Chapter III

Part of Speech Tagset for Kashmiri

3.0 Introduction

Tagset development forms a foundation of any computational processing endeavor. It is generally accepted that, as a prelude to syntactic analysis of natural language by computers, a text must be annotated with tags indicating the POS. The first pre-requisite for automated POS tagging is a tagset that is a set of exhaustive categories into which any token of the language can be placed. While the nature of the language is that there will always be words that are hard to classify, or are ambiguous between two categories,the tagset categories should be designed in such a way so as to minimize such problems. The fundamental problems in POS tagging task stem from the fact that a word can take different lexical categories depending on its context. The tagger has to resolve this ambiguity and determine the best sequence for a sentence. POS tagging also known as morphosyntactic categorization or syntactic word class tagging (Halteren 1999) is the process of assigning a part of speech or other lexical class marker to each word in a corpus. Tags are also applied to punctuation markers, thus tagging for natural language is the same process as tokenization for computer languages, although tags for natural languages are much more ambiguous.

Part of speech tagging has been studied extensively in the past two decades and lot of work has been done in various European languages including many Indian languages like Hindi, Urdu, Sanskrit, Tamil and Kannada. Comparatively speaking, little work has been done in Kashmiri in this field.The only work done in Kashmiri so far in this domain is the development of a flat tagset following

ILMT(Indian Language Machine Translation) guidelines under the project "Development of Language Tools and Linguistic Resources for Kashmiri" at the Department of Linguistics, University of Kashmir.

3.1 General Framework for Kashmiri POS Tagging

For designing a Kashmiri tagset, apart from following the Eagles Guidelinesand thePenn tree bank tagset, manyother Indiantagging guidelines like IL-POST, ILMT and Sanskrit tagset were taken into consideration. The tagging schema for Kashmiri is designed taking into consideration both language features in general and the idiosyncratic features of Kashmiri. After careful consideration a hierarchical tagset was favored. The whole design of the tagset developed so far revolves around three distinct features into which the grammatical schema is distributed. The featuresare:

I. Category

II.Type

III.Attribute

Categories involve major grammatical categories like Nouns, Verbs etc. The type includes the type of those grammatical categories like Common Nounand Proper Noun for Noun category, Main Verb, Auxiliary Verbetc. for Verb category, and so on. The attribute level takes features within each type like Gender (masculine, feminine), Number (singular, plural), Case (dative, ergative, ablative, etc.), Tense and Aspect etc. into consideration. The category list includes all Kashmiri categories that can occur. The type list within a category includes all

types of the category that can occur. The attribute list includes all possible attributes of the type that can occur. A detailed account of the schema is given below:

Category	Types	Attributes
NOUN	Common Noun	Gender, Number, Case, Emphatic[1], Vocative[2]
	Proper Noun	Gender, Number, Case, Emphatic
PRONOUN	Personal Pronoun	Gender, Number, Person, Distance[3],Case, Emphatic, Honorific[4]
	DemonstrativePronoun	Gender, Number, Person, Distance, Case, Emphatic, Honorific
	Reflexive Pronoun	Gender, Number, Case, Emphatic
	Reciprocal Pronoun	Gender, Number, Case, Emphatic
	Relative Pronoun	Gender, Number, Case, Emphatic, Honorific
	Possessive Pronoun	Gender, Number, Person, DistanceCase, Emphatic, Honorific
	Indefinite Pronoun	Gender, Number, Case, Emphatic,
	Interrogative Pronoun	Gender, Number, Case, Honorific
VERB	Main Verb	Gender, Number,Person,Tense, Aspect, Finiteness, Honorific

[1] In the present tagset Emphatic is depicted both as a Type and as an attribute. As a Type it occurs as a free morpheme, and as an attribute it occurs as a bound morpheme as shown above.

[2] Like Emphatic, Vocative is also used as an attribute in as well as a Category.

[3] 3rd Person Pronouns exhibit a two term distinction in terms of distance i.e. Proximate and Remote wherein Remote is further divided into Remote (within sight) and Remote (out of sight).

[4] Honorific can be used as an Attribute as in Noun and Pronoun categories shown above, and also as a Type of the main category where it occurs as a free morpheme.

	Auxiliary[5]	Gender, Number, Person, Tense, Aspect, Finiteness
	Conjunctive Verbs	Gender, Number, Finiteness
	Causative Verbs	Gender, Number,Tense, Aspect,Finiteness, First place, Second Place, Third Place
ADJECTIVE	Qualitative	Gender, Number, Case, Emphatic
	Quantitative	Number, Case, Emphatic
QUANTIFIER	Cardinal	Number, Case, Emphatic
	Ordinal	Number, Case, Emphatic
GERUND	—	—
ADVERB	Time	—
	Place	—
	Manner	—
POSTPOSITION	—	Case[6]
INTENSIFIER	—	—
INTERJECTION	—	—

[5] Auxiliary verbs also include modal thus while tagging all modal will be tagged as auxiliary with its attributes.

[6]Postpositions in Kashmiri take dative and ablative cases with them.

