International Journal of Computer Science and Engineering (IJCSE)
ISSN(P): 2278-9960; ISSN(E): 2278-9979
Vol. 3, Issue 3, May 2014, 179-186

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

IMPROVEMENT OVER IL-POST TAGSET FOR KANNADA

BHUVANESHWARI C. MELINAMATH

Department of Computer and Information Sciences, University of Hyderabad, Hyderabad, Andhra Pradesh, India

ABSTRACT

Compilation of tag set is an important task in all NLP. It is the initial stage in all NLP applications. We are focusing on improvements over the IL-POST (Indian language part of speech tag set) in this paper. Our tag set is fine grained and captures detail information. We have developed this tag set keeping higher NLP applications in mind. Fine grained tag set is useful for NLP applications like chunking, parsing, morphological analyzer and machine translation etc. We follow EAGLES (Expert Advisory Group on Language Engineering Standards) as guideline with modifications as required for our Kannada Language.

The morphology of Kannada is complex as comparable to Turkish and Finnish. This tag set can be adopted for whole Dravidian language family. This is Hierarchical tagset and is largely based on computational needs. We have compiled a tag set of 170 tags. Compilation of tag set is an important task in all NLP and is quite challenging for Languages like Kannada. This paper will look at solving the open issues left unsolved in Microsoft's IL-POST tag set like clitics, auxiliaries. Modal auxiliaries etc. Tagging efficiency rate is more than 90% in Our tag set as compared existing ones.

KEYWORDS: Expert Advisory Group on Language Engineering Standards) EAGLES, Machine Translation (MT), Natural Language Processing (NLP), Part of Speech (POS)

1. INTRODUCTION

Natural Language Processing (NLP) is an area of artificial Intelligence (AI). Natural Language Processing (NLP) is concerned with the computational aspects of the human language. The goal of the NLP is to analyze and understand natural languages used by human beings. Natural language understanding requires extensive knowledge about the outside world and the ability to manipulate such knowledge.

Kannada is a Dravidian language. This is spoken in southern India. Kannada has complex morphology. Words are built up from roots by following fixed patterns that add prefixes, suffixes and infixes to the word. This system that studies how words are constructed from roots, and describes the patterns they follow which in English could be called derivational morphology. Kannada contains three genders masculine, feminine and neuter. Kannada contains three numbers instead of the more common two numbers.

So as well as singular and plural, there is also the dual that is used for describing the actions of two people. All these attributes are taken into account when constructing the tagset. Kannada has complex morphology and designing the tag set for such languages is not an easy job. Kannada language uses 49 phonemic letters, divided into 3 groups: 13 swaragaLu (called vowels in English), 34 vyaNjangaLu (called consonants in English) and 2 yogavaahakagaLu (neither consonants nor vowels), anusvara, namely, "aM" and visarga, namely, "ah".

1.1 What is Tag?

Tag is a grammatical information like verb, noun etc. Tagging or Annotation is the process of adding some additional information (grammatical features like word category, case indicator, other morph features) about the word in the text. The set of all these tags is called a tagset.

We have developed the morphosyntactic tagset using hierarchical approach. The tagset is largely based on computational needs. We have compiled a tag set of 170 tags. We have 10 main categories in the top level and subtypes in second level and further classification in lower levels. We have designed tags to capture all kinds of inflections of the word in our tag set, noun features like case, number, clitics, double clitics are handled, similarly finer distinction in adjectives and adverbs are also made. The tag set is designed keeping parsing in mind. Even though parsing is not attempted in our work. We have used these tag set in building an electronic dictionary of 30000 words and also in our morphological analyzer system in further stages of our research.

The paper is organized into following sections. In section 2 literature survey is explained, in section 3 Justification for our proposed tagset is described. In section 4 comparison of flat versus hierarchical tag set is explained. In section 5 comparison of our Kannada tagset with the Microsoft guideline is explained and section 6 discuses results. Section 7 difficulties faced in designing the tag set and section 8 gives the conclusion.

