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
Indian Institute of Technology Bombay, Mumbai, India
{gkrane45, joshinilesh60, geetanjaleerane, hanumanredkar, malharku and pushpakhbh}@gmail.com

Abstract
Part of Speech (POS) annotation is a significant challenge in natural language processing. The paper discusses issues and challenges faced in the process of POS annotation of the Marathi data from four domains viz., tourism, health, entertainment and agriculture. During POS annotation, a lot of issues were encountered. Some of the major ones are discussed in detail in this paper. Also, the two approaches viz., the lexical (L approach) and the functional (F approach) of POS tagging have been discussed and presented with examples. Further, some ambiguous cases in POS annotation are presented in the paper.

Keywords: Marathi, POS Annotation, POS Tagging, Lexical, Functional, Marathi POS Tagset, ILCI

1 Introduction
In any natural language, Part of Speech (POS) such as noun, pronoun, adjective, verb, adverb, demonstrative, etc., forms an integral building block of a sentence structure. POS tagging is one of the major activities in Natural Language Processing (NLP). In corpus linguistics, POS tagging 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. A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc. The term ‘Part-of-Speech Tagging’ is also known as POS tagging, POST, POS annotation, grammatical tagging or word-category disambiguation.

In this paper, the challenges and issues in POS tagging with special reference to Marathi have been presented. The Marathi language is one of the major languages of India. It belongs to the Indo-Aryan Language family with about 71,936,894 users. It is predominantly spoken in the state of Maharashtra in Western India (Chaudhari et. al., 2017). In recent years many research institutions and organizations are involved in developing the lexical resources for Marathi for NLP activities. Marathi Wordnet is one such lexical resource developed at IIT Bombay (Popale and Bhattacharyya, 2017).

The paper is organized as follows: Section 2 introduces POS annotation; section 3 provides information on Marathi annotated corpora; section 4 describes Marathi tag set; section 5 explains tagging approaches, section 6 presents ambiguous behaviors of the Marathi words, section 7 presents a discussion on special cases, and section 8 concludes the paper with future work.

2 Parts-Of-Speech Annotation
In NLP pipeline POS tagging is an important activity which forms the base of various language processing applications. Annotating a text with POS tags is a standard low-level text pre-processing step before moving to higher levels in the pipeline like chunking, dependency parsing, etc. (Bhattacharyya, 2015). Identification of the parts of speech such as nouns, verbs, adjectives, adverbs for each word (token) of the sentence helps in analyzing the role of each word in a sentence (Jurafsky D. et. al., 2016). It represents a token level annotation wherein it assigns a token with POS category.

3 Marathi Annotated Corpora
Aim of POS tagging is to create a large annotated corpora for natural language processing, speech recognition and other related applications. Annotated corpora serve as an important resource in NLP activities. It proves to be a basic building block for constructing statistical models for the automatic processing of natural languages. The significance of large annotated corpora is widely appreciated by researchers and application developers. Various research institutes in India viz., IIT Bombay, IIIT Hyderabad, JNU New Delhi, and other institutes have developed a large corpus of POS tagged data. In Marathi, there is around 100k annotated data developed as a part of Indian Languages Corpora Initiative (ILCI) project funded by MeitY, New Delhi. This ILCI corpus consists of four domains viz., Tourism, Health, Agriculture, and Entertainment. This tagged data (Tourism - 25K, Health - 25K, Agriculture - 10K, Entertainment - 10K, General – 30K) is used for various applications like chunking, dependency tree banking, word sense disambiguation, etc. This ILCI

1http://language.worldofcomputing.net/pos-tagging/parts-of-speech-tagging.html
2http://www.indianmirror.com/languages/marathi-language.html
3http://www.censusindia.gov.in/
annotated data forms a baseline for Marathi POS tagging and is available for download at TDIL portal\(^9\).

