particular kind may need a somewhat different sort of sub-categorization compared to, say, parsing using LFG or HPSG. Re-usability of tagged data across applications is an issue.

Although rule based approaches may appear to be formidable to start with, once the proper set of rules has been identified through a thorough linguistic study, there are many things to gain. Linguistic approaches can give us deeper and far-reaching insights into our languages and our mind. Knowledge based approaches generalize well, avoiding over-fitting, errors can be detected and corrected easily, improvements and refinements are easier too. In a pure machine learning approach, we can only hope to improve the performance of the system by generating larger and better training data and re-training the system, whereas in linguistic approaches, we can make corrections to the rules and guarantee the accuracy of tagging. Rule based approaches are also better at guessing and handling unknown words [8].

In this paper, we present an approach that does not depend upon statistical or machine learning techniques and there no need for any training data either. No manual tagging work is involved. We can afford to use a large, fine-grained, hierarchical tag set and still achieve high quality tagging automatically. We get both speed and accuracy. In this paper, we have chosen to render all Telugu words in Roman [9].

2 Previous Work in Indian Languages

English morphology is very simple and direct to implement. Morphological features also very few. The number of tags used for English POS tagging system are not that large: it ranges from 45 to 203 (in the case of CLAWS C8 tag-set) [10]. Also, average number of tags per token is low (2.32 tags per token on the manually tagged part of the Wall Street Journal corpus in the Penn Tree-bank) [11]. The number of potential morphological tags in inflectional rich languages are theoretically unlimited [11]. In English many of the unknown words will be proper nouns but in inflectional and/or agglutinate languages such as Indian languages, many common nouns and verbs may be absent in the training corpus. Therefore, a good morphological analyzer helps [12, 13, 1].

POS tagging for English seems to have reached the top level, but full morphological tagging for inflectionally rich languages such as Romanian, Hungarian, is still an open problem [11]. Indian Languages are highly inflectional and agglutinative too.

A Rule based POS tagger for Telugu has been developed by Center for Applied Linguistics and Translation Studies, University of Hyderabad, India [14]. Here there are 53 tags and 524 rules for POS disambiguation. A Rule based POS

tagger for Tamil has been developed by AU-KBC research center, Chennai, India [15]. Here the tag-set developed by IIIT-Hyderabad, consisting of only 26 tags, is used[2]. There are 97 rules of disambiguation. They report a Precision of 92 percent.

Sandipan Dandapat et al proposed a POS tagger for Bangla POS tagging based on Hidden Markov Models (HMM) [16, 17]. The training data set contained nearly 41,000 words and test data set contained 5,127 words. Further, they made use of semi-supervised learning by augmenting the small labeled training set they had with a larger unlabeled training set of 100,000 words. They have also used a morphological analyzer to handle unknown words. They report an accuracy of around 89% on a test data of 10,000 words.

Pattabhi R K Rao et al. [15] proposed a hybrid POS tagger for Indian languages. Handling of unknown words is based on lexical rules. For Telugu the test data used by them consists of 6,098 words, out of which only 3,547 are correctly tagged. Precision and Recall for Telugu were 58.2% and 58.2% respectively.

Asif Ekbal et al. [18] proposed a HMM based POS tagger for Hindi, Bengali and Telugu. Here they make use of pre-tagged training corpus and HMM. Handling of unknown words is based on suffixes and Named Entity Recognition. Reported accuracies are 90.90% for Bengali, 82.05% for Hindi and only 63.93% for Telugu.

Pranjal Awasthi et al. [19] proposed an approach to POS tagging using a combination of HMM and error driven learning. They have used Conditional Random Fields (CRF), TnT, and TnT with Transformation Based Learning (TBL) approaches and have reported F-measures of 69.4%, 78.94%, and 80.74% respectively for the three approaches for Hindi.

Sankaran Baskaran [20] used HMM based approach for tagging and chunking. He achieved a Precision of 76.49% for tagging and 55.54% for chunking using the tag-set developed in IIIT-Hyderabad [2], consisting of only 26 tags.

