Interdisciplinary Journal of Linguistics
Volume [12] 2019, Pp. 62-72

Identification of Part-of-Speech (POS) in Kashmiri: Closed-Class Words and Morphological Markers Strategy

Shahid Yousuf Gilkar*
Nahida Ali*
Prof. Aadil Amin Kak*

Abstract:

Rule-based tagging systems have been developed for different languages such as EngCG tagger for English (Voutilainen 1995, 1999), a tagger for Telugu by Badugu (2014), a tagger for Hindi by Singh et al (2006), a tagger for Turkish by Daybelge and Cicekli (2007), a tagger for Icelandic by Loftsson (2008), a tagger for Pashto by Rabbi et al (2009), a tagger for Arabic by Al-Taani and Al-Rub (2009), a tagger for Hindi by Garg et al (2012), a tagger for Bahasa Indonesia by Rashel et al (2014), a tagger for Marathi by Bagul et al (2014). Most of these systems use syntactic rules, morphological markers and lexicons to identify Part-of-Speech in the corresponding languages. The present paper attempts to identify the closed-class lexical items and lexical items with morphological markers in Kashmiri which can play a crucial role in any rule-based tagging system that may be designed to identity Part-of-Speech (POS) in Kashmiri corpora.

Keywords: Tagger, rule-based systems, closed-class words, morphological endings, lexicon.

1. Introduction

The function words in a language are fixed and are called the closed-class words. These belong to different Parts-of-speech (POS), like, auxiliary verbs, ad positions, conjunctions and pronouns. Kashmiri too contains a comparatively larger group of closed-class words. The nuances in meaning are better captured in Kashmiri as compared to most of the Indo-Aryan languages. Thus Kashmiri has pronouns such as /hu/(ہُ) 'he' (masculine, singular, proximate), /su/ (سُہ) 'he'(masculine, singular, remote), etc.; auxiliaries like /a:sɨ/ /(آسہٕ) 'be' (feminine, past, plural), /tʃʰe/(چھےٚ) 'be' (feminine, present, singular/plural) , /tʃʰum/(چُھم) 'be' (present, singular, neutral, possession), etc. ; post-positions such as /pɛʈʰ/ (پیٹھ) 'on', /niʃ/ (نِش) 'near', etc.; conjunctions such as /beji/ (بیٚیہِ) 'and', /harga/ (ہرگاہ) 'if', etc. Closed-class words can just be listed to form a lexicon which can then be used to tag a corpus.

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/(◌َن) ending, in most of the cases, 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

* University of Kashmir, Srinagar

more information than that. Thus /-an/ (نً) ending further holds the information that the word can either be a plural indirect object or a singular subject. For example:

1./pʰaːt̪**an** kʰʲɔv bat̪ɨ/(فاتن کھیٚو بتہٕ)

فاتن	کھیٚو	بَتہٕ
/pʰa:t̪-**an**/	/kʰʲɔv/	/bat̪ɨ/
Fata (Proper Noun, Third person, Singular, Feminine, Ergative Case)	eat (Verb, Past, Third Person Singular, Masculine)	rice (Noun, Third Person, Singular, Masculine)
"Fata ate rice"		

2. /asi d̪it̪ʲ luk**an** saŋt̪ar/(اَسہِ دِتی لُکن سنٛگتر)

اَسہِ	دِتی	لُکَن	سنٛگتر
/asi/	/d̪it̪ʲ/	/luk-**an**/	/saŋt̪ar/
We (Pronoun, First Person, Plural, Ergative Case)	give (Verb, Past, Third person, Plural, Masculine)	people (Noun, Third Person, Plural, Dative Case)	orange (Noun, Third Person, Plural, Masculine, Common Case)
"We gave people oranges."			

The /-an/(نَ) ending identifies the first word not only as a noun but also as singular active subject in the first sentence above, and as a plural indirect object in the case of the second construction involving a ditransitive verb /d̪it̪ʲ/ (دِتی). This knowledge is absolutely essential to comprehend the sentence.

