Interdisciplinary Journal of Linguistics
Volume [7] 2014, Pp. 248-261

Tagging: A Case Study of Kashmiri

Shahid Gilkar
Nahida Ali
Sumaira Nabi
Mansoor Farooq*

Abstract

Tagging may be defined as assigning words their appropriate parts of speech in a corpus. It can be done manually by using a software tool by human taggers. It can also be accomplished automatically by means of a computer program, with no or little human intervention. The labels attached to words by a human agent or a computer program, are called as Tags. The computer program is called as a Tagger.
This paper will attempt to propose a preliminary architecture for a Rule-based tagger of Kashmiri.

Keywords: Tagger, Corpus, Morphology, Syntax and Hidden Markov Model.

Introduction:
Tagging is one of the ways of annotating a corpus. Annotation can be defined as adding information to a corpus. In part-of-speech tagging the information we add is about the part of speech of each word in a corpus. This paper will confine itself to the part-of-speech tagging of Kashmiri. Although some work has been done in this field e.g. Mehdi(2009), this work has largely been in the area of a type of POS tagging which is called Stochastic Tagging. The present paper will attempt a different approach to the POS tagging of Kashmiri based on Kashmiri

* Department of Linguistics, University of Kashmir, India

morphology and syntax. This type of tagger is usually termed as Rule-based tagging.

POS tagging can be done manually by using a software tool by human taggers. It can also be accomplished automatically by means of a computer program, with no or little human intervention. This paper will restrict itself to the discussion of automatic tagging of Kashmiri corpora. The above mentioned computer program can be based on an algorithm whose logic may be Stochastic, Rule-based or a combination of both Stochastic and Rule-based (Hybrid).

Stochastic Taggers assign a tag based on the probability of a given word having a given tag in a given context. e.g. the Hidden Markov Model or HMM Tagger is one such tagging algorithm which chooses the tag sequence which is most probable given the observation sequence of n words ($w_1{}^n$) i.e. out of all sequences of n tags $t_1{}^n$ the single tag sequence such that $P(t_1{}^n \mid w_1{}^n)$ is highest. Thus, this tagging algorithm does not take into account the knowledge of the language. It is entirely based on probability calculation.

Rule-based Taggers are based on rules (morphological, syntactic, etc.) and lexicons devised by linguists. Thus this type of Tagger takes into account knowledge of language to tag a given word. An example of this type of tagger is the EngCC Tagger (Voutilainen, 1995, 1999). It is based on a two-tier architecture. In the first stage of the tagger each word is given all possible parts of speech after running it through a lexicon of about 56,000 entries. Then a large set of constraints (3734) is applied to the input sentence to eliminate incorrect part of speech.

Hybrid Taggers possess features of both the above mentioned types of taggers. The example of a Hybrid system is Transformation-Based Tagging also called Brill tagging. It is an example of the Transformation-Based Learning (TBL) approach to machine learning (Brill, 1995), and draws an inspiration from both the Rule-based and the Stochastic taggers. The Brill's TBL algorithm is based on three major states. First, every word is given its most likely tag. Then every possible transformation is examined and the one that gets the most improved tagging is selected. Then the data is retagged according to this rule. Last two stages are repeated until some stopping criterion is reached such as insufficient improvement over the previous pass. Thus this type of tagger takes both linguistic knowledge and probability into account.

It is in the last two types of taggers (Rule- based and Hybrid) where linguistic knowledge needs to be encoded into a computer program that the collaboration between the knowledge of programming and linguistics is

essential. The knowledge of linguistics is essential because it supplies the logic behind a computer program written to tag a corpus in a Rule-Based way. The knowledge of programming is needed to encode the linguistic rules into a machine comprehensible form. This collaboration will result in a tagging program which will have the ability to distinguish between different word classes.

This paper will attempt to propose a preliminary architecture for a Rule-based tagger of Kashmiri. This tagger will have three components:

1. A lexicon of closed-class words,
2. A morphological component, and
3. A syntactic component.

A string will pass through the three components in the above order and emerge tagged at the end.

Lexicon:

This component will be composed of all the closed-class words such as pronouns (/hɯ/, /su/, /ji/, etc.) auxiliaries (/a:sun/, etc.), post-positions (/peʈʰ/,/niʃ/, etc.) and conjunctions(/beji/, /harga/, etc.).