PARTICLE	—	—
VOCATIVE	—	—
EQUATIVE	—	—
HONORIFIC	—	—
NEGATIVE	—	—
EMPHATIC	—	—
MULTI WORD EXPRESSIONS	Compound Words	
	Reduplication [7]	
	Echo Words[8]	
	Part of Word	
CONJUNCTION	Coordinating	—
	Subordinating	
PUNCTUATION	—	—

[7] Attributes for reduplication will be same as that of the first word, that is if the first word is a Common Noun then it will take Common Noun attributes and if it is a Verb then verbal attributes are used.

[8] Attributes for Echo words will be same as that of the first word like that of Reduplication. The first word is tagged according to the category which it possesses and the second word is tagged as an Echo word with 'ECH' tag.

SYMBOL	—	—
UNKNOWN WORDS	—	—

Table 3.1: Categories, Types and their Attributes used in the Proposed Kashmiri POS Tagset

3.2 Attribute Description

The table given below gives a detailed description of attributes which are used in the present tagset for annotating the corpus.

Attribute	Value
GENDER	Masculine(mas) Feminine(fem) Neuter(null)[9]
NUMBER	Singular(sng) Plural(plu) Dual(null)[10]
PERSON	I Person(Ip) II Person(IIp) III Person(IIIp)

[9] Neuter gender is not found in Kashmiri that's why it is kept as null in the present tagset. However its presence in the present tagset is because if the same tagset is used for any other language in which neuter gender is present then the same will be activated while using that particular language in order to maintain the cross linguistic usage of the Tagset.

[10]There are no special dual, trial or quadral forms in Kashmiri for any of the Pronouns and the reason for its presence in the present tagset is same as that of Neuter gender.

DISTANCE	III Person Proximate(prx)	
	III Person Remote	within sight (rmn)
		out of sight(rmf)
CASE	Nominative(nom) Dative(dat) Ergative(erg) Ablative(abl) Genitive(gen)	
TENSE	Present(prt) Past(pst) Future(fut)	
ASPECT	Simple(smp) Progressive(prg) perfective(prf)	
FINITENESS	Finite(fn), Non-finite(nfn)	
EMPHATIC	Emph	
CAUSATIVE ATTRIBUTES	Ipl, IIpl,IIIpl	
HONORIFICITY	honorific(hon)	
VOCATIVE	Voc	

Table 3.2: Attributes and their values

3.3 Tag Structure

Using the above mentioned categories, types and attributes, thefollowing tag structure is used for annotating the corpus.

<tag_cat><tag_types><attribute_list>

where<***tag_cat***> is used for the MainCategory, <***tag_types***>is used for the Types of the Main Category and <***attribute_list***>denotes the sub features of the main category.The present tagset is hierarchical in nature and in order to maintain thathierarchy, the main category tags for types and for attributes are placed between '<' and '>' in order to differentiate between the Main Category, Type and Attribute. Furthermore there should be no white space in between the tags of a single word.

Consider *kaT'an* 'boys-dat'. The tag structure of this token will be

Example	Tag structure
kaT'-an	<N><NC><mas><plu><dat>
'boys-dat'	

Table 3.3: Example of Tag structure

The above given example of Tag structure is explained as

Category	Type	Attribute
Noun(N)	Common Noun(NC)	mas,sng,dat

Table 3.4: Schema of Tag structure

3.4 Tag Description

A detailed description of each tag used in Kashmiri tagset is given below:

3.4.1 Noun

Nouns in Kashmiri are broadly classified as Proper and CommonNouns and this distinction is grammatically significant asProper Nouns count as a different category for ergative andpossessive marking.Common Nouns include person, place or a thing, Abstract Noun include emotions, ideas etc, Collective Noun includes group of things, animals, or persons, Countable and Non-countable Nouns.As Nouns (Common and Proper) inflect for number gender and case, in the present tagging scheme Nouns are tagged by using different tags according to their gender- number distinction. Some relevant examples are given below:

NOUN	Example	Tag structure
Common Noun	*Kul*	<N><NC><mas><sng><nom>
	‘tree’	
	kul'-an	<N><NC><mas><plu><dat>
	‘trees- dat’	
	ko:r-i	<N><NC><fem><sng><abl>
	‘girl-abl’	

	ko:r'-an hund	<N><NC><fem><plu><gen>
	‘girls-gen’	
Proper Noun	*parimehal*	<N><NP><mas><sng>
	‘Parimahal’	
	parimehl-an	<N><NP><mas><plu><dat>
	‘Parimahal-dat’	
	Ifat	<N><NP><fem><sng><nom>
	‘Ifat’	
	ift-an	<N><NP><fem><plu><dat>
	‘Ifat-dat’	

Table 3.5: Tag structure of Nouns

3.4.2 Pronouns

All Pronouns are tagged separately. In Kashmiri Pronouns are inflected for number, gender and case like Nouns but are distinguished from Nouns by having a category of person also. Thus all Pronouns will be tagged with their gender, number and person features. Some examples are given below