There are many other morphological endings which can help to identify and differentiate different word classes in Kashmiri. Some of them are listed below:

1. Noun endings (Table 1.1):

S.No.	Morphological Ending	Features Identified	Examples
1.	/-av/ (ٙو)	Noun, Masculine/Feminine, Plural, Ergative Case.	/kʰaːrav/(کھارو) "ironsmiths" /d̪əsʲlav/(دٔسیلو) "masons"
2.	/-as/(ٙس)	Noun, Masculine/ Feminine, Singular, Dative Case	/bat̪-as/(بتَس)"to rice" /ha:pt̪-as/(ہاپتَس)"to a bear"
3.	/-an/(ٙن)	Noun, Masculine/Feminine, Plural, Dative OR Noun, Masculine/Feminine, Singular, Ergative Case	/baːnan/(بانن) "to utensils" /laban/(لَبن) "to walls"

Table 1.1: Noun endings

2. Verb Endings (Table 1.2):

S.No.	Morphological Ending	Features Identified	Examples
1.	/-a:n/(آن)	Verb, Present participle	/kar-a:n/ (کران)“doing” /k^{h}ɛv-a:n/(کھیٖوان) “eating”
2.	/-mut̪/(مُت)	Verb, Masculine, Singular Past participle	/ʃɔŋ-mut̪/ (شۆنٛگمُت) “slept” /zu:al-mut̪/(زولمُت) “burned”
3.	/-na:v/(ناو) /-naː(v) na:v/(ناناو)	Verb, Causative, Imperative	/karnaːv/(کرناو) “cause (someone) to do (something)” /k^{hj}anaːv/(کھیٚاناو) “make someone eat”

Table 1.2: Verb endings

3. Adjective Endings (Table 1.3):

S.No.	Morphological Ending	Features Identified	Examples
1.	/-is/(ِس)	Adjective, Singular, Masculine, Dative Case	/bəɖis neːtʃvis/ “to the elder son” (بٔڈِس نیٚچوِس) /ziːʈ^{h}is puːalas/ “to the long pole” (زیٹھِس پولس)
2.	/-i/(ِ)	Adjective, Singular, Feminine, Dative/ Ergative case	/vazdʒi gaːji/ “to a red cow/ a red cow” (وزجہِ گایہِ) /muatʃi hãːzni/ “to a fat fisherwoman/a fisherwoman” (موچہ ہانٛزنہِ)

Table 1.3: Adjective endings

2. Methodology

Two types of language data were used in this work:

1) The data obtained from the native speakers: The data were transcribed in IPA and in the Perso-Arabic script. They have been used as examples. Each sentence, phrase or word which was used was elicited from at least five native speakers.

2) The written language data collected from various texts belonging to different prose genres (news reports, history, linguistics, fiction, and mythology): Random extracts from different texts were selected and typed into a computer creating fifteen files, which together consisted of over

50,000 tokens. These files were cleaned, normalized and POS-tagged manually using the Bureau of Indian Standards (BIS) tagset. This data was analyzed manually to test the efficacy of the Closed Class Lexicon and the morphological endings in isolation from each other and in collaboration with each other.

3. Analysis

3.1The Closed Class Lexicon

In order to estimate the efficacy of the Closed Class Lexicon in tagging a corpus of Kashmiri texts the counts of all the closed classes were taken from each file. The punctuation marks and symbols were also counted in each file. The counts so obtained in each file were tabulated (Table 1.4):

	V AUX	**PR**	**DM**	**CC**	**PSP**	**RP**	**NST**	**Q-Word**	**QT**	**INTF**	**Total**	**PUNC**	**SYM**	**Total**	**Token Count (TC)**	**TC without P & S**
1	352	323	102	273	399	117	33	12	98	26	1735	301	16	2052	4293	3976
2	185	256	54	170	188	128	125	34	67	17	1224	395	3	1622	3170	2772
3	204	268	59	166	186	98	156	27	40	18	1222	331	0	1553	3083	2752
13	300	262	167	217	327	113	4	28	105	24	1547	273	0	1820	3823	3550
14	188	88	106	147	196	24	24	5	48	5	831	173	32	1036	2253	2048
15	104	174	32	84	123	71	30	23	37	11	689	279	0	968	1995	1716
	3759	4101	1395	2805	3823	1608	1167	501	1057	252	20468	5532	234	26234	52780	47014
	1223 (VM)										1223					
Sums	**4982**	**4101**	**1395**	**2805**	**3823**	**1608**	**1167**	**501**	**1057**	**252**	**21691**	**5532**	**234**	**27457**	**52780**	**47014**