Himanshu Agrawal and Anirudh Mani [21] presented a CRF based POS tagger and chunker for Hindi. Various experiments were carried out with various sets and combinations of features which mark a gradual increase in the performance of the system. A morph analyzer was used to provide extra information such as root word and possible POS tags for training. Training on 21,000 words, they could achieve an accuracy of 82.67%.

Thus, most of the work done so far report accuracies of up to about 90% when tagged with small, flat tag sets. As we shall see, our approach guarantees much higher accuracies although we use a very large, fine grained, hierarchical

tag set. Unlike other systems reported above, our system has been tested on very large data.
A New Approach to Tagging in Indian Languages

3 Morphology Based Tagging

The main difference between our approach and all other work on tagging, whether for Indian languages or for other languages of the world, is the way we define words. The general practice is to tokenize sentences based on spaces and take for granted that these tokens are words. Sequences of characters separated by spaces are not necessarily proper linguistic units. Words have to be defined based on meaning and morphological and syntactic properties. We define a word as a sequence of phonemes bearing a definite meaning and having certain syntactic relations with other words in the given sentence. We need to define a set of syntactic relations that are universally applicable to all human languages. For example, a word which indicates an activity is a verb. If there is one activity, there can be only one verb. Thus 'has been running' is one word, not three. Similarly, 'from the book' is one single word - prepositions, post-positions are not universal word classes, 'from' is not a word in itself, it only adds a morpho-syntactic feature to 'book'. Viewed from this perspective, English morphology is not significantly simpler than the morphology of any other language. Thus, although 'book' and 'books' are both ambiguous between a noun and a verb in English, the words 'from the book' and 'from the books' are both unambiguous and it is morphology which is disambiguating here. This theory of words is a very significant research contribution to NLP and modern linguistics and full details are published elsewhere [22, 23].

Statistical approaches assume that the information necessary for tag assignment comes from the other tokens in the sentence. In many cases, only the tokens 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, in all languages of the world. The crux of tagging lies in morphology. This is clearly true in the case of so called morphologically rich languages but we believe this is actually true of all human languages if only we define words properly, in terms of meanings and universal grammatical properties, rather than in terms of the written form as a sequence of characters delimited by spaces.

A vast majority of the words can be tagged correctly by looking at the internal structure of the word. 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 looking at a sentence as a mere sequence of words is not sufficient. Statistical techniques are perhaps not the best means to capture and utilize complex functional dependencies between words in a sentence. Instead, syntactic parsing will automatically remove most of the tag ambiguities. It must be reiterated that tagging

is intended only to reduce tag ambiguities, not necessarily to eliminate all ambiguities. Syntactic parsing systems are anyway capable of handling ambiguities.

Identifying words is thus a critical task, mere tokenization based on white spaces will not do. In Dravidian languages (including Telugu, Kannada, etc.), as also in Sanskrit, the difference between orthographic tokens and proper words is not too much. Whatever be the case, differences can be handled using several techniques. A pre-processing module can be introduced with the main intention of first tokenizing and then obtaining words from these tokens. In Telugu, we do this using regular expression based pattern matching rules. Languages like English and Hindi may require more complex rules. In certain cases, mainly sandhi (phonetic conflation) and compounds, the morphology module is itself designed to handle these differences. A post-morphology bridge module ensures that we finally have proper words, tagged and ready for further processing such as syntactic parsing.

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, ensuring that the correct structure (and hence meaning) are depicted by the tags. The overall tag structure remains the same throughout, making it so much simpler and easier to build, test and use.

The morph system is implemented as an extended Finite State Transducer. The FST has 398 transitions or arcs. The figure below shows a small part of the FST. A category field has been incorporated so that only relevant transitions are allowed. Derivation is handled by allowing category changes. Transitions are on morphemes, not on individual characters or letters. Dravidian morphology involves complex morpho-phonemic changes at the juncture of morphemes and linguistically motivated rules have been used to handle these [24].

