Part-of-Speech Annotation Challenges in Marathi

Gajanan Rane, Nilesh Joshi, Geetanjali Rane, Hanumant Redkar,
Malhar Kulkarni and Pushpak Bhattacharyya

Center For Indian Language Technology (CFILT)
Indian Institute of Technology Bombay

Presenter: Prof. Malhar Kulkarni, IIT Bombay

at 5th WILD^RE collocated with LREC 2020 – 24th May 2020

Outline

- Introduction
- Marathi Annotated Corpora
- Marathi POS Tag-set
- Lexical and Functional POS Tagging: Challenges and Discussions
- POS Ambiguity: Challenges and Discussions
- Some Special Cases: Challenges and Discussions
- Summary

Introduction

- Parts-Of-Speech (POS) annotation is the process of marking/annotating a word in a text/corpus which corresponds to a particular POS.

- The annotation is done based on its definition and its context, i.e., its relationship with adjacent and related words in a phrase, sentence, or paragraph.

- POS annotation is a standard low-level text pre-processing step before moving to higher levels in the NLP pipeline like chunking, dependency parsing, etc.

- Identification of the POS such as nouns, verbs, adjectives, adverbs for each word the sentence helps in analyzing the role of each word in a sentence.

- Marathi POS tagging was part of an Indian Languages Corpora Initiative (ILCI) project executed at IIT Bombay.

Marathi POS Tagset

- The Bureau of Indian Standards (BIS) has come up with a standard set of tags for annotating data for Indian languages.

- The BIS tag-set aims to ensure standardization in the POS tagging across the Indian languages.

- The tag sets of all Indian languages have been drafted by MeitY and presented as Unified POS standard in Indian languages.

- Marathi POS tag-set has been prepared at IIT Bombay referring to the standard BIS POS Tag-set, IIIT Hyderabad guideline document and Konkani POS Tag-set.

Marathi POS Tagset

Sr. No	Category	Label	Annotation Convention	Examples
1	**Noun (नाम)**	**N**	**N**	
1.1	Common (जातीवाचक नाम)	NN	N_NN	**गाय**\N_NN **गोठ्यात**\N_NN राहते.
1.2	Proper (व्यक्तीवाचक नाम)	NNP	N_NNP	**रामाने**\N_NNP **रावणाला**\N_NNP मारले.
1.3	Nloc (स्थल-काल)	NST	N_NST	1. तो **येथे**\N_NST काम करत होता. 2. त्याने ही वस्तू **खाली**\N_NST ठेवली आहे.
2	**Pronoun (सर्वनाम)**	**PR**	**PR**	
2.1	Personal (पुरुष वाचक)	PRP	PR_PRP	**मी**\PR_PRP येतो.
2.2	Reflexive (आत्म वाचक)	PRF	PR_PRF	मी **स्वतः**\PR_PRF आलो.
2.3	Relative (संबंधी)	PRL	PR_PRL	**ज्याने**\PR_PRL हे सांगितले त्याने हे काम केले पाहिजे.
2.4	Reciprocal (पारस्परिक)	PRC	PR_PRC	**परस्पर**
2.5	Wh-word (प्रश्नार्थक)	PRQ	PR_PRQ	**कोण**\PR_PRQ येत आहे?
2.6	Indefinite (अनिश्चित)	PRI	PR_PRI	**कोणी**\PR_PRI **कोणास**\PR_PRI हासू नये. त्या पेटीत **काय**\PR_PRI आहे ते सांगा.
3	**Demonstrative (दर्शक)**	**DM**	**DM**	**हे** पुस्तक माझे आहे.
3.1	Deictic	DMD	DM_DMD	**तो**\DM_DMD मुलगा हुशार आहे. **हा**\DM_DMD मुलगा हुशार आहे. **ही**\DM_DMD मुलगी सुंदरआहे. **जेथे**\DM_DMD राम होता **तेथे**\DM_DMD तो होता.
3.2	Relative	DMR	DM_DMR	**हे**\DM_DMR लाल रंगाचे असते.
3.3	Wh-word	DMQ	DM_DMQ	**कोणता**\DM_DMQ मुलगा हुशार आहे?
4	**Verb (क्रियापद)**	**V**	**V**	
4.1	Main (मुख्य क्रियापद)	VM	V_VM	तो घरी **गेला**\V_VM.
4.2	Auxiliary (सहाय्यक क्रियापद	VAUX	V_VAUX	राम घरी जात **आहे**\V_VAUX.