2. LITERATURE SURVEY

(Leech and Wilson, 1990) have proposed EAGLES standard for morphosyntactic annotation of European language. EAGLES describe 11 major categories. They proposed 114 tags for English 274 for Italian. (Hardie, 2004) has proposed hierarchical tag set for Urdu. He has developed 280 tags for Urdu. This design followed the description of Urdu grammar given by Schmidt in 1999 for the purpose of tagset design. (Shereen Khoja et al., 2001) have proposed tagset for morphosyntacic tagging of Arabic. They have devised 177 tags, 103 for nouns, 57 verbs, 7 residuals and 1 punctuation. This is an extended tagset and includes, voice, transitive features for verb and derivation for nouns. (Santorini, 1990) has proposed part of Speech tagging guidelines for Penn Treebank Project in English, the tagset is popular even today. It is flat tag set consist of 35 tags. (Garside, 1987) has proposed tag set for English.

It is known as CLAWS (C5) Tag set and has 60 tags. (IIIT Hyderabad, 2007) has proposed tag set for Indian language consists of 26 tags that capture only coarse level categories that do not include finer morpho-syntactic features of Indian languages. the tags seem more suitable for Indo-Aryan languages like Hindi. (AU-KBC, 2006) has proposed tag set for Tamil. The tag set consists of 68 tags. (Baskaran et. al, 2008) have proposed common part of speech tagset framework for Indian languages. It is generally referred as IL-POST. In this work many issues left unsolved like information on transitivity on verbs, clitics information, handling of aspect auxiliaries and modal auxiliaries are not attempted.

From the literature survey, it is observed that not much work is carried out on design of tag set for Kannada. The existing tag sets remained incompatible with each other in terms of morph-syntactic categories and features, tag definitions, levels of granularity, etc. The tag set design plays a vital role when data is tagged according to it and hence it affects the development of NLP tools within and across the languages. This scenario made us to move in the direction of development of hierarchical tag set for Kannada.

3. DESIGN AND DEVELOPMENT OF MORPHOSYNTACTIC TAGSET FOR KANNADA

The Hierarchy of a proposed morphosyntacic tagset is shown in Figure 1.

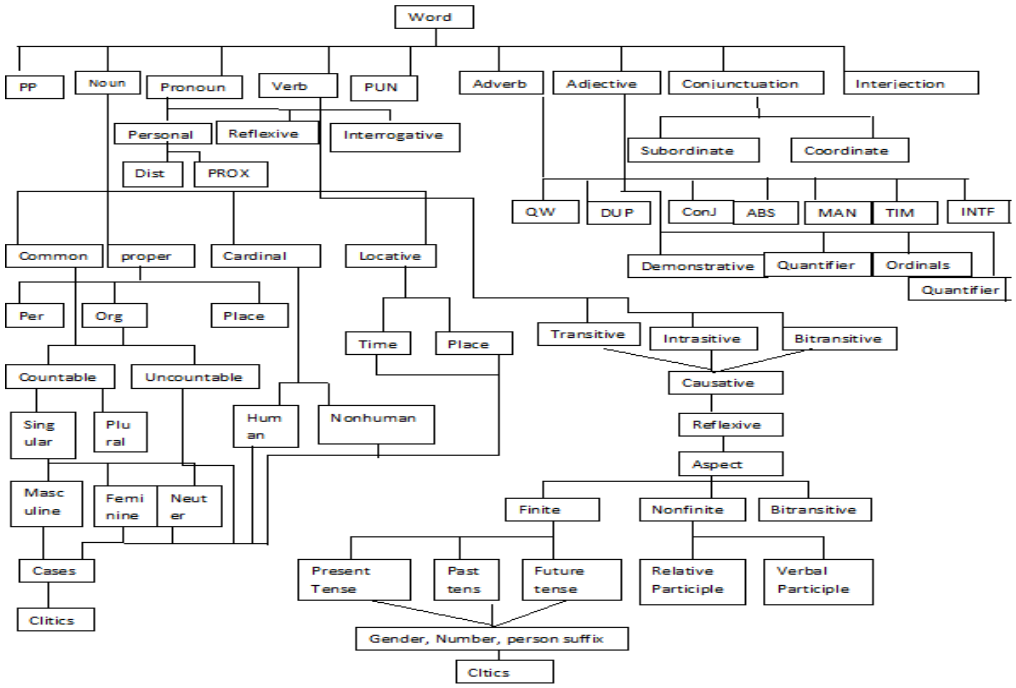

Figure 1: Tagset Hierarchy

The methodology says that major word classes should be in the top level in the tree, followed by sub classifications and lastly morphological features. An issue of general concern is that in an effort to reduce the number of tags we should not miss out the crucial information related to grammatical and other relevant linguistic knowledge which is encoded in a word, particularly in agglutinating languages like Kannada. It is better to encode all features in word.