PRONOUN	Example	Tag
Personal Pronoun	*bI*	<P><PRP><mas><sng><Ip><nom>
	'I'	
	Tse	<P><PRP><mas><sng><IIp><erg>
	'you-erg'	
	hum-is	<P><PRP><mas><sng><IIIp><rmn><dat>
	'he-dat'	
	yim-an	<P><PRP><mas><sng><IIIp><prx><dat>
	'these-erg'	
	as-i	<P><PRP><mas><plu><Ip><dat>
	'we-dat'	
	yem'	<P><PRP><mas><sng><IIIp><prx><erg>
	'this-erg'	
	hum-av	<P><PRP><mas><plu><IIIp><rmn><erg>
	'they-erg'	

	tim-an	<P><PRP><mas><plu><IIIp><rmf><dat>
	‘they-dat’	
	Me	<P><PRP><mas><sng><Ip><erg>
	‘I-erg’	
	Tse	<P><PRP><mas><sng><IIp><erg>
	‘you-erg’	
	hɔ	<P><PRP><fem><sng><IIIp><rmn><nom>
	‘she’	
	sɔ	<P><PRP><fem><sng><IIIp><rmf><nom>
	‘she’	
	Asi	<P><PRP><mas><plu><Ip><dat>
	‘we-dat’	
	tuh'	<P><PRP><fem><plu><IIp><nom>
	‘you-plu/hon’	
	tim-an	<P><PRP><fem><plu><IIIp><rmf><dat>

	‘they-dat’	
	hum-av	<P><PRP><fem><plu><IIIp><rmn><erg>
	‘they-erg’	

Table 3.6: Tag structure of Personal Pronouns

Pronoun	Example	Tag structure
Demonstrative Pronoun	*hum*'	<P><PDM><mas><sng><IIIp><rmn><erg>
	‘he-erg’	
	Su	<P><PDM><mas><sng><IIIp><rmf><nom>
	‘he’	
	hɔ	<P><PDM><fem><sng><IIIp><rmn><nom>
	‘she’	
	sɔ	<P><PDM><fem><sng><IIIp><rmf><nom>
	‘she’	

Table 3.7: Tag structure of Demonstrative Pronouns

Pronoun	Example	Tag
Reciprocal	*akhəkis*	<P><PRC>

Pronoun	‘to one another’	
	pa:nIvə:n'	
	‘amongst each other’	

Table 3.8: Tag structure of Reciprocal Pronouns

Pronoun	Example	Tag
Reflexive Pronoun	*panun*	<P><PRF><mas><sng><nom>
	‘self’s’	
	panIn'*-an*	<P><PRF><mas><plu><dat>
	‘self’s-dat’	
	panIn-i	<P><PRF><fem><sng><dat>
	‘self’s-dat’	
	panIn'*-av*	<P><PRF><fem><plu><erg>
	‘self’s-erg’	

Table 3.9: Tag structure of Reflexive Pronouns

Pronoun	Example	Tag structure
	m'*ə:n-is*	<P><PPO><mas><sng><Ip><dat>

Possessive **Pronoun**	‘ my-dat’	
	cə:n-is ‘your-dat’	<P><PPO><mas><sng><IIp><dat>
	hum'-sund ‘his-gen’	<P><PPO><mas><sng><IIIp><rmn><gen>
	təm'-sund ‘his-gen’	<P><PPO><mas><sng><IIIp><rmf><gen>
	sa:n'-av ‘our-erg’	<P><PPO><mas><plu><Ip><erg>
	tuhund-i ‘you-abl’	<P><PPO><mas><plu><IIp><abl>
	human-hund ‘their-gen’	<P><PPO><mas><plu><IIIp><rmn><gen>
	timan- hund ‘their-gen’	<P><PPO><mas><plu><IIIp><rmf><gen>

	m'ə:n'	<P><PPO><fem><sng><Ip><nom>
	'my'	
	ca:n'-av	<P><PPO><fem><plu><IIp><erg>
	'your-erg'	
	humi-sInz	<P><PPO><fem><sng><IIIp><rmn><gen>
	'her-gen'	
	yem'-sInz	<P><PPO><mas><sng><IIIp><prx><gen>
	'this-gen'	
	sa:n-i	<P><PPO><mas><plu><Ip><abl>
	'our-abl'	
	tuhInz-av	<P><PPO><fem><plu><IIp><erg>
	'your-erg'	
	yiman-hinzI	<P><PPO><mas><plu><IIIp><prx><gen>
	'their-gen'	

	human-hInzI	<P><PPO><mas><plu><IIIp><rmn><gen>
	'their-gen'	

Table 3.10: Tag structure of Possessive Pronouns

Pronoun	Example	Tag structure
Interrogative Pronoun	*Kus*	<P><PIT><mas><sng><nom>
	'who'	
	kImI	<P><PIT><fem><plu><nom>
	'who'	
	kɔsI	<P><PIT><fem><sng><nom>
	'who'	
	kam-av	<P><PIT><mas><plu><erg>
	'who-erg'	

Table 3.11: Tag structure of Interrogative Pronouns

Pronoun	Example	Tag structure
	Yus	<P><PPR><mas><sng><nom>
	'who'	