4 The Marathi POS Tag-Set

The Bureau of Indian Standards (BIS)\(^8\) has come up with a standard set of tags for annotating data for Indian languages. This tag-set is prepared for Hindi under the guidance of BIS. 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 Dept. of Information Technology, MeitY and presented as Unified POS standard in Indian languages\(^11\). Marathi POS tag-set has been prepared at IIT Bombay referring to the standard BIS POS Tag-set, IIIT Hyderabad guideline document (Bharati et al, 2006) and Konkani Tag-set (Vaz et al., 2012). This Marathi POS Tag-set can be seen in Appendix A.

5 Lexical and Functional POS Tagging: Challenges and Discussions

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. In other words, a word can have two roles viz., grammatical role (lexical POS w.r.t. a dictionary entry) and functional role (contextual POS)\(^12\). For 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.

In the initial stage of ILCI data annotation, POS tagging was conducted using the lexical approach. However, over a while, POS tagging was done using the functional approach only. The reason is that, by using the lexical approach we do a general tagging, i.e., tagging at a surface level or token level and by using the functional approach we do a specific tagging, i.e., tagging at a semantic level. This eases the annotation process of chunking and parsing in the NLP pipeline.

While performing POS annotation, many issues and challenges were encountered, some of which are discussed below. Table 1 lists the occurrences of discussed words in the ILCI corpus.

5.1 Subordinators which act as Adverbs

There are three basic types of adverbs. They are time (N_NST), place (N_NST) and manner (RB). Subordinators are conjunctions which are tagged as CCS.

However, there are some subordinators in Marathi which act as adverbs. For example, जोगाणे (tyApramANE, like that), ह्या-प्रामाणे (hyApramANE, like this), जेवा (jeyhA, when) and तेवा (tevA, then), जोगाणे (jyApramANE) and ह्या-प्रामाणे (tyApramANE) are generated from pronominal stems viz., जि (jyA) and ह्या (hyA) hence they are lexically qualified as pronouns, however, they function as adverbs; hence to be functionally tagged as RB at the individual level. However, when these words appear as part of the clause then they should be functionally tagged as CCS.

This distinction was also observed by noted Marathi grammarian, Damle\(^13\) (Damle, 1965) [p. 206-07].

5.2 Words with Suffixes

There are suffixes like मुळे (muLe, because of; due to), साठी (sATHi, for), बरोबरे (barobar, along with), etc. When these suffixes are attached to pronouns they function as adverbs or conjunctions at a syntactic level. For example, words लागाने (tyAmule, because of that), लागाने (tyAmule, because of this), लागाने (tyChyAmule, because of it), लागाने (tyChyAmule, because of them), लागाने (tyChyAmule, further) are formed by attaching the above suffixes to pronouns. These words which are formed are lexically tagged as PRP. However, functionally these words act as conjunctions at the sentence level; therefore, they should be tagged as CCD. Also, consider the words जोगी (tyAveLI, at that time), जोगी (tyAveLI, at this time), जोगी (tyAnaMIara, after this), जोगी (tyAnaMIara, after that). Here, wherever the first string/morpheme appears as जि (jyA) and ह्या (hyA), the tag should be given as PRP, lexically. But functionally, all these words shall be tagged as N_NST (time adverb).

5.3 Words which are Adjectives

Adjectives are tagged as JJ. Consider the example below: व्यावसायिक ही कला परंपरागत बालक जोगी आहे (tyChyAChyAMhye hi kala paraMparAgata balaJogaAhe, 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 बालक जोगी (balaJoga) to be practiced. Hence functionally, the word परंपरागत (paraMparAgata) should be tagged as an RB. Similarly, a word वाईट (vAIIta, bad) has a lexical POS as an adjective (Date-Karve, 1932). But in the sentence माणे वाईट वाटते (maLa vAIIta vAITa, I am feeling bad), it functions as an adverb, as it is qualifying the verb and not preceding the pronoun माणे (maLa, I; me). Therefore, functionally word वाईट (vAIIta) acts as adverb hence should be tagged as RB.

