Morphological Analyzer for AUX Suffixes of Marathi

Swapnil D. Moon

moonswapnil6@gmail.com

Consultant (Non-Medical), ICMR- National Institute of Translational Virology & AIDS Research (ICMR-NITVAR) Pune

Abstract

This research aims to develop a model for analyzing auxiliary (AUX) morphology in Marathi and proposes the design of computational models for various natural language processing (NLP) applications—such as information retrieval, machine translation, and spell-checking—within the framework of finite state technology. A rule-based approach has been adopted for data analysis and the creation of the computational model.

Morphological analysis plays a crucial role in natural language applications like parsing, lemmatization, text generation, machine translation, and document retrieval. Therefore, a robust morphological processing system is fundamental to computational morphology applications.

Marathi, an Indo-Aryan language, exhibits both agglutinative and inflectional morphological characteristics. Despite its linguistic richness, limited efforts have been made to systematically analyze Marathi morphology for the purpose of computational model development. To support this work, foundational linguistic concepts related to morphology and its computational representation are briefly discussed in the following sections.

Introduction:

In linguistics, morphology refers to the identification, analysis, and description of the structure of morphemes—the smallest meaningful units in a language—and other linguistic elements such as words and affixes. Morphological typology is a framework used to classify languages based on how they utilize morphemes. This ranges from analytic languages, which rely on isolated morphemes, to agglutinative languages that

combine morphemes in a "stuck-together" fashion, to fusional languages where morphemes merge together more intricately, and finally to polysynthetic languages, which condense multiple morphemes into single complex words. While words are often considered the smallest syntactic units, in most—if not all—languages, words can be related to one another through systematic morphological rules.

This paper focuses on the morphological analysis of auxiliary (AUX) structures in Marathi, an Indo-Aryan language. Marathi is the official language of Maharashtra and a co-official language in the state of Goa, both located in western India. According to the 2011 Census of India, Marathi is spoken by approximately 83 million people, making it the 19th most spoken language globally and the fourth most spoken native language in India.

Literature Review

Computationally, Bharati et al. (1998) suggested a paradigm-based algorithm for morphological analysis for Hindi. In Hindi, the inflected forms of roots do not allow further attachment of any other suffixes. In contrast, in Marathi once the root is transformed into its inflected form it is followed by suffixes to show its agreement with the other words in the sentence. Some postpositions derive new words which themselves may undergo inflection and allow attachment of other suffixes. This makes the simple paradigm-based model proposed in this work unfit for Marathi morphological analysis (Bharati et al. 1998:5). Dixit et al. (2006) developed a morphological analyzer with the purpose of using it for spell checking. Though their analyzer successfully analyzes the words with a single suffix, its scope is restricted in the handling of only first level suffixes for simple word forms. Shambhavi et al (2001) Introduced Kannada morphology analyzer and generator by using a tire. Ramanathan et.al. (2004) a lightweight stemmer for Hindi. In this research, words conflate terms by suffix removal for information retrieval. Willet.P (2006) proposed the porter stemming algorithm for electronic library and information system. Zahurul, M.D. et al. (2009) developed a lightweight stemmer for Bengali for Bengali language spell checker. Qurat-Ul-Ain Akram and etal. (2009) Assas-band, an affix exception list based Urdu stemmer. It stems the Urdu words using lexical lookup method (Assas-band). Dinesh Kumar and Prince Rana (2010) developed design and development of stemmer for Punjabi. It uses a Brute Force algorithm for stemming the Punjabi words. Vijay Sundar et al. (2010) Introduced Malayalam stemmer for information retrieval. The finite state automata method is used to stem the Malayalam words. In these work researches have given morphological analysis of their respective

languages and they have accounted paradigm of negation.

Verb Inflections

In Marathi, verbal inflection involves the attachment of suffixes to the verb stem. The verb stem may take the form of a base, causative, or passive construction. These inflectional suffixes primarily express inherent verbal features such as tense, aspect, and mood, and they also convey agreement features such as person, number, and gender, based on the subject or object noun with which the verb agrees in the sentence.

These suffixes follow a paradigmatic structure corresponding to the features mentioned above. In certain forms, negative markers become integrated into these inflectional suffixes, resulting in the development of both positive and negative paradigms of verbal inflection. Additionally, there is a second type of negative marker in Marathi that functions as a prefix, placed before the verb stem.

Auxiliary verbs in Marathi also play a significant role in encoding verbal features and are, to a large extent, functionally equivalent to inflectional suffixes. These auxiliary constructions are explored in detail in the following section.

Auxiliary Verbs in Marathi

There are two basic auxiliaries आहे *āhe* 'to be' (the form आहे *āhe* '+ present' in very common) and हो *ho* 'to become / happen'. It inflects present tense with supplementary forms for past and future. Both auxiliaries function as copula as well as tense markers. There is auxiliary अस *as* 'be' which is a future auxiliary mostly used to indicate habitual aspect.