Relative Pronoun	*Yim*	<P><PPR><mas><plu><prx><nom>
	‘these’	
	yɔsI	<P><PPR><fem><sng><nom>
	‘who’	
	yimI	<P><PPR><fem><plu><prx><nom>
	‘these’	

Table 3.12: Tag structure of Relative Pronouns

Pronoun	Example	Tag
Indefinite Pronoun	*kẽ h*	<P><PID><nom>
	‘something’	
	kãh	
	‘someone’	

Table 3.13: Tag structure of Indefinite Pronouns

3.4.3 Verbs

Verbal constructions in languages, apart from one word, may be composed of multi word sequences as well. A multi word verb group sequence contains a main verb and one or more auxiliaries (V AUX ...). In the present work after a careful study of Kashmiri peculiarities it was decided that it would be more

appropriate to use different tags for the main verb and the auxiliary verbs. As per the classification of verbs, they are tagged as main verb, auxiliary, causatives, conjunct, etc. in the present tagset. The features include tense, number, gender and person.

Verbs	Example	Tag
Main Verbs	*Pakun* 'to walk'	<V><VM><fut><nfn>
	pokus 'walked'	<V><VM><mas><sng><Ip><pst><prf><fn>
Auxiliary Verbs	*chus* 'is-M'	<V><VAUX><mas><sng><Ip><prt>
	o:sus 'was-M'	<V><VAUX><mas><sng><Ip><pst>
	ə:sIkh 'was-F'	<V><VAUX><fem><sng><IIp><pst>

	chukh	<V><VAUX><mas><sng><IIp><prt>
	'are-M'	
Causative Verbs	*kh'a:vna:vun*	<V><VCUS><IIpl><nfn>
	'make someone to eat'	
	pakna:vun	<V><VCUS><IIpl><nfn>
	'make someone to walk'	
Conjunctive Verb[11]	*salah d'un*	<N><NC><mas><sng><V><VCNJ><nfn>
	'to give advice'	

Table 3.14: Tag structure of Verbs

3.4.4 Gerunds

Gerunds are formed by adding an infinitive suffix *-un* to verbs. Gerunds are tagged as VBG, for example,

[11]Conjunctive Verbs consists of a Noun lexical item and a Verb lexical item. Thus the first part is tagged as a Noun with its type and attributes and the second part is tagged as a Verb with its type and attributes as given above.

Gerunds	Tag
Yun	<VBG>
‘coming’	
gevun	
‘singing’	
pakun	
‘walking’	

Table 3.15: Tag structure of Gerunds

3.4.5 Adjective

Adjectives are broadly classified into qualitative and quantitativeAdjectives. All modifiers of quality like different colors (*vɔzul* ‘red’, *n'u:l* ‘blue’, *saphe:d* ‘white’, etc.), personal qualities (*ca:la:k* ‘clever’,*da:na:* ‘wise’ *buzdil* ‘coward’, etc.), physical qualities (*thod* ‘tall’, *tshoT* ‘short’, *v'oTh* ‘fat’, *zə:v'ul* ‘slim’ etc.), qualities of taste (*modur* ‘sweet’, *tsok* ‘sour’, *T'oTh* ‘bitter’, etc.) fall under Qualitative Adjectives.

Numerals (cardinals, ordinals, fractions, multiplicatives), intensifiers (*kãh* ‘some’, *sə:ri*: ‘all’, *kam* ‘little’), Demonstrative Adjectives (*yu:t* ‘this much’, *t'u:t* ‘that much’) etc. are included in Quantitative Adjectives. According to the description given above, Adjectives are tagged as follows:

Adjective	Example	Tag structure
Qualitative	*asIl*	<JJ><QL>
	‘good’	
Quantitative	*kam*	<JJ><QN>
	‘less’	

Table 3.16: Tag structure of Adjectives

3.4.6 Quantifier

Quantifiers are tagged as cardinals and ordinals.

Quantifier	Example	Tag
Cardinals	*akh*	<QTF><CRD>
	‘one’	
	ək'	<QTF><CRD><erg>
	‘one-erg’	
Ordinals	*əkim*	<QTF><ORD>
	‘first’,	
	doyim-is	<QTF><ORD><dat>
	‘second-dat’	

Table 3.17: Tag structure of Quantifiers

3.4.7 Cases

Cases in Kashmiri are primarily divided into nominative, dative, ergative, ablative and genitive. The word classes like Nouns, Pronouns, Adjectives, Verbs and Postpositions inflect for cases, and thus it was decided to include cases in the attribute list. To include cases in the present schema for tagging, the following attributes were used.

Cases	Example	Tag
Dative	*shi:l-as* 'sheela-dat'	<dat>
Ergative	*shi:l-an* 'sheela-erg'	<erg>
Ablative	*saku:l-I* 'school-abl'	<abl>
Genitive	*shi:l-un* 'shila-gen'	<gen>

Table 3.18: Cases as Attributes

Genetive markers like *hund, hInd', hInz*and *hInzI*(Group III) are tagged as Postposition in the present tagging scheme as they behave as Postpositions. The

genitive markers like *un, uk, Ik*', *Ic* and *Ici* (Group I and II) are tagged as attributes.