Table 1.4: Closed Class (CCL) Count (due to space constraints whole table could not be included)

The counts of each closed class category for the entire corpus can be found at the bottoms of their respective columns which are summed up with and without the counts of punctuation marks and symbols at the bottoms of the 'Total' columns. The number in the VM row is the number of main verb tokens identical with the auxiliaries in the corpus thus falling within the purview of the Closed Class lexicon. Thus the total numbers of tokens that come under the purview of the Closed Class Lexicon (Coverage of the Lexicon) with and without the punctuation marks and symbols are obtained.

3.1.1 Issues

In order to estimate the accuracy with which the Closed Class Lexicon will tag the portion of corpus that falls under its coverage three main issues had to be noted:

1. The fact that the tokens falling in the category of auxiliaries (VAUX) act as

main verbs when no other verb is present in a string (sentence). It will, thus, be wrongly tagged as auxiliary (VAUX) by the Closed Class lexicon. To deal with this issue, the number of times this occurs was obtained from the corpus and treated as the wrongly tagged instances.

2. The fact that demonstratives (DM) are identical with pronouns (PR) -- mainly third person pronouns. For example, /hu/(ہُ) in the phrase /hu lədki/(ہُ لَڈکہٕ) acts as a demonstrative. This issue was handled by simply subsuming the category of demonstratives under the category of pronouns, because the number of demonstratives is considerably less than that of pronouns as is observed in the table above. The number of demonstratives was then treated as wrongly tagged.

3. The fact that many adverbs of time and place (NST) can also be used as postpositions (PSP). For example, /paṯɨ/ (پتہٕ) acts as a postposition in the phrase /tamipatɨ/(تمی پتہٕ). This issue was handled by subsuming the category of adverbs of time and place under the category of postpositions, because the count of postposition is considerably larger than the count of adverbs of time and place. The number of NST's was then treated as wrongly tagged.

3.1.2 Estimated Accuracy of the Closed-class Lexicon

The estimated result of the Closed Class Lexicon is depicted in the table. (Table 1.5)

	VAUX	PR	CC	PSP	RP	Q-Word	QT	INTF	Total	PUNC	SYM	Total
Number of Correctly Tagged Tokens	3759	4101	2805	3823	1608	501	1057	252	17906	5532	234	23672
Number of wrongly Tagged Tokens	1223 (VM)	1395 (DM)		1167 (NST)								
Coverage	4982	5496	2805	4990	1608	501	1057	252	21691	5532	234	27457
Accuracy	75.452	74.618	100	76.613	100	100	100	100	82.550	100	100	86.215

Table 1.5: Closed Class Lexicon (CCL) Accuracy Estimate

The estimated accuracy is simply the percentage of tokens that will be correctly tagged out of the total number of tokens that will be tagged. The accuracies so obtained with and without taking punctuation marks and symbols into consideration were calculated with the assumption that each token will receive a single tag. If the Closed Class Lexicon is allowed to assign multiple tags to handle the overlap between main verbs (VM) and auxiliaries (VAUX), pronouns (PR) and demonstratives (DM), and postpositions (PSP) and adverbs of time and place (NST), almost every token will have the right tag assigned to it, but sometimes in company with a wrong tag or two. And if the tokens with right tags assigned along with a wrong tag or two are treated as correctly tagged the accuracy will approach 100%.

3.2 Morphological endings

To estimate the efficacy of the morphological endings, 73 morphological endings were selected. For each morphological ending the counts of tokens belonging to each Part-of-Speech category were manually taken from the corpus and tabulated. (Table 1.6)

In order to decide which ending should be used to assign which Part-of-Speech tag, the counts of each ending for each open class part of speech category were observed and the tag of the part of speech category which has the highest count for an ending was selected as the one which that particular ending should assign. If the counts for two or more POS categories are the same and higher than all the other counts, the tag of any of these can be selected as the one the ending should assign. The option of multiple tag assignment (i.e. a single ending assigning more than one tag) was taken into consideration due to some endings whose counts for two or more Part-of-Speech (POS) are equal or the counts for one or more Part-of-Speech (POS) categories are not less than the one third of the highest count. After deciding which ending assigns which Part-of-Speech (POS) tag, the count of an ending for the corresponding Part-of-Speech category was