Sr. No	Category	Label	Annotation Convention	Examples
5	**Adjective (विशेषण)**	**JJ**		**सुंदर**\JJ मुलगी
6	**Adverb (क्रियाविशेषण)**	**RB**		**हळूहळू**\RB चाल.
7	**Conjunction (उभयान्वयी अव्यय)**	CC	CC	
7.1	Coordinator	CCD	CC_CCD	तो **आणि**\ CC_CCD मी.
7.2	Subordinator	CCS	CC_CCS	**जर**\CC_CCS त्याने सांगितले असते **तर**\CC_CCS हे काम मी केले असते.
7.2.1	Quotative	UT	CC_CCS_UT	**असे**\CC_CCS_UT **म्हणून**\C_CCS_UT तो पुढे गेला.
8	**Particles**	**RP**	**RP**	
8.1	Default	RPD	RP_RPD	मी **तर**\RP_RPD खूप दमले.
8.2	Interjection (उद्गार वाचक)	INJ	RP_INJ	**अरेरे**\RP_INJ ! सचिनची विकेट ढापली.
8.3	Intensifier (तीव्र वाचक)	INTF	RP_INTF	राम **खूप**\RP_INTF चांगला मुलगा आहे.
8.4	Negation (नकारात्मक)	NEG	RP_NEG	नको, न
9	**Quantifiers**	**QT**	**QT**	
9.1	General	QTF	QT_QTF	**थोडी**\QT_QTF साखर द्या.
9.2	Cardinals	QTC	QT_QTC	मला **एक**\QT_QTC गोळी दे.
9.3	Ordinals	QTO	QT_QTO	माझा **पहिला**\QT_QTO क्रमांक आला.
10	**Residuals (उर्वरित)**	**RD**	**RD**	
10.1	Foreign word	RDF	RD_RDF	
10.2	Symbol	SYM	RD_SYE	$, &, *, (,),
10.3	Punctuation	PUNC	RD_PUNC	. (period), ,(comma), ;(semi-colon), !(exclamation),? (question), : (colon), etc.
10.4	Unknown	UNK	RD_UNK	Not able to identify the Tag.
10.5	Echo-words	ECH	RD_ECH	जेवण बिवण, डोके बिके

Marathi Annotated Corpora

- In Marathi, there is around 100k annotated data developed at IIT Bombay as a part of ILCI project funded by MeitY, New Delhi.

- This ILCI corpus consists of four domains viz., Tourism, Health, Agriculture, and Entertainment.
  - Tourism - 25K (parallel)
  - Health - 25K (parallel)
  - Agriculture - 10K (parallel)
  - Entertainment - 10K (parallel)
  - General – 30K (monolingual)

- This tagged data is used for various applications like chunking, dependency tree banking, word sense disambiguation, etc.

- This ILCI annotated data forms a baseline for Marathi POS tagging and is available for download at TDIL portal.

Lexical and Functional POS Tagging

- Lexical POS tagging (Lexical or L approach) deals with tagging of a word at a token level.
- Functional POS tagging (Functional or F approach) deals with tagging of a word as a syntactic function of a word in a sentence.
- Example: In the phrase ‘golf stick’, the POS tag of the word ‘golf’ could be determined as follows:
    - Lexically it is a noun as per lexicon.
    - Functionally it is an adjective as it is a modifier of succeeding noun.