Some examples of nouns with and without explicit case markers are given below.

Example	**Gloss**		**Example with cases**	**Gloss**
Kul	'tree-S'	Vs.	*kul-is*	'tree-dat.M.S'
shur'	'children-P'	Vs.	*shur*'*-an*	'children-dat.M.P'
insa:n	'human-S'	Vs.	*insa:n-as*	'human-dat.M.S'
Kagaz	'paper-S'	Vs.	*kagz-an*	'paper-erg.M.P'
Pash	'roof-S'	Vs.	*pash-I*	'roof-abl.M.S'
ko:ri	'girls-P'	Vs.	*ko:r*'*-av*	'girls-erg.F.P'
me:z	'table-S'	Vs.	*me:z-I*	'table-abl.M.S'
Pash	'roof-S'	Vs.	*pash-uk*	roof-gen .M.S

Table 3.19: Examples of Nouns with Cases

3.4.8 Particles

Particle is a function word that is not assignable to any of the traditional grammatical word class (such as Nouns, Conjunctions, etc.). Particles are uninflected and can float within a sentence and can be removed from the sentence without changing the grammaticality of the sentence. Expressions like *ti* 'also', *toti* 'still/even then' are tagged as Particles in this tagging scheme. The tag used for Particles is PRT.

3.4.9 Adverbs

Adverbs are words that modify part of speech like Adjectives and Verbs. Adverbs are mainly divided into three types viz manner, place and time Adverbs. The tag structure for Adverbs is given below.

Adverb	Example	Tag structure
Temporal Adverbs	*yeli* 'when'	<ADV><ATM>
Locative Adverbs	*yeti* 'here'	<ADV><ALO>
Manner Adverbs	*yithpə:Th*' 'in this way'	<ADV><AMN>

Table 3.20: Tag structure of Adverbs

3.4.10 Postpositions

Kashmiri has Postpositions. Apart from regular features, Postpositions in Kashmiri also express certain grammatical functions such as cases like Locative, Instrumental, Sociative and Allative. On the basis of their distribution Postpositions are divided into two groups.

Postpositions governing the dative case

Postpositions governing the ablative case

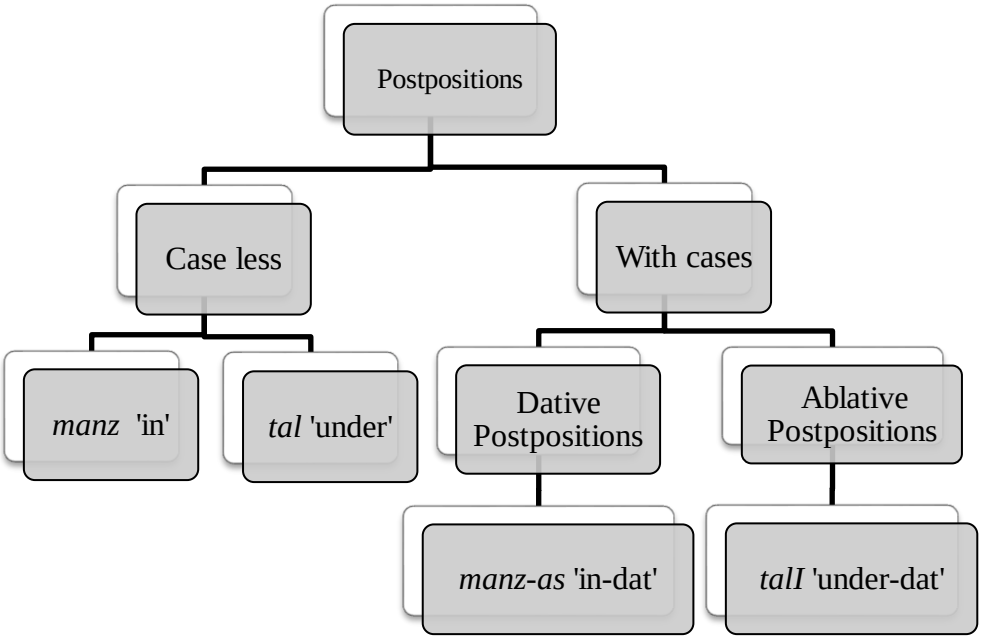

Fig.3.1: Postpositions with their attributes

Postposition	Example	Tag structure
Dative Postposition	*manz-as* 'room-dat'	<PSP><dat>
Ablative Postposition	*manz-I* 'room-abl'	<PSP><abl>

Table 3.21: Tag structure of Postpositions

3.4.11 Conjunctions

Conjunctions are divided into two types subordinating and coordinating Conjunctions. The tag 'CC' with 'SB' and 'CO' as its type tags is used for all Conjuncts, for example,

Conjuncts	Example	Tag Structure
Coordinating Conjunction	*Magar*	<CC><CO>
	‘but’	
	zantI	
	‘as if’	
	hargah	
	‘in case’	
Subordinating Conjunction	*ti*	<CC><SB>
	‘that’	
	zi	
	‘that’	
	ki	
	‘that’	

Table 3.22: Tag structure of Conjunctions

3.4.12 Intensifiers

The expressions intensifying an Adjective or an Adverb is tagged as INTF. For example,