	Endings	**NN**	**POS NN**	**NNP**	**PR**	**POS PR**	**VM**	**VAUX**	**JJ**	**QT**	**Q Word**	**RB**	**RP**	**CC**	**PSP**	**INTF**	**Totals**
1	◌َس	840		212	194		50	14	18	6	20	65			6	3	1428
2	◌ِس	232	8	24	280	58	33	15	46	87	4	4			126	4	921
3	◌َن	873		283	163	1	101	45	14	48	5	32		65	32	2	1664
4	◌َو	92		2	7		14		1		1	4					121
71	مس	42		28			10		3			6					89
72	نس	142		24	63		137	14									380
73	نم						8										8
	Totals	**6319**	**176**	**1116**	**2552**	**385**	**5708**	**1128**	**516**	**514**	**233**	**923**	**913**	**2054**	**906**	**90**	**23533**

Table 1.6: Morphological endings Count (due to space constraints whole table could not be included) treated as the number of tokens that will be correctly tagged.

3.2.1. Estimated Accuracy of the Morphological endings

Four kinds of estimated accuracies were calculated for each individual ending:

1) Accuracy over the sum of open class category counts along with the sum of closed class category counts with multiple tag assignment.

2) Accuracy over the sum of open class category counts along with the sum of closed class category counts with single tag assignment.

3) Accuracy over the sum of open class category counts without the sum of closed class category counts with multiple tag assignment.

4) Accuracy over the sum of open class category counts without the sum of closed class category counts with single tag assignment

The accuracy is calculated by dividing the sum of tokens that will be correctly tagged by the sum of all the tokens that the ending will tag in each of the four cases mentioned above. The following tables (Tables 1.7, 1.8, 1.9) will make the above mentioned clear.

3.3 Overall Estimated Accuracy

The closed class lexicon and the morphological endings will work serially in the order in which they have been mentioned. The combined accuracy of these two components both in the case of the single tag assignment and the multiple tag assignment can simply be obtained by calculating the percentage of the total number of correctly tagged tokens out of the total number of tokens tagged (Coverage). The totals are obtained by adding the totals of closed-class lexicon and the morphological endings. Because all the closed class tokens will be dealt with by the closed class lexicon, only the total excluding the count of closed class tokens from the morphological-endings-accuracy table (Table 1.7) is used in calculating the overall accuracy of the two components. Moreover, the accuracies have been calculated with and without taking the counts of the punctuation marks and symbols into consideration.

3.3.1 Overall Estimated Accuracy with all the morphological endings

3.3.1.1. Overall estimated accuracy with single tag assignment (Table 1.8).

3.3.1.2. Overall estimated accuracy with multiple tag assignment (Table 1.9).

S.No.	Endings	Tags Per Token	Tag/s	Correctly Tagged Tokens		Accuracy With CCL		Accuracy without CCL		Total Number of Tagged Tokens	Total Number of Tags in MTA	Total No. of Tagged Tokens without CCL	Total No. of Tags without CCL in MTA
				Multiple Tags	Single Tag	Multiple Tags	Single Tag	Multiple Tags	Single Tag				
1.	◌َس	1	NN	1052	1052	73.669	73.669	88.776	88.776	1428	1428	1185	1185
2.	◌ِس	1	NN	264	264	28.664	28.664	76.081	76.081	921	921	347	347
3.	◌َن	1	NN	1156	1156	69.471	69.471	88.718	88.718	1664	1664	1303	1303
4.	◌َو	1	NN	94	94	77.685	77.685	83.185	83.185	121	121	113	113
71.	مس	1	NN	70	70	78.652	78.652	78.652	78.652	89	89	89	89
72.	نس	2	NN, VM	303	166	79.737	43.684	100	54.785	380	760	303	606
73.	نم	1	VM	8	8	100	100	100	100	8	8	8	8
				Sums						**Sums**			
				12634	**11101**					**23533**	**35720**	**14758**	**20861**

