



DEPARTMENT OF COMPUTER SCIENCE
SCHOOL OF PHYSICAL SCIENCES
ASSAM UNIVERSITY SILCHAR
(A CENTRAL UNIVERSITY CONSTITUTED
UNDER ACT XIII OF 1989)
Silchar-788011, Assam, India

Date:

DECLARATION

I, Aadil Ahmad Lawaye, bearing Registration No. Ph. D/ 1782/11 Dated 22.09.2011, hereby declare that the subject matter of the thesis entitled “Design and Implementation of Part of Speech Tagger for Kashmiri” is the record of works done by me and that the contents of the thesis did not form the basis for award of any other degree to me or to anybody else to the best of my knowledge. The thesis has not been submitted in any other University / Institute.

This thesis is being submitted to Assam University for the degree of Doctor of Philosophy in Computer Science.

(Aadil Ahmad Lawaye)
Research Scholar

Place:

Date:

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my supervisor, guide and mentor Prof. Bipul Syam Purkayastha for the continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better supervisor and mentor for my Ph.D study. I am also thankful to other faculty members and non-teaching staff of the Department of Computer Sciences for their support and cooperation.

I convey my sincere gratitude to Prof. Pushpak Bhattacharyya of IIT Bombay and Prof. Aadil Amin Kak of Kashmir University for their precious help and encouragement.

Credit goes to my family members especially to my beloved Aba, Naana, Daddy and Baiji for constant support and encouragement. I am also thankful to Didi and Rani di for believing in me and nudging me on. I should also mention Sadaf, Choti, Aifa, Insha, Uzma, Qamran, Arfat and Shailah for always being there for me.

I am extremely thankful to my friend's Aabid, Aashiq, Nazima, Sehar, Mansoor, Aijaz, Huma, Samee, Imran and many others whose names might have escaped me this time for their love and support.

And, last and definitely not the least, it was the Grace of the Almighty that made it possible, and many many thanks to Him.

Date:

(AADIL AHMAD LAWAYE)

CONTENTS

<i>List of Tables and Figures.....</i>	<i>i</i>
<i>Kashmiri Consonants and Vowels.....</i>	<i>v</i>
<i>List of Abbreviations.....</i>	<i>vi</i>

Chapters

1.	Introduction	1
	1.0 Introduction.....	1
	1.1 Tagset.....	5
	1.2 Types of Tagset	6
	1.3 Part of Speech Tagger	8
	1.4 Motivation	12
	1.5 Objectives	13
	1.6 Methodology	13
	1.7 Main contributions.....	14
	1.8 Thesis Outline.....	14
	1.9 Conclusion.....	18
2.	Part of Speech Tagset	20
	2.0 Introduction.....	20
	2.1 English Tagsets	21
	2.2 Claws1 Tagset	22
	2.3 The TOSCA Scheme	24
	2.4 The ICE Tagset	25
	2.5 The Pen Tagset.....	25
	2.6 Eagles Guidelines	26
	2.7 Indian Tagsets.....	27
	2.8 Conclusion.....	34

3.	Part of Speech Tagset for Kashmiri	36
	3.0 Introduction.....	36
	3.1 General Framework for Kashmiri POS Tagging.....	37
	3.2 Attribute Description.....	41
	3.3 Tag Structure	43
	3.4 Tag Description.....	44
	3.5 Proposed Tagset for Kashmiri.....	77
	3.6 Conclusion.....	80
4.	Part of Speech Tagger for Kashmiri	82
	4.0 Introduction.....	82
	4.1 Hidden Markov Model.....	83
	4.2 Our Approach.....	89
	4.3 Experiments	98
	4.4 Conclusion.....	103
5.	Conclusion.....	105
	Bibliography	108
	List of Publications.....	117