Intensifier	Tag
va:rya:h	
'very'	
	<INTF>
t'u:t	
'very/so much'	

Table 3.23: Tag structure of Intensifiers

3.4.13 Interjections

Interjections will be marked as INJ. As such very few expressions are used for Interjections in Kashmiri, for example,

Interjection	Tag
o:ho:	<INJ>
'oh'	

aha:	
‘ah’	

Table 3.24: Tag structure of Interjections

3.4.14 Negatives

Negation is expressed by the negative Particle *na*. The imperative *mI* ‘don’t’, *matI* ‘don’t’ and the negative conditional *nai* ‘if not’ is also used to express negation. Negative expression like *na* ‘no’, *kihinnI* ‘nothing’, *zIhIn*' ‘never’, etc. will be tagged as negatives and thetag used will be NEG.

Negation	**Tag**
Na	
‘no’	
kihin	<NEG>
‘nothing’	
mI	
‘don’t’	

Table 3.25: Tag structure of Negatives

3.4.15 Question Words

All question words like *kya* ‘what?’ and *k'a:zi*‘ why?’ and Interrogative Adverbs like *kithI pə:Th*' ‘how?’, *kati* ‘where?’etc. are tagged as ‘QW’ in the present tagset.

3.4.16 Equatives

Equatives are used to show comparison between Nouns. To tag equatives ‘EQ’ is used, for example,

Example	Tag
khɔtI	
‘than’	
	<EQ>
Nishi	
‘in comparison of’	

Table 3.26: Tag structure of Equatives

3.4.17 Vocatives

Vocatives behave sometimes as Interjections but they have an independent function as well when used separately, thus it needs a separate tag. For vocatives ‘VOC’ is used as a tag. Vocatives can also be depicted as an attribute when they are attached to a word as a bound morph and the tag used is ‘VOC’.

Vocative Particle	Tag Structure
hey	<VOC><mas>
'hey'	
hat.sa	
'hello'	
haye:	<VOC><fem>
'hey (woman/girl)'	
hatI bin'	
'hey girl'	

Table 3.27: Tag structure of Vocative words

Vocatives	Tag Structure
aslamo:	<N><NP><mas><sng><voc>
'hey Aslam'	
shi:lay	<N><NP><fem><sng><voc>
'hey Sheela'	

Table 3.28: Tag structure of Vocatives as an attributes

3.4.18 Emphatic Words

There are no distinct Emphatic words that modify the Nouns, Pronouns, Verb, Adverbs and Adjectives. Emphatics are either bound morphemes attached to a word itself, or they can be free morphs also as in Emphatic Particles.

Noun	Emphatic marker	Emphatics
bI	*+Iy*	*bIy*
'I'	Emphatic	'I-emph'
shur'	*+i:*	*shuri:*
'children'	Emphatic	'children-emph'
vɔzul	*+ui*	*vɔzlui*
'red'	Emphatic	'red-S.emph'

Table 3.29: Examples of Emphatics

Since Emphatics themselves fall into different grammatical categories, Emphatics are tagged according to the category they posses that is if a word is an Emphatic Pronoun then the word is tagged as a Pronoun with its types and attributes (here Emphatic is used as an attribute). Tag used for Emphatic Particles

is ‘EMPH’ and tag ‘emph’ is used while tagging Emphatics as attributes. Consider the following examples

Emphatic Particle	Tag structure
pa:nay	<EMPH>
‘by self’	

Table 3.30: Tag structure of full word Emphatics

Emphatics	Example	Tag Structure
Emphatic Pronoun	*bIy*	<P><PRP><Ip><mas><sng><nom><emph>
	‘me only’	
Emphatic Noun	*shur'nIy*	<N><NC><mas><plu><dat><emph>
	‘only by the kids’	
Emphatic Adjective	*vɔzl'avIy*	<JJ><QN><mas><plu><erg><emph>
	‘only by the reds’	

Table 3.31: Tag structure of Emphatics as an attribute

3.4.19 Honorific

This tag includes honorific free morphemes (words) like *hez, mahra* (honorific indicators), etc. which exist independently, as opposed to honorific attributes (which are bound) like *tuh*' 'you- honorific'. Tag used is 'HON' for free morphemes and 'hon' for attributes.

3.4.20 Symbol

All special characters like #, @, &, $, %, etc. are tagged as 'SYM'. This tag is similar to the Penn tag set tag 'SYM'.

3.4.21 Punctuation

All punctuation marks like '،''-''؛'':'are tagged as 'PUNC'.

3.4.22 Unknown

The unknown tagis assigned for those words where the category is not known or if there is a doubt regarding the category of a particular word.Tag used is 'UNK'. The tag used here is a temporary tag.

3.4.23 Multi Word Expressions

Multi word expressions include Echo words, Compounding and Reduplication, and Part of Words.

3.4.23.1 Echo Words

Indian languages have a highly productive usage of echo words. Echo words are very frequently used in Kashmiriand such words are tagged as 'ECH'. First part of the Echo Word is tagged according to the category which it possesses and second part of it is tagged as ECH.