4. IMPROVEMENT OVER MICROSOFT IL-POST FRAMEWORK

(Baskaran et al., 2008) have proposed Common Frame work guidelines for Indian languages using EAGLES as basis model. It is referred as IL-POST. Our design of tag set is developed keeping parsing in mind. So it is felt that, semantic information is also required to some extent for example what kind of argument the verb takes and marking of transitivity, intransitive, bitransitive information are required in higher application like Machine translation, Anaphora resolution application, parsing etc. IL-POST tagset does not capture such information. Another major issue is, in Kannada the interrogative sentences are formed by using clitics this information is also missing in IL-POST framework. Handling of aspect auxiliaries, formation of conjunctive verbs and compounding are missing in IL-POST tag set. To start with we focus on clitics first we have identified 4 types of clitics.

- ***Clitics*** information is not handled in Baskran et al. guidelines. Clitics provide lot of information for higher level analysis for example during parsing. Four clitics are identified viz. Emphatic clitic (ee), Interrogative clitic(aa), Inclusive clitic(uu) and Indefinite clitic (oo). Clitics do occur in next level i.e. clitic emphatic is followed by interrogative in the below example. This was an open issue in Baskaran et al. tagset, which we have overcame here.

Table 1: Comparison of IL-POST and Our Kannada HPOS

S. No	KHPOS for Kannada				IL-POST for Indian Language			
	Top Level	Sub Category (Level 1)	Sub Sub Category (Level 2)	Tag	Top Level	Sub Cat Gory (Level 1)	Sub Sub Category (Level 2)	Tag
1	Noun			N	Yes			N
		Common		COM		Yes	yes	C
			Countable/ Uncountable	COU/UNC			No	-
		We have decomposed Common noun as countable and uncountable. This distinction is necessary because uncountable noun have only case inflection but not number i.e. plural inflection. This information is useful in word form generation, otherwise morphological generation systems keeps on generating ungrammatical forms like below. This information is missing in Microsoft guidelines. Consider the word as given in Example. a) **¤ÃgÀÄUÀ¼ÀÄ** niirugaLu (water) in plural not correct grammatically						
		Proper			Yes			P
			Person	PER		No		-
			Location	LOC		No		-
			Organization	ORG		No		-
		Locative		LOC	Yes, Merged			NST
			Time	TIM		No		
			Place	PLA		No		
		Cardinals		CARD	Numerals			
			Human	HUM		Cardinals		NUMC
			Non human	NHUM		Ordinals		NUMO
		In (Baskaran et al, 2008). Framework ordinals (like MAzÀ£ÉÃAiÀÄ (oMdaneeya), (“first”) are kept under Quantifiers, and cardinals (MAzÀÄ (oMdu), “one”) are also kept under quantifiers this distinction is not correct, because though ordinals and cardinals are representing numbers, they are playing different roles in the sentences, ordinals are not inflected for cases and cardinal are inflected for case and cardinals can be head of the noun phrase, while ordinals cannot and ordinals are more like adjectives, they are used in deriving noun like other adjectives. Considering these features, ordinals are kept under adjectives since they act as noun modifiers like adjectives in our Kan-HPOS Tag set.						
2	Pronoun			PRP	Yes			PPR
			Personal			Yes		
				PROX				
				DIST				
			Reflexive			Yes		PRF
			Interrogative			Yes		PWH
			Kept under cardinals as human		<=	Reciprocal		PWC
	Reciprocals are placed under pronoun in IL-POST is not correct. But reciprocals like M¨ÉÆâ§ (obbobba) “each each person” has two functionality i.e. they are used as noun modifiers for example consider a sentence obbobba vyakatiyannu noo Duve, “I will see each each person”. In this example reciprocal is modifying noun but in general theory of all languages pronoun replaces noun is a accepted rule but pronoun modify noun is not accepted theory, hence in our tag set, such words are treated as adjectives and also they are treated specially as protonoun since they are further derived as nouns like M¨ÉÆâ§£ÀÄ (obbobbanu) (indicate masculine single person), obbobaLu(indicate feminine).							
3	Adjective			ADJ	Nominal	Adjective		JJ
						Quantifier		JQ
		Demonstrative		DEM		No		
		Quantifiers		QNTF		No		
		Ordinals		ORD		No		
		Absolute		ABS		No		
	IL-POST tag set has no finer distinctions in adjectives. However this distinction is necessary because the order of occurrence adjectives is useful information in a noun phrase grouping, like which adjective follows which kind of another adjective; it is observed that all true adjectives, quantifiers and ordinals cannot be kept in the same order. Chunking rules are Determined based on occurrence of constituents in a phrase.							