Lexical and Functional POS Tagging: Challenges and Discussions

- Subordinators which act as Adverbs
  - ज्याप्रमाणे (*jyApramANe*, likewise), त्याप्रमाणे (tyApramANe, like that), ह्याप्रमाणे (*hyApramANe*, like this), जेव्हा (*jevhA*, when) and तेव्हा (*tevhA*, then).
  - ज्याप्रमाणे (*jyApramANe*) and ह्याप्रमाणे (*tyApramANe*) are generated from pronominal stems viz., ज्या (*jyA*) and ह्या (*hyA*)
  - They are lexically qualified as pronouns, hence lexically tagged as pronouns
  - However, they function as adverbs; hence to be functionally tagged as RB.
  - When these words appear as part of the clause then they should be functionally tagged as CCS.

- Words with Suffixes
  - There are suffixes like मुळे (*muLe*, because of; due to), साठी (*sAThI*, for), बरोबर, (*barobara*, along with), etc.
  - When these suffixes are attached to pronouns are lexically tagged as PRP. However functionally they are tagged as CCD.

- Words which are Adjectives
  - Consider the example below: त्याच्यामध्ये ही कला परंपरागत चालत आली आहे (*tyAchyAmadhye hI kalA paraMparAgata chAlataAlIAhe*, this art has come to him by tradition).
  - Lexically, the word परंपरागत (*paraMparAgata*, traditional) is an adjective,
  - But, in the above sentence, it qualifies the verb चालत येणे (*chAlatayeNe*, to be practiced). Hence functionally, the word परंपरागत (*paraMparAgata*) should be tagged as an RB.

Lexical and Functional POS Tagging: Challenges and Discussions

- Adnominal Suffixes Attached to Verbs
  - The adnominal suffix जोगं (*jogaM*) and all its forms (जोगा, जोगी, जोगे, जोग्या; *jogA*, *jogI*, *joge*, *jogyA*) are always attached to verbs.
  - For example, word करण्याजोग्या (*karaNyAjogyA*, doable) is lexically tagged as a verb.
  - However, word करण्या (*karaNyA*) is a Kridanta form of a verb करणे (*karaNe*, to do) and suffix जोगं (*jogaM*) is an adnominal suffix attached to Kridanta form; hence, a verb with all the forms of जोगं (*jogaM*) should functionally be treated as adjectives. Therefore verbs with adnominal suffix should be tagged as JJ.

- Words जसे (*jase*) तसे (*tase*)
  - words जसे (*jase*, like this) and तसे (*tase*, like that) are lexically tagged as adverbs.
  - All these words function as a relative pronoun in a sentence. Hence, the words and their variations should be functionally tagged as PRL.

- Word तसेतर (*tasetara*)
  - A word तसेतर (*tasetara*, as it is seen) is the same as तसे पाहिले तर (*tase pAhile tara*, as it is seen).
  - Lexically, it can be tagged as a particle (RPD)
  - but since it has a function of conjunction; it should be tagged as CCD.

Lexical and Functional POS Tagging: Challenges and Discussions

- Word मात्र (mAtra)
  - A word मात्र (*mAtra*) is very ambiguous in its various usages;
  - It is difficult to functionally identify the POS of this word at a sentence level.
  - Various meanings of word मात्र (*mAtra*) are given in lexicon. It should be tagged as per its usage in the sentence.
  - When the word मात्र (mAtra) conveys the meaning of ही, देखील, सुद्धा (hI, dekhIla, suddhA; also) then it should be tagged as RB functionally.
  - When a word is related to the preceding word तेथे (tethe, there) and its function is an emphatic marker च (cha) then it should be tagged as RPD functionally.
  - When word मात्र (mAtra) appears in the form of conjunction then it should be marked as CC functionally.
  - If it is modifying the succeeding noun, then it should be tagged as JJ functionally.
  - If it is modifying the preceding word, then the tag will be RPD as a particle functionally.