आहे *āhe* shows the number and person inflections. In continuous tense it occurs with main verb and verb normally takes the continuous marker त *t*.

General Rules:

<stem>आह *āh* <present tense to be>

ॐ -*e* <SG+FST/TRD>#

ो *-o* <PL+FST>#

ा *ā* <PL+SND>#

ेस *es* <SG+SND>#

ोत *ot* <PL+FST>#

ात *āt* <PL+SND>#

ेत *et* <PL+TRD>#

तो गा-त आहे

to gā-t āhe

'He is singing'

तो जा-त-ो आहे

to ʣā-t-o āhe

'He is going'

<stem> होत *hot* <past tense to be>

ॊ *a* <NUT+SG+TRD>#

ो *o* <MASC+SG+FST>, <MASC/FEM+PL+FST>#

ी *ī* <FEM+SG+FST/SND/TRD> <NUT+PL+TRD>#

ा *ā* <MASC+SG+SND/TRD>, <MASC+PL+SND>#

े *e* <FEM+SG+FST>, <NUT+SG+TRD>, <MASC+PL+SND/TRD>#

ीस *īs* <FEM+SG+SND>#

ास *ās* <MASC+SG+SND>#

Exception rule no 3: stem हो +त्या <FEM+PL+SND/TRD>#

Exception rule no 4: stem हो +त्यात <FEM+PL+SND>#

<Stem> अस *as* <Future tense to be>, <Past Habitual to be>, <Predictive Aspect to be असाच *asāts*>.

ेन *en* <SG+FST>#

असशील *asaśīl* <SG+SND>#

ेल *el* <SG+TRD>#

ाल *āl* <PL+SND>#

तील *atil*<PL+TRD>#

ु *u* <PL+FST>, < PL FST Past Habitual>

े *e* <SG+FST/SND/TRD Past Habitual>

त *at* <PL+SND/TRD Past Habitual>

<Stem> असायच *as-āyts*

ो *o* <MASC+SG+FST>, <PL+FST>

े *e* <NUT+SG/PL+TRD>, <MASC+PL+SND/TRD>

Note: If FUTINDF Verb is ending C# then <MASC+SG+FST> || <FEM+SG+FST>

marker is ीील || ीिन. If FUTIDNF Verb is ending V# than <MASC+SG+FST> || <FEM+SG+FST> marker is ईल || ईन.

Flowcharts:

Input - Word.

Output - Root + suffixes.

Formal Specification of the Morphological Parsing Flowchart

Start:

⬇

Input:Word.
 ⬇
 Check if word exists in lexicon

➡Yes→Output:Root+Suffixes→End.
 ➡No→ApplyMorphologicalRules.
 ⬇
 Apply Suffix Stripping.
 ⬇
 Identify Root + Suffixes
 ⬇
 Validate Root in Lexicon
➡Yes → Output: Root + Suffixes → End.
➡ No → Output: Unknown or Invalid Word → End.

Formal Specification of the Morphological Parsing Flowchart

·   Input: Any inflected word in Marathi

·   Process:

o Match against lexicon

- o If not found, apply morphotactic rules

- o Strip suffixes based on known patterns

- o Identify root

- o Validate root against lexicon

- Output: Root + Identified Suffixes or "Unknown Word"

Flowchart: Morphological Parsing Rules

Start

 |

Input Word

 |

Is word in Lexicon

 |

 |-- Yes --> Output: Root + Suffixes

 |

 |-- No --> Apply Morphological Rules

        |

     Strip Suffixes

        |

     Is Root in Lexicon

     |

     |-- Yes --> Output: Root + Suffixes

|-- No --> Output: Unknown Word

|

Check Data Base

|

Output: Root + Suffixes

|

End

Computational Morphology Rules for AUX

1)   Rule no: (1)

Rule for PNTTNS AUX verbs SUFFIX #

Process specification (algorithm):

Step 1: check the word which is ending with ॆ *e* || ा *ā* || ॊ *o* || ॊत *ot* || ॆत *et* || ॆस *es* || ात *āt* #

Step 2: strip the found suffix from the word#

Step 3: store the found suffix#

Step 4: Find PNTTNS AUX verb stem from stem database#

Step 5: Find a suffix ॆ *e* from suffix database <MASC+SG+FST || <MASC+SG+TRD> <FEM+SG+FST> || <FEM+SG+TRD> || <NUT+SG+TRD>#

(or)

Step 5: Find a suffix ा *ā* from suffix database <MASC+PL+SND> || <FEM+PL+SND>#

(or)

Step 5: Find a suffix ॊ *o* from suffix database <MASC+PL+FST> || <FEM+PL+FST>#

(or)