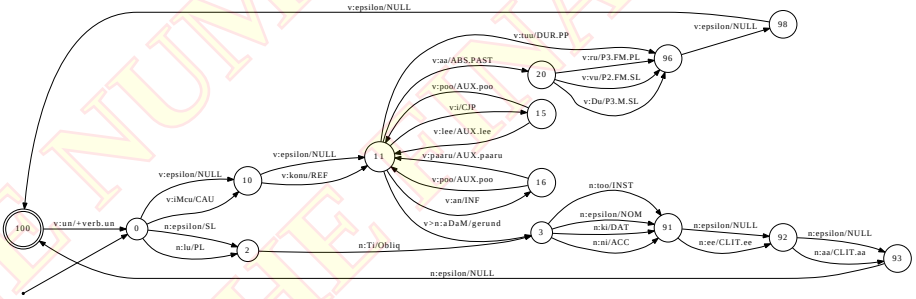

Fig. 1. Sample FST Grammar

We find that in any running text approximately 40% of the words are found directly in the dictionary. Less than 2% of the words in the dictionary are ambiguous. About one third of these are ambiguous between noun and verb. Since nominal and verbal morphologies are more or less completely disjoint in Telugu, and since these words occur mostly in inflected forms (more than 92% of times), morphology can resolve most of these cases of ambiguity. Morphology can also resolve ambiguity between nouns / verbs other categories such as adjectives and adverbs. Thus, morphology has a very important role in tagging. If we work with proper words instead of tokens, we believe we will get a similar picture in other languages. Certain kinds of systematic structural ambiguities in a language can lead to multiple tag assignments, calling for further disambiguation.

4 Tag Set Design and Tagging

Tags must be assigned to words, not to tokens. This is where we differ from all others. Once we have a precise definition of what constitutes a word and once we have a clear idea of universal word classes, the main grammatical categories and tags can be defined accordingly. The main categories should ideally be semantically motivated and hence universal and language independent. Nouns and verbs are universal categories with an independent and clear lexical meaning. Adjectives and manner adverbs have dependent lexical meaning and can also be taken as universal categories. Pronouns are variables, they do not have a fixed lexical meaning, but their meaning can be resolved in context. These five are the universal lexical categories. Conjunctions are typical of functional categories.

Although the major categories are semantically motivated, it must be noted that in the actual analysis process, we start from characters, build tokens and hence words, and work bottom-up through dictionary look-up / morphological analysis towards syntactic analysis leading to semantics. Since computers cannot work directly with meanings, we will have to work keeping lexical, morphological and syntactic properties in mind. Subcategories are thus dependent to some extent on the intended purpose and architectural and design issues. Each tag should then be precisely defined and supported with examples, need and justification. We give here the summary of our tagging scheme - see [25] for more details.

N (NOUN)	ADV (Adverbs)
COM(Common)	MAN(Manner)
PRP(Proper)	CONJ(Conjunctive)
-PER(Personal)	PLA(Place)
-LOC(Location)	TIM(Time)
-ORG(Orgzn.)	NEG(Negative)
-OTH(Others)	QW(Question Word)
LOC(Locative)	INTF(Intensifier)
CARD(Cardinal)	POSN(Post-Nominal
PRO (Pronoun)	Modifier)

	PER(Personal)		ABS (Absolute)
	INTG(Interrogative)	CONJ (Conjunction)	
	REF(Reflexive)		SUB(Subordinating)
	INDF(Indefinite)		COOR(Coordinating)
ADJ (Adjective)		V (Verb)	
	DEM(Demonstrative)		IN(Intransitive)
	QNTF(Quantifying)		TR(Transitive)
	ORD(Ordinal)		BI(Bitransitive)
	ABS(Absolute)		DEFE(defective)
SYMB (Symbol)		INTJ (Interjection)	

Table 1: LERC-UoH Tag Set

Here are some examples of tags in the dictionary.

```
baDi||N-COM-COU-N.SL-NOM                    muduru||ADJ-ABS||V-IN
aMdamaina||ADJ-ABS                          telusu||V-DEFE
adhikaari||N-COM-COU-FM.SL-NOM              tinu||V-TR
ataDu||PRO-PER-P3.M.SL-DIST-NOM             paatika||N-CARD-NHU-NOM
```