Lexical and Functional POS Tagging: Challenges and Discussions

- Word अन्यथा (anyathA)
  - Lexically word अन्यथा (*anyathA*, otherwise; else; or) is an adverb/indeclinable.
  - However, it behaves like conjunction at the sentence level and hence it should be tagged as CCD.

- Different Forms of कसा (kasA)
  - As per BIS Tag-set, words कसा, कशी, कसे (kasA, kashI, kase; how) shall be tagged as PRQ.

  - However, the PRQ tag is only for pronoun category and the word कसा (kasA) is not a pronoun; it can behave as an adverb or as a modifier.

  - Consider the examples below:
    - तो माणूस कसा आहे हे त्याच्याशी बोलल्यावरच कळेल (to mANUsa kasA Ahe he tyAchyAshI bolalyAvaracha kaLela, we will come to know about him only after talking to him) [adnominal]
    - सरकारी ठरावाने कायद्याचे कलम कसे रद्द होणार (sarakArI tharAvAne kAyadyAche kalama kase radda hoNAra, How can this clause of law be prohibited by Government Resolution?) [adverbial]

  - In the 1st case, word कसा (kasA, how) functionally acts as a pronoun, hence to be tagged as PRQ. While, in the 2nd case, it acts as an adverb, hence to be functionally tagged as RB

POS Ambiguity: Challenges and Discussions

- Ambiguous Words: ते (te) and तेही (tehI)
  - The word ते (te) has different grammatical categories like pronoun (they), demonstrator (that) and conjunction (to). Examples:
  - ३० ते ४० (30 te 40, 30 to 40)
  - The word ते (te) lexically and functionally acts as conjunction, hence to be tagged as CCD.
  - ते म्हणाले (te mhaNAle, they said)
  - Here word ते (te) acts as personal pronoun, hence to be tagged as PR_PRP
  - ते कुठे आहेत? (te kuThe Aheta?, where are they?)
  - Here word ते (te) acts as relative demonstrator, hence to be tagged as DM_DMR
  - राकेशने पोलीसांना फोन केला आणि ते दोन्ही चोर पकडले गेले (rAkeshane polIsAMnA phona kelA ANi te donhI chora pakaDale gele, Rakesh called police and those two thieves got caught).
  - Here, word ते (te) is modifying its succeeding noun चोर (chora, thief) so it is Deictic demonstrator, hence to be tagged as DM_DMD.
  - त्यांना हे कधीच पसंत नव्हते, त्यांच्या मुलाने संगीत शिकावे आणि तेही नृत्य (tyAMnA he kadhIchapasaMtanavhate, tyAMchyAmulAnesaMgItashikAveANitehInRRitya, He never wanted his son to learn music and that too the dance form)
  - Here, the word तेही (tehI) is an ambiguous word. It is modifying succeeding noun or previous context. Here, ही (hI) is a bound morpheme and conveys the meaning ‘also’. Therefore word तेही (tehI) should be tagged as DM_DMR.

POS Ambiguity: Challenges and Discussions

Several POS level ambiguity issues were faced by annotators while annotating the Marathi corpus. Following are some POS specific ambiguity problems encountered while annotating.

- Ambiguous POS: Adjective or Noun?
  - Examples: वयस्कर (vayaskara, the aged)
    - कुटुंबाच्या वयस्कर सदस्यांनी मतदान केले (kuTuMbAchyA vayaskara sadasyAMnI matadAnakele, all the aged members of the family voted).
    - सर्व वयस्करांनी मतदान केले (sarva vayaskarAMnI matadAna kele, all the aged people voted).
  - Here, the word वयस्कर (vayaskarAMnI) lexically acts as an adjective as well as a noun.
  - At the syntactic level, in the first example, it is functioning as adjective hence to be tagged as JJ, while in the second example it is functioning as a noun hence to be tagged as N_NN.
- Ambiguous POS: Demonstrators
  - Demonstrators such as हा, ही, हे, तो, ती, ते ((hA, hI, he), this), (to, tI, te), that)
  - Simple guideline can be followed is, if the demonstrator is directly following noun, then tag it as DMD, otherwise tag it as DMR i.e., if the demonstrator is referring to previous noun/person.
- Ambiguous POS: Noun and Conjunction
  - Example: word कारण (kAraNa, reason; because).
  - At semantic level, the word कारण (kAraNa) has two meanings, one is ‘a reason’ which acts as a noun and another is ‘because’ which acts as a conjunction.