Morphological Component:

The morphological affixes of a word can be used to identify the part-of-speech of a word. Thus, for example, when a word ends with /-an/ ending, it identifies the word as a noun and thus limits the syntactic and semantic possibilities of the word. Morphological ending of a word can contain more information than that. Thus /-an/ ending above further holds the information that the word can either be a plural object or a singular subject. For example,

/pʰa:ʇan khɔv baʇi/ "Fati ate rice" (singular subject)

/pʰa:ʇan/	/khɔv/	/baʇi/
Fati <third person> <singular> <feminine> <nominative>	eat <present> <feminine> <singular>	rice <accusative>
"Fati ate rice"		

The /-an/ ending identifies the first word not only as a noun but also as singular active subject. This knowledge is absolutely essential to comprehend the sentence. This type of system coupled with a morphological synthesizer is also absolutely essential during the process of language synthesis.

There are many other morphological endings which can help a computer program to identify and differentiate different word classes in Kashmiri. Some of them are listed in the following table:

S.No.	Morphological Ending	Features Identified	Examples
1.	/-ɖ:ar/ "bearer or pertaining to"	Noun, Masculine, Singular	/zə:l-ɖa:r/ "headman" /tə:sil-ɖa:r/ "revenue officer /ɖuka:n-ɖa:r/ "shopkeeper" /ʤama:-ɖ:ar/ "head-sweeper"
2.	/-vo:l/ "pertaining to or related to"	Noun, Masculine, Singular	/duaɖi-vo:l/ " milk man" /sabzi:-vo:l/ "vegetable seller" /ʧa:i-vo:l/ "tea seller" /gevan-vo:l/ "singer" /ɖuvan-vo:l/ "sweeper"
3.	/-en'/ "pertaining to or related to"	Noun, Feminine, Singular	/da:ndr-en/ "vegetable seller or grower" /ka:ndr-en/ "baker"
4.	/-va:ʤen'/ "pertaining to or related to"	Noun, Feminine, Singular	/sabzi-va:ʤen'/ "vegetable seller" /duvan-va:ʤen/ "sweeper" /ga:ɖi-va:ʤen/ "fisherwoman"
5.	/-ba:j/ "pertaining to or related to"	Noun, Feminine, Singular	/gu:rʲ-ba:j/ "milkmaid" /ma:sʈər-ba:j/ "teacher" /ɖa:kʰʈər-ba:j/ "doctor" /raŋrirʲ-ba:j/ "dyer"

6.	/-a:n/	Verb present participle	/kar-a:n/ "doing" /kʰjv-a:n/ "eating" /ɖav-a:n/ "running" /giɳɖ-a:n/ "playing" /par-a:n/ "reading"
7.	/-muʈ/	Verb past participle	/ʃɔŋ-muʈ / "slept" /zu:al-muʈ/ "burnt" /mu:anɖ-muʈ/ "kneaded" /su:av-muʈ/ "made to sleep"
8.	/-pɨ:atʰj/	Adverb	/asɨl-pɨ:atʰj/ "well" /ɨtʰai-pɨ:atʰj/ "like this" /hutʰ-pɨ:atʰj/ "like that" /va:r-pɨ:atʰj/ "thoroughly"
9.	/-as/	Noun, Dative	/baʈ-as/ "to rice" /ha:pʈ-as/ "to a bear" /kalm-as/ "to pen" /ɖarva:z-as/ "to door" /a:b-as/ "to water"

10.	/-na:v/ /-na:na:v/	Verb, Causative	/kar-na:v/ "cause (someone) to do (something)" /kʰja-na:v/ "make someone eat" /vaɖ-na:v/ "make someone weep" /ʃuaŋ-na:na:v/ "cause someone to make someone sleep" /kʰja-na:na:v/ "cause someone to make someone eat"
11.	/-is/	Noun, Dative	/kul-is/ "to a tree" /hu:n-is/ "to a dog" /mə:l-is/ "to father" /ə:l-is/ "to a nest"

Syntactic Component:

Besides morphology, syntax can also be encoded into machine readable rules. Thus, a machine can identify a word class by means of rules based on syntax. For example,

1. A rule can be devised which states that if the word /mˈo:n/ or any other word belonging to the same class of possessive pronouns occurs at the penultimate position of a sentence, the word immediately following must be a noun. For example,

/ji ʧˈu mˈo:n kalam/ "This is my pen."