Here PRO-PER-P3.M.SL-DIST-NOM as a whole is called a tag. A tag consists of a series of tag elements separated by hyphens. The first element is always the main category and the next one or two levels indicate syntactic or morphological subcategories. The rest are morphological or semantic features. There is a more or less one-to-one correspondence between these elements and the morpheme structure of words. When a morpheme indicates more than one feature, the individual features are indicated as tag atoms within the given element, as in the case of P3.N.SL. In our Telugu dictionary, there are 274 unique tags made up of 143 tag elements and 121 atoms. Morph refines and/or adds more information. For example, 'ceppu' is a verbal root listed in the dictionary and 'ceppinavaaDu' is a pronominalized form derived by morphology and the corresponding tags are:

```
ceppu||N-COM-COU-N.SL-NOM||V-TR12

ceppinavaaDu||ceppu||V-TR12.v-PAST.RP-.adj
                    -PRON.vaaDu.P3.M.SL-.n-NOM
```

In the final analysis there are more than 20,000 tags for nouns (including number, case, clitics, vocatives, pronominalized forms, etc.) and nearly 15 Million different tags for verbs (including inflection, derivation, clitics etc.) Our morph is capable of generating and analyzing all these word forms. The tags contain all the necessary lexical, morphological, syntactic and relevant semantic information for carrying out syntactic analysis etc. without need for getting back to the dictionary or morphology.

Most of the other works on morphology for Indian languages are based on the Paradigm Model where lists of word forms are manually created for each paradigm based on morpho-phonemic considerations but as reflected in the orthography. It is next to impossible to create complete lists of all word forms manually given the richness of morphology of our languages. Nor is this an intelligent or wise approach. It is very unlikely that the human mind simply lists all forms of all words in tables. Also, morphology is reduced to arbitrary string manipulation in this paradigm approach. For example, in Telugu, 'maniShi' (person) becomes 'manuShulu' (persons) in plural. In the paradigm approach, 'man' is identified as the common prefix and 'maniShi' is broken into 'man' and 'iShi'. Then, 'manuShulu' is obtained by adding 'uShulu' to 'man'. Since 'man', 'iShi', 'uShulu' are all totally arbitrary, meaningless, linguistically unacceptable units, this is really not morphology at all. Ours is perhaps the first, linguistically motivated, psychologically plausible, nearly complete, computationally efficient morphological system for any Indian language. It may be noted that many other works for various languages across the world are also based on arbitrary character level manipulations. A proper system of morphology will be of great help not only in tagging but also for spell checking, stemming / lemmatization etc. More importantly, it will provide insights into the way the language works. A proper system of morphology will be useful for language teaching and learning too.

Morph can resolve a major portion of tag ambiguities. For example, the Telugu word 'ceppu' has two meanings: 1) 'to say or to tell' 2) shoe or slipper. The examples below show how morphology can resolve the noun-verb ambiguity. In the case of derivations, note how our tags depict the complete flow of category changes. This is essential for syntactic parsing.

```
ceppu||N-COM-COU-N.SL-NOM||V-TR12

ceppaaDu||ceppu||V-TR12-ABS.PAST-P3.M.SL

ceppinavaaDu||ceppu||V-TR12.v-PAST.RP-.adj
                    -PRON.vaaDu.P3.M.SL-.n-NOM

ceppulanu||ceppu||N-COM-COU-N.PL-ACC
```

When morph fails to disambiguate, syntactic considerations such as chunking constraints, predicate-argument structure and selectional restrictions can resolve the ambiguities in most cases. Less than 1% of words will remain ambiguous as can be seen from our experiments below.

Disambiguation by purely statistical methods have also been used by researchers [26]. Although all words can be disambiguated, there can be no guarantee of correctness, even in cases where clear disambiguation rules exist linguis-

tically. A rule-based disambiguation will usually leave out only those ambiguities which are genuine.