POS Ambiguity: Challenges and Discussions

- Ambiguous word: उलटा (ulaTA)

  Examples:
  - उलटे टांगून सुकवले जाते (ulaTe TAMgUna sukavale jAte). Here, उलटे (ulaTe, upside down is behaving as manner, not a noun, hence to be tagged as RB.
  - उलटे भांडे सुलटे कर (ulaTe bhAMDe sulaTe kara). Here उलटे (ulaTe) it is modifying succeeding noun, hence it is an adjective, hence to be tagged as JJ.
  - In the above examples, annotator should identify word behavior in the sentence and tag accordingly.

- Ambiguous words: कितीही (kitIhI), ना का (nA kA) and असू दे ना का (asU de nA kA)

  Examples
  - संगणक हा कितीही प्रगत किंवा चतुर असू दे ना का, तो केवळ तेच काम करू शकतो ज्याची विधी (पद्धत) आपल्याला स्वत: माहित आहे. (saMgaNaka hA kitIhI pragata kiMvA chatura asU de nA kA, to kevaLa techa kAma karU shakato jyAchI vidhI (paddhata) ApalyAlA svata: mAhita Ahe¸ The computer how much ever may be advanced and clever, it only does that work whose method we only know). Here, कितीही (kitIhI, how much) is a quantifier, hence to be tagged as QTF.
  - In the phrase असू दे ना का (asU de nA kA), the token ना (nA) is a part of verb असू दे (asU de, let it be) and should be tagged as VM, hence the phrase should be tagged as VM, while the token का (kA) is acting as a particle in this phrase and not as a question marker, therefore का (kA) should be tagged as RPD.
  - किती माणसे जेवायला होती? (kitI mANase jevAyalA hotI, how many people were there for a meal?). Here, किती (kitI, how many) is a question so it should be tagged as DMQ.

POS Ambiguity: Challenges and Discussions

- Ambiguous word: तर (tara)

  Examples:

  - Conjunction: जर मी वेळीच गेलो नसतो तर हा वाचला नसता (jara mI veLIcha gelo nasato tara hA vAchalA nasatA, if I had not gone on time he would have not survived).
  - Particle: 'हो! आता मी जातो तर!' = 'मी अजिबात जाणार नाही'('ho! AtA mI jAto tara!' = 'mI ajibAta jANAra nAhI', 'yes! now I am leaving then' = 'I am not at all leaving').
  - In the above sentences, the word तर (tara) is used as a supplementary or stressable word so somewhat special as to give meaning in the sentence. (Date-Karve, 1932). Hence it should be treated as CCD.
  - तुम्ही तर लाख रुपये मागतां व मी तर केवळ गरीब पडलो (tumhI tara lAkha rupaye mAgatAM va mI tara kevaLa garIba paDalo, you are asking for lakh rupees and I am a poor person). In this sentence, the word तर (tara) indicates opposition with respect to meaning between two connected sentences. (Date-Karve, 1932). Hence, it should be treated as a RPD.

Some Special Cases

- Words आम्लयुक्त (Amlayukta), मलईरहित (malaIrahita), मेदरहित (medarahita), दुष्काळग्रस्त (duShkALagrasta) are combinations of noun plus adjective suffix such as युक्त (yukta), ग्रस्त (grasta) and रहित (rahita). In such cases, even though noun is a head string and adjective part is a suffix, the whole word shall be tagged as JJ.