Table 1.7: Estimated Accuracy of Morphological endings (due to space constraints whole table could not be included)

	Coverage without PUNC & SYM	**Coverage with PUNC & SYM**	**Number of Correctly Tagged Tokens without PUNC & SYM**	**Number of Correctly Tagged Tokens with PUNC & SYM**	**Accuracy without PUNC & SYM**	**Accuracy with PUNC & SYM**
Closed Class Lexicon	21691	27457	17906	23672	82.550	86.215
Endings Module	14758	14758	11101	11101	75.220	75.220
Lexicon plus Endings	36449	42215	29007	34773	79.582	82.371

Table 1.8: Overall estimated accuracy with single tag assignment

	Coverage without PUNC & SYM	**Coverage with PUNC & SYM**	**Number of Correctly Tagged Tokens without PUNC & SYM**	**Number of Correctly Tagged Tokens with PUNC & SYM**	**Accuracy without PUNC & SYM**	**Accuracy with PUNC & SYM**	**No. of Tags Assigned without PUNC & SYM**	**No. of Tags Assigned with PUNC & SYM**	**Tags/Token without PUNC & SYM**	**Tags/Token with PUNC & SYM**
Closed Class Lexicon	21691	27457	21691	27457	100	100	37159	42925	1.7	1.56
Endings Module	14758	14758	12634	12634	85.607	85.607	20861	20861	1.4	1.4
Lexicon plus Endings	36449	42215	34325	40091	94.17	94.969	58020	63786	1.6	1.5

Table 1.9: Overall estimated accuracy with multiple tag assignment

In calculating the overall accuracy in the case of multiple tag assignment, a token tagged with a correct tag along with one or more incorrect tags is treated as correctly tagged; and the value for “tags assigned per token” is obtained by dividing the total number of tags assigned by the total number of tokens tagged. No change is observed in the values calculated for the morphological endings in the above two tables (Table 1.8, 1.9) with the inclusion or exclusion of the counts of punctuation marks and symbols, because they are completely outside the scope of the morphological endings.

In the case of single tag assignment, the overall accuracy of the Closed Class Lexicon along with the morphological endings without punctuation marks and symbols is 79.582%, which rises to 82.371% when the punctuation marks and symbols are included.

The overall accuracy increases to 94.17% in the case of multiple tag assignment with 1.6 tags assigned per token when the punctuation marks and symbols are excluded. The accuracy increases to 94.969% with about 1.5 tags assigned per token if the punctuation marks and symbols are included.

4. Conclusion

This paper explores the role of the closed class lexicon and morphological endings or suffixes in annotating Kashmiri corpora with Part-of-Speech information.

Following results have been obtained from the analysis:

1) The closed class lexicon's estimated accuracy is over 80% in the case of single tag assignment and approaches 100% in the case of multiple tag assignment with a coverage of 21,691 tokens without punctuation marks and symbols, and 27,457 tokens with punctuation marks and symbols. The morphological endings achieve a dismal accuracy of 47.172% in single tag assignment with all the 73 endings when the lexicon does not feed it. It's estimated accuracy improves slightly to 53.686% in multiple tag assignment with all the 73 endings without the lexicon feeding into it. In both the above cases it has a coverage of 23,533 tokens. In multiple tag assignment the morphological endings assign less number of tags per token when acting in isolation (about 1.5 tags per token) than the lexicon (1.7 tags/token without punctuation marks and symbols, and about 1.56 tags/token with punctuation marks and symbols). The accuracy of the morphological endings in isolation is lower than that of the lexicon even when a selected group of most accurate endings is employed (about 73%) and coverage is only 7318.

2) The morphological endings become more accurate when the closed class tokens covered by the morphological endings are not considered when calculating their accuracy. The accuracy improves from 47.172% to 75.220% in single tag assignment, and from 53.686% (1.5 tags/token) to 85.607% (1.4 tags/token) in multiple tag assignment, when all the 73 endings are used. The coverage, however, decreases from 23,533 tokens to 14,758 tokens. The accuracy when only the seven selected endings are used is improved even further to 89.145%, but with a coverage of only 5,988 tokens. This is the accuracy with which the morphological endings will work when acting down the line from the closed class lexicon. This indicates the importance of the closed class lexicon in tagging in Kashmiri corpus.