5 Experiments and Results

There are no publicly available standard data sets available for Telugu. We have developed our own Telugu text corpus of about 50 Million words [27]. We have tested our system on a corpus of 15 Million words. Performance of the morph analyzer on randomly selected sentences from this corpus is shown below:

#Sent	#Tokens	Found in Dict	Identified by Morph	Unknown
101	861	(376) 44%	(402) 46%	(83) 10%
500	4788	(2058) 43%	(2330) 49%	(400) 8%
1000	9269	(3869) 42%	(4691) 50%	(709) 8%
1500	14092	(5860) 42%	(7105) 50%	(1127) 8%

Table 2. Results of Morph Analysis on Telugu Corpora

Eight to ten percent of the words remain un-analyzed. We have options for guessing but here we show results without guessing. It is found on close inspection that most of the un-analyzed words are spelling errors, loan words, named entities and compounds. Among the words analyzed, it is found that around 10% of words are assigned more than one tag. In most cases of ambiguity, words get only two tags, not more. More importantly, the correct tag is almost always included.

Since ours is a manually created rule based system, there is no scope for chance errors. Incorrect analysis is very rare and occurs only due to complex interactions involving spelling errors, loan words, named entities etc. In order to evaluate the Precision and Recall, random samples have been manually checked. A random sample of 202 sentences consisting of 1776 words has been tagged and manually checked carefully. Of these, 1626 words (91.5%) were tagged, the rest remain untagged. Only 5 words (0.3%) were found to be incorrectly tagged. This gives us a Precision of 99.69% and a Recall of 91.27%. In these calculations, a word has been taken to be correctly tagged if the correct tag is included, along with possibly other tags.

In cases of ambiguous tag assignments, we use a set of 17 rules based on local syntactic context to disambiguate the tags. About 90% of ambiguities can be resolved using these local rules. Finally, we find that we can tag more than 93% of all words in a raw corpus, with less than 1% of the words assigned more than one tag, and with a guarantee of more than 99% correctness.

6 Conclusions

A New Approach to Tagging in Indian Languages

In this paper we have presented a new approach to tagging based on our new theory of words, using a morphological analyzer and a fine-grained hierarchical tag-set. We have shown that it is possible to develop high performance tagging system without need for any training data or machine learning or statistical inference. Since the whole system is rule governed, the results can be guaranteed to be correct. Manual verification has validated this claim. We have demonstrated the viability and merits of our ideas through actually developed system for Telugu. The same ideas and methods have been used to develop a system for Kannada and the performance of our Kannada system is similar. The method is being applied for other languages too.

References

1. Atwell, E.: Development of Tag Sets for Part-of-Speech Tagging. In Ludeling, A., Kyto, M., eds.: Corpus Linguistics An International Handbook, Mouton de Gruyter (2008) 501–526
2. IIIT-Hyderabad: A Part-of-Speech Tagset for Indian Languages. `http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf`
3. AU-KBC: POS Tagset for Tamil. `http://nrcfosshelpline.in/smedia/images/downloads/Tamil_Tagset-opensource.odt`
4. Sankaran, B., Bali, K., Choudhury, M., Bhattacharya, T., Bhattacharyya, P., Jha, G., Rajendran, S., Saravanan, K., Sobha, L., Subbarao, K.V.: A Common Parts-of-Speech Tagset Framework for Indian Languages. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, European Language Resources Association (ELRA) (2008) 1331–1337
5. RamaSree, R.J., Rao, G.U., Murthy, K.V.M.: Assessment and Development of POS Tagset for Telugu. In: Proceedings of the Sixth Workshop on Asian Language Resources, 3rd International Joint Conference on Natural Language Processing (IJCNLP-08), IIIT Hyderabad, Hyderabad, India (2008) 85–88
6. Elworthy, D.: Tagset Design and Inflected Languages. In: In EACL SIGDAT workshop From Texts to Tags: Issues in Multilingual Language Analysis. (1995) 1–10
7. Schmid, H., Florian, L.: Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging. In: COLING. (2008) 777–784
8. Abney, S.: Part-of-Speech Tagging and Partial Parsing. In: Corpus-Based Methods in Language and Speech, Kluwer Academic Publishers (1996) 118–136
9. Murthy, K.N., Srinivasu, B.: Roman Transliteration of Indic Scripts. In: 10th International Conference on Computer Applications, University of Computer Studies, Yangon, Myanmar (28-29 February 2012)
10. Garside, R.: The CLAWS Word-Tagging System. In Garside, R., Leech, G., Sampson, G., eds.: The Computational Analysis of English, Longman (1987) 30–41
11. Hajič, J.: Morphological Tagging: Data vs. Dictionaries. In: Proceedings of the 6th Applied Natural Language Processing and the 1st NAACL Conference, Seattle, Washington (2000) 94–101