- Before tagging अभंग (abhaMga, verses), ओव्या (ovyA, stanzas), काव्य (kAvya, poetry), etc., annotator shall first read between the lines; understand the meaning which it conveys and then decide upon the grammatical categories of each token. For example, in sentence कळावे तयासी कळे अंतरीचे कारण ते साचे साच अंगी (kaLAve tayAsI kale aMtarIche kAraNa te sAche sAcha aMgI) the POS tagging should be done as साचे\V_VM साच\N_NN अंगी\N_NN, etc.

- Doubtful cases of word कोणता (koNatA)

  Examples:
  - कोणता मुलगा हुशार आहे (koNatA mulagA hushAra Ahe)?
  - वाहतुकीच्या दरम्यान कोणतीही हानी झालेली नाही (vAhatukIchyA daramyAna koNatIhI hAnI jhAlelI nAhI).
  - ह्यांच्या बोलण्याचा माझ्यावर कोणताही परिणाम झाला नाही (hyAMchyA bolaNyAchA mAjhyAvara koNatAhI pariNAma jhAlA nAhI).
  - शेतकऱ्यास कोणत्याही वर्षी पाण्याची कमतरता भासणार नाही (shetakaryAsa koNatyAhI varShI pANyAchI kamataratA bhAsaNAra nAhI).
  - Here, in the 1st example, the word कोणता (koNatA, which one) undoubtedly is DMQ. In rest of the examples कोणतीही (koNatAhI, whichever, whomever), कोणताही (koNatAhI, whichever, whomever), कोणत्याही (koNatyAhI, whichever, whomever) are DM adjective (DMD).

Summary

- Marathi POS tagging is an important activity for NLP tasks.
- While tagging, several challenges and issues were encountered.
- In this paper, Marathi BIS tag-set has been discussed.
- Lexical and functional tagging approaches were discussed with examples.
- Further, various challenges, experiences, and special cases have been presented.
- The issues discussed here will be helpful for annotators, researchers, language learners, etc. of Marathi and other languages.

- In future, more issues such as tagging for words having multiple senses; words having multiple functional tags will be discussed.
- Also, tagset comparison of close languages will be done.
- Further, the evaluation of lexical and functional tagging using statistical analysis will be done.

References

Akshar Bharati, Dipti Misra Sharma, Lakshmi Bai, Rajeev Sangal. (2006). AnnCorra : Annotating Corpora Guidelines For POS And Chunk Annotation For Indian Languages. *Language Technologies Research Centre*, IIIT, Hyderabad.

Chitra V. Chaudhari, Ashwini V. Khaire, Rashmi R. Murtadak, Komal S. Sirsulla. (2017). Sentiment Analysis in Marathi using Marathi WordNet. *Imperial Journal of Interdisciplinary Research (IJIR)* Vol-3, Issue-4, 2017 ISSN: 2454-1362.

Damle, Moro Keshav. (1965). Shastriya Marathi Vyakran. *A scientific grammar of Marathi,* 3rd edition. Pune, India: RD Yande.

Daniel Jurafsky & James H. Martin. (2016). Speech and Language Processing.

Edna Vaz, Shantaram V. Walawalikar, Dr. Jyoti Pawar, Dr. Madhavi Sardesai. (2012). BIS Annotation Standards With Reference to Konkani Language. *24th International Conference on Computational Linguistics (COLING 2012)*, Mumbai.

Lata Popale and Pushpak Bhattacharyya. (2017). Creating Marathi WordNet. *The WordNet in Indian Languages. Springer*, Singapore, 2017. 147-166.

Pushpak Bhattacharyya, (2015). Machine Translation, *Book published by CRC Press,* Taylor and Francis Group, USA.

Yashwant Ramkrishna Date, Chintman Ganesh Karve, Aba Chandorkar, Chintaman Shankar Datar. (1932). Maharashtra Shabdakosh. *Published by H. A. Bhave, Varada Books*, Senapati Bapat Marg, Pune.

Thank You