5.4 Adnominal Suffixes Attached to Verbs

The adnominal suffix जंग (jogaM) and all its forms (jonga, jonga, jonga, jonga, jogA, jogI, jogE, jogvyA) are always attached to verbs. For example, word करण्यांक्रया (karaNyA-jogvyA, doable) is lexically tagged as a verb. However, word करण्या (karaNyA) is a Kridanta form of a verb क्रया (karaNe, to do) and suffix जंग (jogaM) is an adnominal suffix attached to Kridanta form; hence, a verb with all the

---

\(^{8}\) https://www.tdil-dc.in/

\(^{9}\) http://www.bis.gov.in/

\(^{10}\) http://tdil-dc.in/tdildcMain/articles/134692Draft POS Tag standard.pdf

\(^{11}\) https://www.cicling.org/2018/cicling18-hanoi-special-event-23mar18.pdf

\(^{12}\) http://www.cfilt.iitb.ac.in/damale/index.html
forms of जोग (jogam) should functionally be treated as adjectives. Therefore verbs with adnominal suffix should be tagged as JJ.

5.5 Words जो (jose) तसे (tase)
As per Damle (1956), words जो (jose, like this) and तसे (tase, like that) are tagged as adverbs. However, if they appear with nouns in a sentence, they are influenced by the inflection and gender property of that nominal stem. For example, words जो (jose, like this) and तसे (tase, like that) have inflected forms like जासा (jasa, like him), अजी (ajhi, like her), तसा (tasA, like this), अजी (ajhi, like this), तसे (tase, like this), etc. All these words function as a relative pronoun in a sentence. Hence, the words and their variations should be functionally tagged as PRQ.

5.6 Word तसेतरा (tasetara)
A word तसेतरा (tasetara, as it is seen) is the same as तसे (tase) पहले (tase pahle 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. For example, in a sentence तसेतरा लागवामुखी काही बलय देतला (tasetara ta- NAvAmuLeH kALI valaTa yeTaka, as it is seen that black circles appear because of stress as well), word तसेतरा (tasetara) functions as conjunction and hence should be tagged as CCD instead of tagging it as RPD.

5.7 Word अन्यत्था (anathA)
The standard dictionaries give POS of the word अन्यत्था (anathA, otherwise; else; or) as an adverb/indeclinable. For example, consider a sentence अन्यत्था ते येणार माणे (anathA to yeNara mAtra, Otherwise, he will not come). Here, while annotating अन्यत्था (anathA) there is a possibility that annotator can directly tag this word as an adverb at a lexical level. However, it behaves like conjunction at the sentence level and hence it should be tagged as CCD.

5.8 Different Forms of कसा (kasA)
As per BIS Tag-set, words कसा, कणी, कणे (kasA, kashI, kase; how) shall be tagged as PRQ. However, the PRQ tag is only for pronom category and the word कसा (kasA) is not a pronoun; it can behave as an adverb or as a modifier. Consider the examples below:

1. तो मापुस कसा आहे हे लोकाच्या शैलीवाच्या काणेल (to mANusa kasA Ahe he tyAchyAshI bolalAyA- varacha kaLaLe, we will come to know about him only after talking to him) [adnominal]
2. सरकारी ठरावाने कायाचे कलम कसे रद होणार (sarakAri TharAvAnE kAyaDyeAche kalam 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.

5.9 Word भार (mAtra)
A word भार (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 भार (mAtra) are given in Data-Karve dictionary. Some of the different senses of भार (mAtra) are discussed here:

- When the word भार (mAtra) conveys the meaning of ही, देखील, हील (hi, dekhIla, suddha; also) then it should be tagged as RB functionally.
- When a word is related to the preceding word ते (tethe, there) and its function is an emphatic marker च (cha) then it should be tagged as RPD functionally.
- When word भार (mAtra) appears in the form of conjunction then it should be marked as CC functionally.
- If the word is modifying the succeeding noun, then it should be tagged as JJ functionally.
- If the word is modifying the preceding word, then the tag will be RPD as a particle functionally.