12. Huihsin, T., Jurafsky, D., Christopher, M.: Morphological Features help POS Tagging of unknown Words across Language Varieties. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, Jeju Island, Korea, Association for Computational Linguistics (October 2005) 32–39
13. Sawalha, M., Atwell, E.: Fine-grain Morphological Analyzer and Part-of-Speech Tagger for Arabic Text. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta (2010) 1258–1265
14. SreeGanesh, T.: Telugu POS Tagging in WSD. In Journal of Language in India **6** (August 2006)
15. Pattabhi, R.K.R., SundarRam, R.V., Krishna, R.V., Sobha, L.: A Text Chunker and Hybrid POS Tagger for Indian Languages. In: Proceedings of International Joint Conference on Artificial Intelligence Workshop on Shallow Parsing for South Asian Languages, IIIT Hyderabad, Hyderabad, India (2007)
16. Dandapat, S., Sarkar, S., Basu, A.: A Hybrid Model for Part of Speech Tagging and its Application to Bengali. In: Proceedings of International Conference on Computational Intelligence, Instanbul, Turkey (2004) 169–172
17. Dandapat, S., Sarkar, S.: Part of Speech Tagging for Bengali with Hidden Markov Model. In: Proceedings of NLPAI Machine Learning Workshop on Part of Speech Tagging and Chunking for Indian languages, IIIT Hyderabad, Hyderabad, India (2006)
18. Ekbal, A., Mandal, S.: POS Tagging using HMM and Rule based Chunking. In: Proceedings of International Joint Conference on Artificial Intelligence Workshop on Shallow Parsing for South Asian Languages, IIIT Hyderabad, Hyderabad, India (2007)
19. Awasthi, P., DelipRao, Ravindran, B.: Part of Speech Tagging and Chunking with HMM and CRF. In: Proceedings of NLPAI Machine Learning Workshop on Part of Speech Tagging and Chunking for Indian languages, IIIT Hyderabad, Hyderabad, India (2006)
20. Baskaran, S.: Hindi Part of Speech Tagging and Chunking. In: Proceedings of NLPAI Machine Learning Workshop on Part of Speech Tagging and Chunking for Indian languages, IIIT Hyderabad, Hyderabad, India (2006)
21. Agarwal, H., Mani, A.: Part of Speech Tagging and Chunking with Conditional Random Fields. In: Proceedings of NLPAI Machine Learning Workshop on Part of Speech Tagging and Chunking for Indian languages, IIIT Hyderabad, Hyderabad, India (2006)
22. Murthy, K.N.: Language, Grammar and Computation. Central Institute of Indian Languages (CIIL), Mysore (Forthcoming)
23. Murthy, K.N.: What Exactly is a Word? Special Issue of International Journal of Dravidian Language (Forthcoming)
24. Krishnamurthi, B., Gwynn, J.P.L.: A Grammar of Modern Telugu. Oxford University Press, New Delhi (1985)
25. Murthy, K.N., Srinivasu, B.: On the Design of a Tag Set for Dravidian Languages. In: 40th All India Conference of Dravidian Linguists, University of Hyderabad, Hyderabad, India (18-20 JUNE 2012)
26. Steven, J., DeRose: Grammatical Category Disambiguation by Statistical Optimization. Computational Linguistics **14**(1) (1988) 31–39
27. Kumar, G.B., Murthy, K.N., Chaudhuri, B.B.: Statistical Analysis of Telugu Text Corpora. In International Journal of Dravidian Languages **36**(2) (June 2007) 71–99