/ji/	/ʧˈu/	/mˈo:n/	/kalam/
this <singular> <neuter>	to be <present> <masculine> <singular	first person genitive	pen
		possessive pronoun	noun

"This is my pen"

2. Another syntactic rule can be formulated as:
Rule A: In a three token string with /a:sun/ ("to be") as one of the tokens, the other tokens must be a combination of:
Noun – Verb
e.g. /mu:adʒʲ ʧʰe asa:n/ "Mother is laughing"

/mu:adʒʲ/	/ʧʰe/	/asa:n/
mother <third person> <singular> <feminine> <nominative>	to be <present> <feminine> <singular>	laugh <continuous>
"Mother is laughing"		

/gul ʧʰu para:n/ "Gul is studying"

/gul/	/ʧʰu/	/para:n/
Gul <third person> < singular> <masculine> <nominative>	to be <present> <masculine> <singular>	study <continuous>
"Gul is studying"		

/fa:ʈi ʧʰe mərmits/ "Fati is dead"

/fa:ʈi/	/ʧʰe/	/mərmits/
Fati <third person> <singular > <feminine> <nominative>	to be<present> <feminine> <singular>	die <past perfect>
"Fati is dead"		

Pronoun – Verb
e.g. /su ʧʰu ɖu:ara:n/ "He is running"

/su/	/ʧʰu/	/ɖu:ara:n/
he <third person> <singular> <masculine> <nominative>	to be<present> <masculine> <singular>	run <continuous>
"He is running"		

/ʈim ʧʰi tsəlmitʲ/ "They are gone"

/t̪im/	/ʧʰi/	/tsəlmitɹ/
they <third person> <plural> <masculine> <nominative>	to be<present> <plural> <masculine>	go <past perfect>
"They are gone"		

/hua ʧʰi vad̪a:n/"She is crying"

/hua/	/ʧʰi/	/vad̪a:n/
she <third person> <singular> <feminine> <nominative>	to be<present> <singular> <feminine>	cry <continuous>
"She is crying"		

In the above examples, one of the two words other than /ʧʷu/ ("is") or any other word belonging to the same closed-class category will easily be identified as a Verb because of the morphological endings /-a:n/ and various forms of /-mut̪/ or /-mits/. This leaves the other word which is necessarily either a Noun or a Pronoun.

Noun – Adjective

e.g. /na:d̪i: ʧʰe kʰu:bsu:rat̪/ "Nadi is beautiful"

/na:d̪i:/	/ʧʰe/	/kʰu:bsu:rat̪/
Nadi <third person> <singular> <feminine> <nominative>	to be <present> <feminine> <singular>	beautiful <adjective>
"Nadi is beautiful"		

/sua ʧʰe tʰəz/ "She is tall"

/sua /	/ʧʰe/	/tʰəz/
She <third person> <singular> <feminine> <nominative>	to be <present> <feminine> <singular>	tall <adjective>
"She is tall"		

Noun – Noun

e.g. /sulɨ ʧʰu insa:n/ "Sul is a human"

/sulɨ/	/ʧʰu/	/insa:n/
sul <third person>	to be <present>	a human <noun>

<singular> <masculine> <nominative>	<masculine> <singular>	
"Sul is a human"		

/guli ʧʰu kʰar/ "Gul is a donkey"

/guli/	/ʧʰu/	/kʰar/
Gul <third person> <plural> <masculine> <nominative>	to be<present> <singular> <masculine>	a donkey <noun>
"Gul is a donkey"		

Pronoun – Noun

e.g. /ji ʧʰu kul/ "This ia a tree"

/ji/	/ʧʰu/	/kul/
this <singular> <neuter>	to be<present> <singular> <masculine>	a tree <noun>
"This ia a tree"		

Noun – Question word

e.g. /nabi kus ʧʰu/ "who is Nabe?"

/nabi/	/kus/	/ʧʰu/
Nabe <third person> <singular> <masculine> <nominative>	who <singular> <masculine>	to be<present> <singular> <masculine>
"who is Nabe?"		

Pronoun – Question word

e.g. /su kus ʧʰu/ "who is he?"

/su/	/kus/	/ʧʰu/
he <third person> <singular> <masculine> <nominative>	question word <singular> <masculine> <third person>	to be<present> <singular> <masculine>
"who is he?"		

/ji kja: ʧʰu/ "what is this?"

/ji/	/kja:/	/ʧʰu/
this <singular>	question word	to be<present>

<neuter>		<singular> <masculine>
"what is this?"		