Echo words	Tag
ca:y sha:y	<N><NC><fem><sng>
'tea and the like'	<ECH>

Table 3.32: Tag structure of Echo Words

3.4.23.2 Compounding

Compound words are made up of two constituents, each constituent either belongs to a Noun, Verb, Adjective or a Postposition. The compound itself may belong to the category Noun, Adjective or Verb. The compound is categorized on the basis of its head. A compound may be endocentric (when there is a definable head inside the group that has the same distribution as the compound as a whole) or it can be exocentric (i.e., when there is no definable head that relates to the compound as a whole). According to the present scheme of tagging, components of the compound words like Noun compounds, Adjectival compounds, Verbal compounds are treated as individual words and are tagged separately rather than giving a single tag to the whole word sequence (compound). That is, if a word is a compound then both of its components are tagged according to the category they

posses, that is, if a word is an Adjective- Noun compound then the first component is tagged as an Adjective (with its type and attributes) and the second component is tagged as a Noun (with its type and attributes) and if it is part of a Compound Adjective (Adjective-Adjective Compound) then the category 'Adjective' is used for both the components and so on and so forth. This decision was arrived at after great deal of thought and taking into consideration the inadequacy of handling multiple words (of a compound) under one tag (compound tag) without losing some intrinsic individual word (compound components) attributes and properties. Furthermore, it would be much more feasible to handle compounds at the Chunking level.

Some of the Kashmiri compound examples are given as under:

<table>
<tr><th>Compounds</th><th colspan="2">Example</th><th>Tag</th></tr>
<tr><td rowspan="6">Compound Noun</td><td>candI</td><td>tsu:r</td><td rowspan="3"><N><NC><mas><sng><nom><N><NC><mas><sng><nom></td></tr>
<tr><td>pocket</td><td>thief</td></tr>
<tr><td colspan="2">'pickpocket'</td></tr>
<tr><td colspan="3"></td></tr>
<tr><td>o:m</td><td>dɔd</td><td rowspan="3"><JJ><QL><mas><sng>
<N><NC><mas></td></tr>
<tr><td>raw</td><td>milk</td></tr>
<tr><td colspan="2">' raw milk'</td></tr>
<tr><td colspan="4"></td></tr>
<tr><td rowspan="2">Adjectival Compound</td><td>Tsok</td><td>modur</td><td rowspan="2"><JJ><QN><mas><sng><JJ><QN><mas><sng></td></tr>
<tr><td>sour</td><td>sweet</td></tr>
</table>

	‘sour and sweet’		
Verb Compound	*she:run*	*pə:run*	<V><VM><mas><sng><nfn><V><VM><mas><sng><nfn>
	to make better	to decorate	
	‘to decorate’		
Adverb Compound	*ja:yi*	*ja:yi*	<ADV><LOC>
	place	place	
	‘everywhere’		
Pronoun Compound	*kãh*	*kãh*	<P><PID><mas><sng><nom>
	someone	someone	
	‘someone’		

Table 3.33: Tag structure of Compounds

3.4.23.3 Reduplication

When the same word is written twice with or without an internal change in it, and the resulting reduplicated form is grammatically or semantically significant, the process is known as reduplication. For such word combinations category, type and attributes of the first word is written followed by the tag ‘RDP’

Reduplication	Example		Tag
Pronoun Reduplication	*panIn'-an*	*panIn'-an*	<P><PRF><mas><plu><dat> <RDP>
	self-dat	self-dat	
	'of self's'		
Noun Reduplication	*ta:n-as*	*ta:n-as*	<N><NC><mas><sng><dat> <RDP>
	organ-dat	organ-dat	
	'every organ/body part'		
Verb Reduplication	*asa:n*	*asa:n*	<V><VM><mas><sng><prt><prog><fn> <RDP>
	laughing	laughing	
	' while laughing/easily'		
Adverb Reduplication	*jaldi:*	*jaldi:*	<ADV><AMN> <RDP>
	quickly	quickly	
	'very quickly'		
Postposition	*sI:t'*	*sI:t'*	<PSP>

Reduplication	with	with	<RDP>
	‘close/together’		
Adjective Reduplication	*lɔkIT*"	*lɔkIT*"	<JJ><QN><mas><plu><nom> <RDP>
	little	little	
	‘very little’		

Table 3.34: Tag structure of Reduplication

3.4.23.4 Part of Word

Part of word is used for those morphological words which are separated by an orthographic space and thus divided into two tokens as discussed below. For such words POW tag is used.

3.4.23.4.1 Word Boundaries (Token Division)

In Kashmiri, because of orthographic compulsion a single morphological word may be divided into two tokens i.e. a single word can be broken with a white space in between unlike in languages like English, Hindi, etc. In tagging every orthographic white space is considered as a word break even if it occurs within a lexical word.