Step 5: Find a suffix ॊत *ot* from suffix database <MASC+PL+FST> || <FEM+PL+FST>#

(or)

Step 5: Find a suffix ॆत *et* from suffix database <MASC+PL+TRD> || <FEM+PL+TRD> || <NUT+PL+TRD>#

(or)

Step 5: Find a suffix ॆस *es* from suffix database <MASC+SG+SND> || <FEM+SG+SND>#

(or)

Step 5: Find a suffix ात *āt* from suffix database <MASC+PL+SND> || <FEM+PL+SND> #

Step 6: Get the POS of stem and Suffix#

Example: आहे *āhe* <SG FST>

आहोत *āhot* <PL FST>

Conclusion:

From the above analysis/ discussion, it is understood or understandable that the morphological processing improves the recovery performance for Marathi Language. An important observation is that the suffixes in Marathi can also contribute to the semantics of the document and hence improves the retrieval performance.

The current morphological analysis does not handle derivational morphology. In Marathi, derivational morphology is a very productive way of forming words. Handling derivational morphology can also increase the system performance. This paper presented a highly accurate morphological analysis of AUX in Marathi which is very efficiently finds the Root word of a given word and recognizes the Gender of the sentence with the

inputs.

Looking ahead, further work is needed to address the parsing of derivational morphemes for grammatical categories beyond verbs.

References:

Walmbe, M. R. (2010). Sugam Marathi Vyakaran Lekhan. Pune, Nitin prakashan.

Bhagwat, S. (2003). Tumache Amche Marathi vyakran. Latur, Widyabharti prakashan.

Ramesh V. Dhongde and Kashi Wali (2009). Marathi. North America, John Benjamins Publishing Company.

Rajeshwari V. Pandharipande (1997). Marathi. London and New York.

Kelkar, A. R. (1958). The Phonology and Morphology of Marathi A Thesis. Presented to the Faculty of theGraduate School of Cornell University for the Degree of Doctor of Philosophy.

Andrew Carstairs McCarthy. (1992). Current Morphology. London and New York, Routledge.

Khandekar, D. (2012). Verb Dictionary. Pune.

Bhatia, T. K. (1979). Negation in South Asian Languages. Urbana-Champaign, South Asian Languages AnalysisVol. 1, Dept. of Linguistics, University of Illinois. 1-12.

Bhatia, T. K. (1993). Punjabi: a cognitive-descriptive grammar. Descriptive Grammars, Routledge, London.

Bhatt, R. and Munshi, S. (2002). Negation Movement and Optional Verb Second in Kashmiri. Handout of talkgiven at SALA 22 at the University of Iowa.

Bhatt, R. M. (1999). Verb Movement and the Syntax of Kashmiri. Studies in Natural Language and Linguistic Theory 46, Kluwer, Dordrecht.

Cardona, G. (1965). A Gujarati Reference Grammar. Philadelphia,The University of

Pennsylvania Press.

Dasgupta, P. (1987). Sentence Particles in Bangla. Selected Papers from SALA 7, Indiana University LinguisticsClub, Bloomington, 49-75.

Kumar, R. (2003). The Syntax of Negation in Hindi. Doctoral dissertation. University of Illinois-Urbana,Urbana-Champaign, Illinois.

Lahiri, U. (1998). Focus and Negative Polarity in Hindi. Natural Language Semantics. 6:1, 57-123.

Mahajan, A. K. (1990). LF Conditions on Negative Polarity Licensing. Lingua 80:4, 333-348.

Masica, C. (1991). The Indo-Aryan Languages. Cambridge Language Surveys, Cambridge University Press,Cambridge, England.

Prado, Z. N. and Gair, J. (1994). The position of negation in Bengali: An account of synchronic and diachronic variation," in A. Davison and F. M. Smith, eds., Papers from the Fifteenth SALA Roundtable Conference 1993, South Asian Studies Program, University of Iowa, Iowa City, 234-250.

Rahman, S. (2002). Verb Movement and Negation in Bengali. Handout of a talk given at SALA 22 at theUniversity of Iowa.

Ramchand, G. C. (2001). Tense and Negation in Bengali. Linguistic Structure and Language Dynamics inSouth Asia: Papers from the Proceedings of the SALA XVIII Roundtable, MLBD Series in Linguistics 15, Motilal Banarsidass, Delhi, 308-326.

Ramchand, G. C. (2003). Two Types of Negation in Bengali. Clause Structure in South Asian Languages,

Loksatta

Maharashtra Times

Sakal

Tarun Bharat

Lokmat

Citation:Moon, *Swapnil D. (2025). Morphological Analyzer for AUX Suffixes of Marathi, HindiTech: A Blind Double Peer Reviewed Bilingual Web-Research Journal, 16 (2), 1-12. URL:*

https://hinditech.in/morphological-analyzer-for-aux-suffixes-of-marathi/