Therefore, it is noticed that the word भार (mAtra) does not have one single POS tag functionally and it depends upon the appearance in a sentence. Hence, should be tagged as per the usage.

| Token     | Lexical | Functional | Occurrences |
|-----------|---------|------------|-------------|
| भार       | Adverb  | Post-position Conjunction Particle | 1496        |
| तसेतरा    | Particle | Conjunction | 426         |
| कसा       | Wh-word | Adverb      | 269         |
| लागवामुळे   | Conjunction | 530         |
| अजी        | Pronoun | Conjunction | 424         |
| जसा        | Conjunction | 8           |
| तसे         | Time adverb | 244         |
| तसे         | Time adverb | 33          |
| तसे         | Time adverb | 71           |
| तसे         | Time adverb | 298          |
| तसेतरा    | Adverb  | 104         |
| भार        | Adverb  | 246         |
| अजी        | Conjunction | 24           |
| अजी        | Relative pronoun | 1007         |
| अजी        | Relative pronoun | 511          |
| भार        | Verb     | 97          |

Table 1: Occurrences of discussed words and lexical v/s functional tags assigned to these words

6 POS Ambiguity: Challenges and Discussions
Ambiguity is a major open problem in NLP. Several POS level ambiguity issues were faced by annotators while annotating the Marathi corpus. Following are some POS
specific ambiguity problems encountered while annotating.

6.1 Ambiguous POS: Adjective or Noun?
Examples: वयस्कर (vayaskara, the aged)

- कुठाया वयस्कर सदस्यांनी मतादन केलेला (kuTumhIChya vayaskara sadasyAMnI matadAnakele, all the aged members of the family voted).
- सर्वे वयस्करांनी मतादन केले (sarva vayaskarAMnI matadAnakele, all the aged people voted).

In the above examples, the word वयस्कर (vayaskaraAMnI) lexically acts as an adjective as well as a noun. However, 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. This is one of the challenges while annotating adjectives appearing in nominal form. Annotators usually fail to disambiguate these types of words at the lexical level; therefore such words should be disambiguated at syntactic level. Hence, annotators need to take special care while annotating such cases.

6.2 Ambiguous POS: Demonstrators
While annotating demonstrators such as हा, ही, हे, ती, ते ((ha, hi, he, this), (to, ti, te), that) annotators often get confused whether to tag them as DMD or DMR. 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.

6.3 Ambiguous POS: Noun and Conjunction
Example: word कारण (karaNa, reason; because). At semantic level, the word कारण (karaNo) has two meanings, one is ‘a reason’ which acts as a noun and another is ‘because’ which acts as a conjunction. Annotators have to pay special attention while tagging such cases.

6.4 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)
- ते नाहीं (te mhanAile, they said)
- ते हे (te kaThe Aheta?, where are they?)
- ते कुठे आहे? (te kuThe Aheta?, where are they?)
- Here word ते (te) acts as relative demonstrator, hence to be tagged as DM_DMR
- राखेश्ने मोल्लीला कोण नेतीचे आणि ते दोनही चॉरा पकडले गेले (rakeshane moliAMnA phona keI ANI te donhl chora pakaDale gele, Rakesh called police and those two thieves got caught).
- Here, word ते (te) is modifying its succeeding noun चॉरा (chora, thief) so it is Deictic demonstrator, hence to be tagged as DM_DMD.

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.

6.5 Ambiguous word: उलटा (ulaTA)
Examples:

- उलटे टैगून सुकवले जाते (ulaTe TAgUnA sukavale jAt). Here, उलटे (ulaTe), upside down is behaving as manner, not a noun, hence to be tagged as RB.
- उलटे भांडे सुकते कर (ulaTe bhAMDe sulaTe kara). Here उलटे (ulaTe) 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.