List of Tables and Figures

List of Tables

Table 1.1	Accuracy of Kashmiri POS Tagger	17
Table 3.1	Categories, Types and their Attributes used in the proposed Kashmiri POS tagset	41
Table 3.2	Attributes and their values	42
Table 3.3	Example of Tag structure	43
Table 3.4	Schema of Tag structure	43
Table 3.5	Tag structure of Nouns	45
Table 3.6	Tag structure of Personal Pronouns	48
Table 3.7	Tag structure of Demonstrative Pronouns	48
Table 3.8	Tag structure of Reciprocal Pronouns	49
Table 3.9	Tag structure of Reflexive Pronouns	49
Table 3.10	Tag structure of Possessive Pronouns	52
Table 3.11	Tag structure of Interrogative Pronouns	52
Table 3.12	Tag structure of Relative Pronouns	53

Table 3.13	Tag structure of Indefinite Pronouns	53
Table 3.14	Tag structure of Verbs	55
Table 3.15	Tag structure of Gerunds	56
Table 3.16	Tag structure of Adjectives	57
Table 3.17	Tag structure of Quantifiers	57
Table 3.18	Cases as Attributes	58
Table 3.19	Examples of Nouns with Cases	59
Table 3.20	Tag structure of Adverbs	60
Table 3.21	Tag structure of Postpositions	61
Table 3.22	Tag structure of Conjunctions	62
Table 3.23	Tag structure of Intensifiers	63
Table 3.24	Tag structure of Interjections	64
Table 3.25	Tag structure of Negatives	64
Table 3.26	Tag structure of Equatives	65
Table 3.27	Tag structure of Vocative words	66
Table 3.28	Tag structure of Vocatives as an attributes	66

Table 3.29	Examples of Emphatics	67
Table 3.30	Tag structure of full word Emphatics	68
Table 3.31	Tag structure of Emphatics as an attribute	68
Table 3.32	Tag structure of Echo Words	70
Table 3.33	Tag structure of Compounds	72
Table 3.34	Tag structure of Reduplication	74
Table 3.35	Examples of Token division	75
Table 3.36	Tag structure of POW	77
Table 3.37	Proposed Tagset for Kashmiri	80
Table 4.1	Accuracy of Kashmiri POS tagger.	102

List of Figures

Figure 1.1	Tag structure of Flat Tagset	7
Figure 1.2	Tag structure of Hierarchal Tagset	8
Figure 1.3	Kashmiri Corpus: POS learning curve	17
Figure 3.1	Postpositions with their attributes	61
Figure 4.1	Modified Hidden Markov Model	87
Figure 4.2	The HMM based POS tagging architecture.	90
Figure 4.3	Optimization process of TnT using Kashmiri Corpus	101
Figure 4.4	Kashmiri Corpus: POS learning curve.	103
Figure 5.1	Schematic diagram of Tagging process.	107

Kashmiri Consonants and Vowels

Manner of Articulation	State of Glottis	Place of Articulation							
		Bilabials	Labio-dentals	Dentals	Alveolars	Palato-alveolars	Palatals	Velars	Glottals
Stops	voiceless	p, ph		t, th	T, Th			k, kh	
	voiced	b		d	D			g	
Fricatives	voiceless				s	s			h
	voiced		v		z				
Affricates	voiceless				ts, tsh	c, ch			
	voiced					j			
Nasals	voiced	m			n			N	
Laterals	voiced				l				
Trills	voiced				r				
Semivowels	voiced	w					y		

Chart of Kashmiri Consonants

		Front	Central	Back
High	Short	i	ɪ	u
	Long	i:	ɪ:	u:
Mid	Short	e	ə	o
	Long	e:	ə:	o:
Low	Short	ɛ	a	ɔ
	Long	ɛ:	a:	ɔ:

Chart of Kashmiri Vowels

List of Abbreviations

M	Masculine
F	Feminine
S	Singular
P	Plural
nom	Nominative
dat	Dative
abl	Ablative
erg	Ergative
gen	Genitive
1p	First person
2p	Second person
3p	Third person
prt	Present
pst	Past
fut	Future
prog	Progressive
POS	Part-of-Speech
ILMT	Indian Language Machine Translation
IL-POST	Indian Language Part of Speech Tagging
TnT	Trigrams'n'Tags
TBL	Transformation Based Learning
HMM	Hidden Markov Model
BNC	British National Corpus
TOSCA	Tools for Syntactic Corpus Analysis