Noun – Adverb

e.g. /mansu:r ʧʰu jaṯi/ "Mansoor is here"

/mansu:r/	/ʧʰu/	/jaṯi/
Mansoor <third person> <singular> <masculine> <nominative>	to be<present> <singular> <masculine>	here <adverb>
"Mansoor is here"		

/ṯa:re:k ʧʰu huṯen/ "Tariq is there"

/ṯa:re:k/	/ʧʰu/	/huṯen/
Tariq <third person> <singular> <masculine> <nominative>	to be<present> <singular> <masculine>	there <adverb>

/ṯa:re:k ʧʰu gari/ "Tariq is at home"

/ṯa:re:k/	/ʧʰu/	/gari/
Tariq <third person> <singular> <masculine> <nominative>	to be<present> <singular> <masculine>	at home <adverb>
"Tariq is at home"		

Pronoun – Adverb

e.g. /su ʧʰu jaṯi/ "He is here"

/su/	/ʧʰu/	/jaṯi /
he <third person> <singular> <masculine> <nominative>	to be<present> <singular> <masculine>	here <adverb>
"He is here"		

/sua ʧʰa huṯen/ "She is there"

/sua/	/ʧʰe/	/huṯen/
she <third person> <singular> <feminine> <nominative>	to be<present> <singular> <feminine>	there <adverb>
"She is there"		

Thus the word other than /ʧ͡ʰu/ or any other word belonging to this closed-class category in this type of sentence is a noun, a pronoun, a verb, an adverb or an adjective. All this multiplicity of patterns can be reduced by first applying the closed-class lexicon and the morphological rules on a given string. Question words, pronouns, the auxiliary and some adverbs will be tagged when a given string passes through the closed-class lexicon. This will leave adjectives, nouns and verbs. The verbs and some of the nouns will be tagged in the morphological rule section. This leaves nouns and adjectives with no apparent morphological markers for the syntactic part. Thus, a more specific rule in this regard can be formulated as:

Rule B: If in a three-token string with a form of auxiliary /a:sun/ and an /-a:n/ or /mut̪/ token (which will already be tagged as AUX and VERB by the closed-class lexicon and the morphological rules) the remaining untagged token will be tagged as Noun.
For example, the below string will go through the following stages:

/za:hid̪ ʧ͡ʰu para:n/ **closed class lexicon**

/za:hid̪/	/ ʧ͡ʰu/	/para:n/
	AUX	

/za:hid̪/	/ ʧ͡ʰu/	/para:n/
	AUX	

Morphological rules

/za:hid̪/	/ ʧ͡ʰu/	/para:n/
	AUX	Verb

/za:hid̪/	/ ʧ͡ʰu/	/para:n/
	AUX	Verb

Syntactic Rules

/za:hid̪/	/ ʧ͡ʰa/	/para:n/
Noun	AUX	Verb

3. Another syntactic rule which can be used to disambiguate the syntactic category of a word is: if two consecutive words having any of the endings /-as/ or /-is/ or /-en/ or /-i/, the first word is an adjective and the next one is a noun.
e.g. /krihnis hu:nis/ "to a black dog"

/krihnis/	/ hu:nis/
black <dative>	dog <dative>

"to a black dog"

/asli ko:ri/ "to a good girl"

/asli/	/ko:ri/
good <dative>	girl <dative>
"to a good girl"	

/vazlis puaʃas/ "to a red flower"

/vazlis/	/puaʃas/
good <dative>	girl <dative>
"to a red flower"	