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

- संगणक हा कितीही प्रमाण किंवा चूज हवेच असू दे ना का, तो केवल तेच काम करतात कसतील ज्ञानी विचार (pahul) आपघात यें: माहितेच आहेत. (saMaGaNaka ha kitIhI pragata kiMVa chatura asU de nA kA, to kevaLa techa kAma karU jyAchI vidhl (paddha-ta) 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 RPM.
- कितीही माणसे ज्ञानितो? (kitIhI maNase jevAyalA hotI, how many people were there for a meal?). Here, कितीही (kitIhI, how many) is a question so it should be tagged as DMQ.

6.7 Ambiguous word: तरा (tara)
Examples:

- Conjunction: जर मी वेळीच गेलो नसतो तर हा वाचला नसतो (jara mI veLhA gelo nasato tara hA va-chalA nasata, if I had not gone on time he would have not survived).
- Particle: हो! आता मी जातो तर! = ‘मी अविश्वास जाणार नाही’ (ho! AIA mI jAto tara! = ‘mI ajibAIA jANA-
ra 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.

• तुम्हीं तर लाख केंद्रे मागण (tumhi tara lakha kende mara) की तर केंद्र गरीब पडली (tumhi tara lakha kende mara koatalam va ml tara kevala garib 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.

7 Discussions on Some Special Cases

• Words आमलयुक्त (Amlayuktā), मलाल्रहित (malalrahitā), मेदरहित (medarahitā), दुःखालग्रस्त (duShkaLagrasī) are combinations of noun plus adjective suffix such as युक्त (yuktā), ग्रस्त (grasī) and रहित (rahitā). 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), ओख्या (ovyā, stanzas), काव्य (kAvyā, 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 कोणतीही कमतरता केली अंतरीचे कारण ते सांप सांबी अभंग (kO NatIhI kO NatA hAnI jhAlelI nAhI) the POS tagging should be done as सांप_व्ह_सांबी_अभंग_न_अभंग_न_नन, etc.

• Doubtful cases of word कोणता (koNatA)

Examples:

- कोणता मुलगा हुस्यार आहे (koNatA mulagA husharyaA Ahe)?
- वाहतुकीय दरम्यान कोणतीही हानी झालेली नाही (vAtahtkIChya daramyAna koNatIhI hAnI jhAlelI nAhI).
- ह्या कोणत्या माहित्याचे ग्रामानात कोणताही परिणाम झालेला नाही (hyAMchyA bolaNyaAchA mAhyAyAvara koNatAhh pariNAmjA hAhA nAhI).
- गोत्यांच्या कोणत्याच्या वरीयांच्या पाण्याची कमतरता भावाच्याच्या नाही (shaetakaryAasa koNatAhh varShI pAnyAchI kamaratah hAsaNArA nAhI).

Here, in the 1st example, the word कोणता (koNatA, which one) undoubtedly is DMQ. In rest of the examples कोणताही (koNatAhh, whichever, whomever), कोणताही (koNatAhh, whichever, whomever), कोणताही (koNatAhh, whichever, whomever), कोणताही (koNatAhh, whichever, whomever) are DM adjective (DMD).

8 Conclusion and Future Work

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, IIT, 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.

Damel, 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, Shantanam V. Walawalkar, 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). Maharashatra Shabdakosh. Published by H. A. Bhave, Varada Books, Senapati Bapat Marg, Pune.
Appendix A