Table 1: Contd.,

4	**Adverb**		**ADV**			
		Time	**TIM**		**No**	
		Place	**PLA**		**No**	
		Reduplication	**RDP**	**No division**	**No**	
		Question	**QW**	**In adverbs**	**No**	
		Intensifier	**INTF**		**No**	
		Absolute	**ABS**		**No**	
		Conjunction	**CONJ**		**No**	
	In IL-POST tag set guidelines, there is no detail distinction for adverb also. However this division is necessary for handling of onomatopoeic forms like¸ ÀgÀ¸ ÀgÀ£É (sarasarane) “fast fast”, §gÀ§gÀ£É (barabarane) “fast fast”. which indicates emphasis is required. The finer distinction is necessary since all types of adjective or adverbs occur in particular order, which subtype follows which subtype, this information is useful in chunking while identifying noun group and verb groups. We have resolved many unsolved open issues mentioned by IL-POST tag set. And prove that our tag set is an exhaustive tag set.						
5	**Interjection**	-		**INTJ**		**Yes**	**IN**
6	**Conjunctuation**	Coordinating		**COOR**		-	-
		Subordination		**SUB**			
7	**Punctuation**			**PUN**			**PU**
8	**Verbs**	Transitive Intransitive.		**V**	**No**	**No**	**V**
			Finite			**Finite**	
			Nonfinite	**Aspect auxilaries (table 3)**		**Non finite**	**FIN**
			Infintive	**Modal auxiliaries (table 4)**			**NFN**
						Infinitive	**INF**
	Aspect Marker: In Kannada a set of suffixes may be added to verbal participle to give certain semantic nuances to the meaning of the sentence. Aspect markers are very similar to main verbs in their morphology and syntax. In fact they are derived from certain main verbs. But semantically they do not express the lexical meaning as that their main verbs express. The aspectual biDu ’completive’ does not mean the same as main verb biDu “leave,” In Kannada, the verbal aspect marker is usually added to the past verbal participle, then tense markers, modals etc. This feature is not captured in (Baskaran et. al, 2008). Handling of aspect auxiliaries is important factor because using this feature infinite number of derived verb stems can be generated. The list of aspect auxiliaries is shown in table 3.						
	Modal Auxiliaries: Kannada has a number of modal auxiliary verbs that are attached to /al/ form of the infinitive, different modal auxiliary give the notion of could, can, may, might, should, capable, not able modality of ability, compulsive, completion, prohibitive, permissive and opportunity and their negatives. This feature is currently not handled in (Baskran et.al, 2008). This feature is also useful in generation of many derived verb roots. We have handled this feature; this is also one of the factors leading to good tagging rate of KHOS tag set. The different modals are as table 2.. ªÀiÁqÀÄ + C¯ï+ §ºÀÄzÀÄ = ªÀiÁqÀ®Ä§ºÀÄzÀÄ maaDu + alu+ **bahudu = maaDalubahudu** **do + to + can = can be done** Modal auxiliaries contribute different shades of grammatical meaning. The different possible modal auxiliaries are shown in table 2						
9	**Postposition**		**PP**				
		Genitive	**PP-GEN**		**No**		
		Ablative	**PP-ABL**		**No**		
		Dative	**PP-DAT**		**No**		
		Accusative	**PP-ACC**		**No**		
		Comparative	**PP-COMP**		**No**		
		Purposive	**PP-PUR**		**No**		
		Similarities	**PP-SIM**		**No**		
		Associative	**PP-SOC**		**No**		
		Others	**PP-OTH**		**No**		
10	**Residual**		**RD**	**Yes**			**RD**

- Another open issue proposed in Microsoft guidelines is that, it is hard to distinguish between adjectival participle and verbal nouns for Bangla. But however this problem can be solved in Kannada. In Kannada there is productive rule for deriving noun from adjectives by adding third person pronoun suffix avanu (he) as shown in previous example **cikkavanu**. Relative participles obey this rule but verbal noun does not satisfy this rule so one can distinguish between adjective participle and verbal nouns easily.