3) In the case of single tag assignment, the overall estimated accuracy of the closed class lexicon along with the morphological endings (with all the 73 endings) without punctuation marks and symbols is 79.582%, which rises to 82.371% when the punctuation marks and symbols are included. The overall estimated accuracy increases to 94.17% in the case of multiple tag assignment with 1.6 tags assigned per token when the punctuation marks and symbols are excluded. The accuracy increases to 94.969% with about 1.5 tags assigned per token if the punctuation marks and symbols are included.

References:

Al-Taani, T Ahmad and Salah Abu Al-Rub. "A rule-based approach for tagging non-vocalized Arabic words." *Int. Arab J. Inf. Technol*. Vol. 6, no.3, 2009, pp. 320-328.

Badugu, Srinivasu. "Morphology Based POS Tagging on Telugu." *Proceedings of International Journal of Computer Science Issues (IJCSI). Vol.* 11, no.1, 2014, p. 181.

Bagul, P.Mishra.et al. "Rule Based POS Tagger for Marathi Text. Proc." *Int. J. Comput. Sci. Inf. Technol. (IJCSIT),* vol. 5, no.2, 2014, pp. 1322-1326.

Bhaskaran, S.Bali. "A Common Parts-of-Speech Tagset Framework for Indian languages." *Proceeding of 6th Language Resources and Evaluation Conference (LREC),* vol. 8, 2008.

Daybelge, Turhan and Ilyas Cicekli. "A rule-based morphological disambiguator for Turkish." *Proceedings of Recent Advances in Natural Language Processing*, 2008.

Garg, Navneet, Vishal Goyal, and Suman Preet. "Rule Based Hindi Part of Speech Tagger." *Proceedings of COLING*, 2012, pp. 163-174.

Loftsson, Hrafn. "Tagging Icelandic text: A linguistic rule-based approach." *Nordic Journal of Linguistics*, vol. 31, no. 1, 2008, pp. 7-72.

Rabbi, Ihsan, A. M. Khan, and Rahman Ali. "Rule-based part of speech tagging for Pashto language." *Conference on Language and Technology, Lahore, Pakistan*, 2009.

Rashel, F. Luthfi, A.Dinakaramani and A Manurung, R. "Building an Indonesian rule-based part-of-speech tagger." *Asian Language Processing (IALP), International Conference,* 2014, pp. 70-73.

Singh, S. Gupta, K, Shrivastava, M, & Bhattacharyya, "Morphological richness offsets resource demand-experiences in constructing a POS tagger for Hindi." *Proceedings of the COLING/ACL on Main conference poster sessions,* 2006, pp. 779-786.

Voutilainen, Atro. "A syntax-based part-of-speech analyser." *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics,* 1995, pp. 157-164.

Voutilainen, Atro. *Hand-crafted rules. Syntactic wordclass tagging*. Netherlands, Springer, 1999.

Appendix:

Tag labels used in analysis:

S. No.	Category	Label
1.	Common Noun	NN
2.	Proper Noun	NNP
3.	Possessive Noun	POS NN
4.	Pronoun	PR
5.	Possessive Pronoun	POS PR
6.	Main Verb	VM
7.	Auxiliary Verb	VAUX
8.	Adjective	JJ
9	Adverb of Manner	RB
10.	Adverb of Time and Place	NST
11.	Postposition	PSP
12	Conjunction	CC
13.	Particles	RP
14.	Intensifier	INTF
15.	Quantifier	QT
16.	Question Word	Q Word
17.	Punctuation	PUNC
18.	Symbol	SYM

These are based on the BIS Tagset which has been prepared for the Indian languages by the POS Tag Standardization Committee of Department of Information Technology (DIT), New Delhi, India. A couple of new labels (POS NN and POS PR) have been used and some labels which are included under other labels in the original BIS tagset have been treated separately. For example, Intensifiers (INTF) has been considered separately from Particles (RP). In the case of Question Words, not only a new label is used (Q Word) but it has also been treated separately from Pronouns (PR) and Demonstratives (DM). Furthermore, Adverbs of Time and Place (NST) have been considered separately from the category of Nouns (NN).