Example:

Example	Gloss
دٔپیِ زِ	‘you should say’
dəp'zi	
مِلِ ژار	‘cooperation’
milItsa:r	
ہَتہِ بوٗد	‘hundreds’
hatIbod	
کھُلہِ ڈُلہِ	‘spacious’
khulIDulI	
نالہِ موٗت	‘embrace’
nalImot	

Table 3.35: Examples of Token division

Thus, the problem here was how to tag two tokens which make up a single morphological word. After considering various solutions to this problem it was decided to use two separate tags to follow the tagging rules i.e. for such words first part will be tagged according to the category which the whole word (the two word components together) possesses and the second part of the same word will be tagged as POW (part of the word). That means if we have a Noun or an

Adjective which is separated by an orthographic space then the first part of the word will be tagged as <**N><….><…>**or**<JJ><….><…>**and second part of the same word will be tagged as**<POW>**.

POW will also be used for those Nouns which take *-vo:l* as a suffix with them e.g*dɔdIvo:l* ‘milkman’, *sabzi:vo:l* ‘vegetable seller’. Such constructions will be tagged as **‘<N><type><attribute> [SPACE]<POW>’**. Some examples are given as under.

Word	مِلہٕ ژار	نالہٕ موٚت	ریٚتہٕ وادٕ
	milItsa:r	*na:lImot*	*retIva:dI*
Tag	<N><NC><mas><sng> <POW>	<N><NC><mas><sng> POW>	<N><NC><mas><plu> <POW>
Gloss	‘co-operation’	‘embrace’	‘months’
Word	اوٚنٛد پوٚک	بُتھی لاگے	کھُلہٕ ڈلہٕ
	Ondpok	*buth'la:gay*	*khulIDulI*
Tag	<ADV><ALO><POW>	<N><NC><fem><sng> <POW>	<JJ><QN><POW>
Gloss	‘surrounding’	‘defy’	‘spacious’
Word	دودٕ وول	سَبزی وول	ماکان وول

[12] **[SPACE]** represents single white space

	dɔdIvo:l	*sabzi:vo:l*	*maka:nvo:l*
Tag	<N><NC><mas><sng><POW>	N><NC><mas><sng><POW>	N><NC><mas><sng><POW>
Gloss	'milkman'	'vegetable seller'	'Land Lord'

Table 3.36: Tag structure of POW

3.5 Proposed Tagset for Kashmiri

The final tagset which is proposed for Kashmiri is given in a tabular form below

S. No	Category	Tag name
1	Noun	N
1.1	Common Noun	NC
1.2	Proper Noun	NP
2	Pronoun	P
2.1	Personal Pronoun	PRP
2.3	Demonstrative Pronoun	PDM
2.4	Reflexive Pronoun	PRF
2.5	Reciprocal Pronoun	PRC
2.6	Possessive Pronoun	PPO
2.7	Relative Pronoun	PRL

2.8	Indefinite Pronoun	PID
2.9	Interrogative Pronoun	PIT
3	Adjective	JJ
3.1	Qualitative	QL
3.2	Quantitative	QN
4	Gerund	VGB
5	Verb	V
5.1	Main Verb	VM
5.2	Auxiliary Verb	VAUX
5.3	Conjunctive Verbs	VCNJ
5.4	Causative Verbs	VCUS
6	Adverb	ADV
6.1	Adverb of Location	ALO
6.2	Adverb of Time	ATM
6.3	Adverb of Manner	AMN
7	Particle	PRT
8	Postposition	PSP

9	Conjunction	CC
9.1	Coordinating Conjunction	CO
9.2	Subordinating Conjunction	SB
10	Interjection	INJ
11	Emphatic	EMP
12	Honirific	HON
13	Vocative	VOC
14	Question Words	QW
15	Equative	EQT
16	Intensifier	INTF
17	Negative	NEG
18	Quantifier	QTF
18.1	Cardinal	CRD
18.2	Ordinal	ORD
19	Reduplication	RDP
20	Echo Words	ECH
21	Part of word	POW

22	Symbol	SYM
23	Punctuation	PUNC
24	Unknown	UNK

Table 3.37: Proposed Tagset for Kashmiri

3.6 Conclusion

The aim of this tagset is to provide clear instructions for annotating the Kashmiri corpus. The tagset developed so far is hierarchical in nature as it is divided into main word categories, types of the categories and their sub features or attributes as discussed above. This tagset follows the Eagles guidelines, and here also word categories are obligatory, types are recommended and attributes are optional. The need for designing a hierarchical tagset is to capture the morphological richness of the language, and moreover it is effective in reducing confusions like ambiguities and inconsistenciesthat come up while annotating the corpus. A hierarchical design gives greater flexibility, extensibility and re-usability. Since hierarchical tagsets are more elaborative and comprehensive in nature, consequently the same tagset can be used at all levels - POS tagging, chunking, dictionary, morphological analysis and deep parsing. As far as the design of the present Kashmiri tagset is concerned, it is designedin a way where there is linguisticcompatibility with other languages as well. However, the tagset designed here is developed keeping in view the morphosyntactic features of the language thus the tagset here is more compatible with Kashmiri language.However, there are certain feature(attributes) which are not present in

Kashmiri but are included in the above mentioned guidelines taking into consideration the cross linguistic usage of the tagset. In this way we can say the tagset developed here can be used for other languages as well with modifications and additions, wherever required.