Table 2: Table Showing Modal Auxiliaries

Modal auxiliary	English Meaning
¨ÉÃPÀÄ (beeku)	MUST (Want)
§ºÀÄzÀÄ (bahudu)	PROH(Should not)
¨ÉÃqÀ (beeDa)	NEG(IMP)
PÀÆqÀzÀÄ (kuuDadu)	PERM(May)
¯ÁgÀ(laara)	NCAP(might not)
§®è(Balla)	CAP(capable)
¥ÀqÀÄ (paDu)	PASS(Passive voice)
DUÀÄ (aagu)	Finality

Table 3: Inventory of Aspect Auxiliaries

Aspect Marker	Aspect Meaning	Aspect Marker	Aspect Meaning
©qÀÄ (biDu)	Completion	**DUÀÄ** (aagu)	Finality
ºÉÆÃUÀÄ (Hoogu)	Completion	**EgÀÄ** (iru)	Perfective
DqÀÄ (aaDu)	Continuity	**ºÁPÀÄ (**Haaku)	Exhaustive
PÉÆqÀÄ (koDu)	Benafactive	**PÉÆ¼ÀÄî(** koLLu)	Reflexive
£ÉÆÃqÀÄ (nooDu **)**	attemptive		

- Relative participles and conjunctive participles are kept under verb non finite forms in Microsoft guidelines work. In our work Even though relative participles act as adjectives but are not placed under adjectives since these can take negative suffix and Kannada do not have negative adjectives in Dravidian Languages. Hence these should be treated separately. Consider a word baarada (not coming one), here (baarada is a negative relative participle acting as adjective.)

We have solved many unsolved open issues mentioned by IL-POST tag set. And prove that our tag set is an exhaustive tag set.

6. EXPERIMENTS AND RESULTS

We have selected paragraph of text for tagging from famous Kannada daily news paper, and one sample from Kannada website Kannada yahoo.com another randomly generated and tagged these samples using IIIT-H tag set, IL-POST tag set and our KHPOS tagging scheme. We observed that tagging efficiency for our Kannada tag set is good as compared IIIT-H, tag set and IL-POST tag set.

Table 4: Tagging Rate Using Different Tag Sets

Tag Set	Synthetic (39 Words)	Yahoo Text (109 Words)	Prajavani (89 Words)
IIITH	33 tagged	83 tagged	81 tagged
IL-POST	27 tagged	91 tagged	77 tagged
KHPOS	39 tagged	104 tagged	89 tagged

Table 5: Showing Precision vs. Recall

	IL-POST	IIIT-H	KHPOS
Precision	82%	88%	99%
Recall	97%	96%	98%
F-measure	88%	91%	98%

However IIIT-H tags set are not useful in MT application, because MT applications require fine detail information about each word. If you simply tag all nouns as NN, the information like whether the noun is masculine/ feminine/neuter/plural all these information are not obvious.

But IL-POST are useful in MT application since they are fine grained and captures more information but fails to handle clitic information, new compound verbs, auxiliary and modal auxiliaries hence efficiency may be affected.

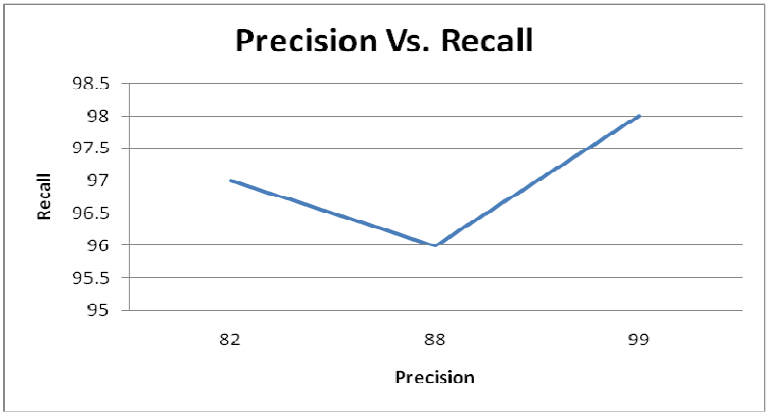

Figure 2: Precision versus Recall

Any noun tag which combines an N for noun with other characters to indicate other features of the word is decomposable. The tag "N-COM-COU-M.SL-NOM" is a single tag, this tag is decomposable and is analyzed as N=noun, COM=common, COU=countable, M.SL= Masculine singular, NOM=nominative, the decomposable elements of the tag set will indicate features in a hierarchy. The following illustrates few examples of the words and their KHPOS tags.