Marathi Parts of Speech-Set with Examples

| Sl. No | Category            | Label | Annotation Convention | Examples                      |
|-------|---------------------|-------|-----------------------|-------------------------------|
| 1.1   | Common (साधारण वाक्य) | NN    | N, NN                 | गाय, N, NN गोड़यात, N, NN राहे. |
| 1.2   | Proper (स्वतंत्र वाक्य) | NNP   | N, NNP                | सामान्य N, NNP स्वयं N, NNP मारी. |
| 1.3   | Nloc (स्थान-काल)    | NST   | N, NST                | 1. तो पैडे N, NST काम करत होला. 2. माझे ही मुलगी N, NST नवीनी आहे. |
| 2     | Pronoun (अर्थमान)   | PR    | PR                    | मो, PR, PR गंगो. |
| 2.1   | Personal (प्रथम वाक्य) | PRP   | PR, PRP               | मी, PR, PRP आलो. |
| 2.2   | Reflexive (आत्मार्थ) | PRF   | PR, PRF               | मी, PR, PRF आलो. |
| 2.3   | Relative (संबंधी)   | PRL   | PR, PRL               | ज्या प्रे PR, PRL हे सांगितले व्ह्हो हे काम केले पाहिजे. |
| 2.4   | Reciprocal (साम्य)  | PRC   | PR, PRC               | परसरार |
| 2.5   | Wh-word (क्रिया)     | PRQ   | PR, PRQ               | कोण PR, PRQ पाहो आहे? |
| 2.6   | Indefinite (अनुस्मरण) | PRI   | PR, PRI               | कोणी PR, PRI काम PR, PRI हसू नय. हा पैडे काम PR, PRI आहे ते वाना. |
| 3     | Demonstrative (वर्ण) | DM    | DM                    | हे मुलगाव माहे आहे. |
| 3.1   | Deictic (देय)        | DMD   | DM, DMD               | हो DM, DMD मुलगा हवार आहे. |
| 3.2   | Relative             | DMR   | DM, DMR               | हे DM, DMR गात रागात असते |
| 3.3   | Wh-word               | DMQ   | DM, DMQ               | कोणतत DM, DMQ मुलगा हवार आहे? |
| 4     | Verb (क्रियापद)      | V     | V                     | तो DM, DMD मुलगा हवार आहे. |
| 4.1   | Main (मुख्य क्रिया)   | VM    | V, VM                 | तो पैडे मेसा V, VM. |
| 4.2   | Auxiliary (सहायक क्रिया) | VAUX | V, VAUX               | रम पैडे जात आहे V, VAUX. |
| 5     | Adjective (वर्णण)    | JJ    | JJ                    | बुध JJ मुलगा |
| 6     | Adverb (क्रियाविद्युषण) | RB    | RB                    | हे बुधहुजु, RB वाय. |
| 7     | Conjunction (उभयाव्य) | CC    | CC                    | तो आणि CC, CC भाई |
| 7.1   | Coordinator          | CCD   | CC, CCD               | तो आणि CC, CC भाई |
| 7.2   | Subordinator         | CCS   | CC, CCS               | बर CC, CCS राहे सर्वांतले असते तर CC, CCS हे काम भाई केले असते. |
| 7.2.1 | Quotative            | UT    | CC, CCS, UT           | असे CC, CCS, UT मुलगा CC, CCS, UT तो पैडे मेसा. |
| 8     | Particles            | RP    | RP, RPD               | तो हा RP, RPD खाली दसत. |
| 8.1   | Default              | RP    | RP, RPD               | तो हा RP, RPD खाली दसत. |
| 8.2   | Interjection (इंटरजेक्शन) | INJ  | INJ                  | अपूर्त RP, INJ | सज्जनी क्रिया दावारी. |
| 8.3   | Intensifier (तीव्र प्रकट) | INTF | RP, INTF             | रम खुप्प RP, INTF चांगला मुलगा आहे. |
| 8.4   | Negation (दंडरक)    | NEG   | RP, NEG               | नका, न |
| 9     | Quantifiers           | QT    | QT                    | पैडे QT, QT चांगला ज्या. |
| 9.1   | General               | QTF   | QT, QTF               | तो हा QT, QTF सांगितला ज्या. |
| 9.2   | Cardinals             | QTC   | QT, QTC               | तो हा QT, QTC गांव्या ज्या. |
| 9.3   | Ordinals              | QTO   | QT, QTO               | मध्ये पाहिजे QT, QTO क्रमांक आला. |
| 10    | Residuals (खोजी)    | RFD   | RD                    | असे CC, CCS, UT मुलगा CC, CCS, UT तो पैडे मेसा. |
| 10.1  | Foreign               | RFD   | RD                    | असे CC, CCS, UT मुलगा CC, CCS, UT तो पैडे मेसा. |
| 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               | अज्ञात स्क्रिन, डाय विक. |