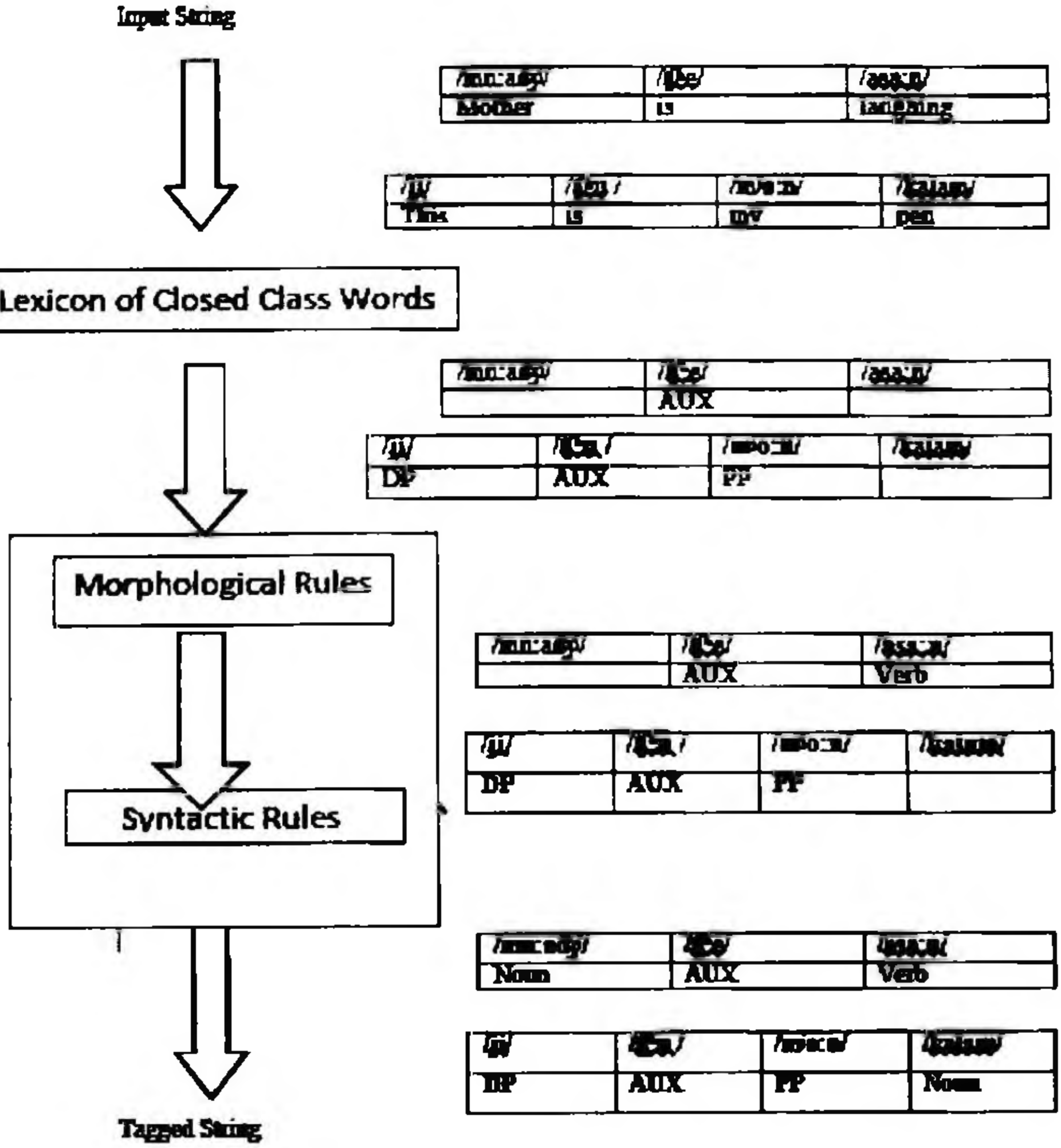
/zə:vilen ʃuren/ "to lean children"

/zə:vilen/	/ʃuren/
lean <dative>	children <dative>
"to lean children"	

In the above four examples the first one is an adjective and the second one a noun as specified by the rule.

Conclusion:

Both the syntactic and morphological rules can be encoded into a program to form a tagger whose proposed architecture is illustrated by the block diagram below. As the block diagram shows, an input string is first passed through an in-built closed-class lexicon of the tagger which tags all the closed-class words in the input string. For example, the words like, /ji/, /ʧʰe/, / ʧʰu/, /mjo:n/ in the example sentences below are tagged as DP (Demonstrative Pronoun), AUX (auxiliary verb), AUX (Auxiliary verb) and PP Possessive Pronoun) respectively because these are all closed-class words. The string then passes through the morphological rules section which tags words according to morphological endings e.g. the words like /asa:n/ which have the morphological ending /-a:n/ are identified as Verbs. The string then goes onto the syntactic rules section which tags according to syntactic rules. Thus, the words /mu:aʤʲ/ and /kalam/ are tagged as Nouns in accordance with the syntactic rules 2 and 1 respectively. Thus, the output of the tagger is a completely tagged string.



References

Antony and Soman. 2013. Part of Speech Tagging for Indian Languages: A Literature Survey. *International Journal of Computer Applications* 34.8 (2011): 22-29.

Brill, Eric. A Simple Rule-Based Part of Speech Tagger. *ANLC '92 Proceedings of the Third Conference on Applied Natural Language Processing*(1992):152-155.

Garside, R., and Smith, N. A Hybrid Grammatical Tagger: CLAWS4. *Corpus Annotation: Linguistic Information from Computer Text Corpora.*

Grishman, Ralph. *Computational Linguistics: An Introduction.* Cambridge: Cambridge University Press, 1986.

Jurafsky, Daniel and James Martin. 2009.*Speech and Language Processing : Computational Linguistics and Speech Recognition.*2nd ed. New Jersey: Pearson Prentice Hall.

Kak , A.A., etal. 2009. What Should Be And What Should Not Be: Developing A POS Tagset For Kashmiri. *Interdisciplinary Journal of Linguistics.Vol.02.Department of Linguistics, University of Kashmir. 287-294.*

Voutilainen, A.1999. An Experiment on the Upper Bound of Interjudge Agreement: The Case of Tagging. *Proceedings of EACL '99*: 204-208.

□□□