Example 1: HuDuga "boy": N-COM-COU-M.SL-NOM.

Example 2: Niiru "water": N-COM-UNC-N.SL-NOM.

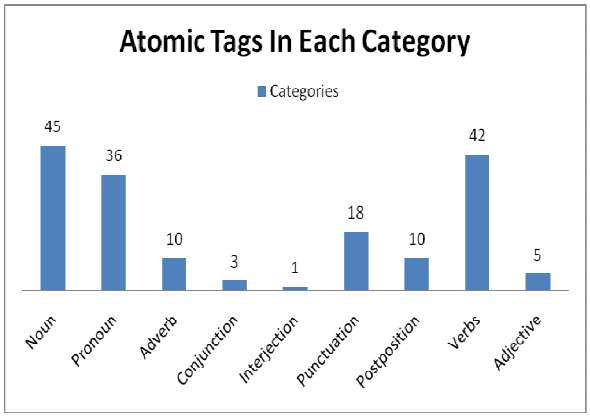

Figure 3: Graph Showing Tagging Different Sample Text

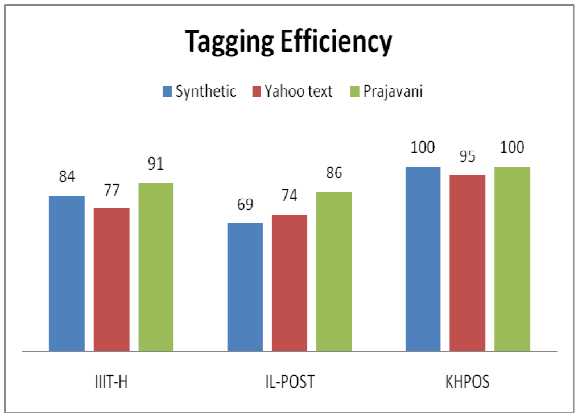

Figure 4: Category Wise Tags in KHPOS Tag Set Efficiency Using

7. CONCLUSIONS

HPOS tagset offers advantages such as flexibility, cross-linguistic compatibility, reusability, and decomposability. We have developed a dictionary of 30000 words using this hierarchical tag set. But IL-POST are useful in MT application since they are fine grained and captures more information but fails to handle clitic information, new compound verbs, auxiliary and modal auxiliaries hence efficiency may be affected. KHPOS tags are useful in MT applications and are fine grained and handled clitic, modal auxiliaries, conjunct verbs etc. and tagging efficiency is more here as compared to other two tag sets. Therefore the first step must be to make linguistically ideal tagset: The tagset which we would like to apply to our text in a real world. This ideal tagset will be the largest within the parameter laid out by hierarchical design principles, on the basis that it is always easier to remove distinctions than to add them. We have tried to overcome the issues left unsolved in IL-POST Tag set.

REFERENCES

1. Leech G and Wilson, "A. Recommendations for morph syntactic annotation of corpora". EAGLES Technical Report, 1996.
2. IIIT-tagset, "A Part-of-Speech Tagset for Indian. 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-open source.odt
4. A. Hardie. "The computational Analysis of Morphosyntactic Categories in Urdu". Ph.D thesis, Department of Linguistics, Feb 2004.
5. Baskran et al. "Designing a common pos-tagset framework for Indian languages". Proc. of 6th Workshop on Asian Language Resources, Language Technologies. Research Centre, IIIT, Hyderabad, 2008
6. A. Hardie. "Developing a tagset for automated part-of speech tagging in Urdu. Proc. Of the Corpus Linguistics 2003 conference, 16, 2003.
7. Santorini B. "Part of Speech tagging guidelines for the Penn Treebank Project". Technical report MS-CIS 90-47. Department of Computer and Information Science, University of Pennsylvania. 1990
8. Greene. B and Rubin G M "Automatic grammatical tagging of English", Providence, R.I:. Department of Linguistics, Brown University, 1981.
9. Garsdie, R. "The CLAWS word tagging system". In The Computational Analysis of English, ed. London, 1987.
10. Shreen Khoja. "A tagset for the morphosyntactic tagging of Arabic", Proc. of Corpus Linguistics. Lancaster University, Lancaster, UK, March 2001,