Pratiksha Gawade, Deepika Madhavi, Jayshree Gaikwad, Sharvari Jadhav, Rahul Ambekar / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622
www.ijera.com Vol. 3, Issue 2, March -April 2013, pp.322-326

Morphological Analyzer for Marathi using NLP

Pratiksha Gawade Deepika Madhavi Jayshree Gaikwad Sharvari Jadhav Rahul Ambekar
I. T. Department,
Padmabhushan Vasantdada Patil Pratishthan's College Of Engineering.
Sion (East), Mumbai-400 022

Abstract

Morphology is a part of linguistic that deals with study of words, i.e internal structure and partially their meanings. A morphological analyzer is a program for analyzing morphology for an input word, it detects morphemes of any text. In current technique, only provides dictionary which defines the meaning of the word, but does not give the grammatical explanation regarding that word. In propose system, we evaluate the morphological analyzer for Marathi, an inflectional language and even a parsed tree i.e a grammatical structure. We plug the morphological analyzer with statistical pos tagger and chunker to see its impact on their performance so as to confirm its usability as a foundation for NLP applications.

Keywords: Analyze , Inflection, Lexicon, Marathi morphology, Natural language Processing.

1. INTRODUCTION

1.1 Marathi morphology:

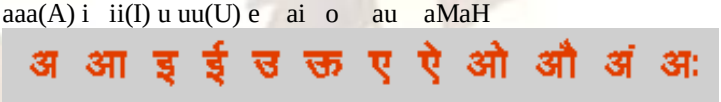
In linguistics, morphology[2] is the identification, analysis and description of the structure of a given
language's morphemes and other linguistic units, such as words, affixes, parts of speech, intonation/stress, or implied context (words in a *lexicon* are the subject matter of *lexicology*). Morphological typology represents a method for classifying languages according to the ways by which morphemes are used in a language —from the analytic that use only isolated morphemes, through the agglutinative ("stuck-together") and fusional languages that use bound morphemes (affixes), up to the polysynthetic, which compress lots of separate morphemes into single words.

While words are generally accepted as being (with clitics) the smallest units of syntax, it is clear that in most languages, if not all, words can be related to other words by rules (grammars). For example, English speakers recognize that the words *dog* and *dogs* are closely related — differentiated only by the *plurality morpheme* "-s", which is only found bound to nouns, and is never separate. Speakers of English (a fusional language) recognize these relations from their tacit knowledge of the rules of word formation in English. They infer intuitively that *dog* is to *dogs* as *cat* is to *cats*; similarly, *dog* is to *dog catcher* as *dish* is to *dishwasher*, in one sense. The rules understood by the speaker reflect specific patterns, or regularities, in the way words are formed from smaller units and how those smaller units interact in speech. In this way, morphology is the branch of linguistics that studies patterns of word formation within and across languages, and attempts to formulate rules that model the knowledge of the speakers of those languages.

1.2 The Alphabets
Marathi script consists of 16 vowels and 36 consonents making a total of 52 alphabets.

1.3 Vowels
The vowels are grouped in two groups. The first group consists of 12 vowels as follows:

aaa(A) i ii(I) u uu(U) e ai o au aMaH



The first 10 vowels are very widely used. The last two are less commonly used.

1.4 Linguistic Resources
The linguistic resources required by the morphological analyzer include a lexicon and inflection
rules for all paradigms. These are few linguistic resources:

1.4.1 Lexicon
In linguistics, the description of a language is split into two parts, the grammar consisting of rules describing correct sentence formation and the lexicon listing words and phrases that can be used in the sentences. The lexicon (or wordstock) of a language is its vocabulary. Statistically, most lexemes contain a single morpheme. Lexemes composed of multiple morpheme also known as compound words such as idiomatic expressions and colocations are also considered part of the lexicon. In practical applications, such as

language learning the lexicon is represented by a dictionary, which lists words alphabetically and provides definition.

1.4.2 Inflection Rules

Inflection rules [7] specify the inflectional suffixes to be inserted (or deleted) to (or from) different
positions in the root to get its inflected form. An inflectional rule has the format: <*inflectional Suffixes, morphosyntactic features, label*>. The element *morphosyntactic features* specifies the set of morphosyntactic features associated with the inflectional form obtained by applying the given inflection rule. Following is the exhaustive list of morphosyntactic features to which different morphemes get inflected:

1) Case: Direct, Oblique
2) Gender: Masculine, Feminine, Neuter ,Non-specific
3) Number: Singular, Plural, Non-specific
4) Person: 1st, 2nd, 3rd
5) Tense: Past, Present, Future
6) Aspect: Perfective, Completive, Frequentative, Habitual, Durative, Inceptive, Stative
7) Mood: Imperative, Probabilitive, Subjunctive, Conditional, Deontic, Abilitive, Permissive

1.5 Category Wise Morphological Formulation

The grammatical categories observed in Marathi include nouns, pronouns, verbs, adjectives, adverbs, conjunctions, interjections and postpositions. The morphemes belonging to different categories undergo different treatment.

1.5.1 Postposition Morphology

Paradigms of postpositions are created based on their linguistic behavior. They include case markers (vibhaktipratyay) and a class of postpositions called shabdayogiavyay. The latter are attached to singular and plural forms of nouns and pronouns. Some shabdayogiavyays exhibit specific behavior.

For example, some postpositions need to be written separately when they follow syllable (cyaa) (chya), which is a case marker. Some shabdayogiavyays can be suffixed with case markers
cao(che),caI,(chi),caa(cha),cyaa(chya).Some shabdayogiavyays can be composed of others.PostpositionshI(hI) andca(cha) can be attached before some shabdayogiavyays, but not before vibhaktipratyays. Some shabdayogiavyays can be attached to different oblique forms of verbs.

1.5.2 Noun Morphology

Changes due to the attachment of postpositions are different for singular and plural forms of nouns. The changed form of a noun to which such attachment is done, is called saamaanyaroop (oblique form) of that noun. For example, in morphological transformation of word rama(ram) to word ramaalaa(ramala), the samanyaroop of rama(ram) is ramaa(rama).

1.5.3 Pronoun Morphology

Exhaustive list of all possible (over 550) inflections of all pronouns is prepared because pronouns show very irregular behavior. The ratio of inflectional rules to actual forms in the case of pronouns is close to one. A pronoun has a specific single oblique form to which all shabdayogiavyays are attached

1.5.4 Verb Morphology

Aakhyaata Theory is the basis of verb morphology analysis. It systematically segments the verb forms into verb roots and terminating suffixes called Aakhyaatas. Aakhyaata represents information about TAM and GNP. They are named according to the phonemic shape such as taakhyaata, vaakhyaat and laakhyaata. A regular verb root generates over 80 forms. In addition to regular verbs, there are over 35 irregular verbs. The rules are represented in the form of tables.

1.5.5 Adjective Morphology

Adjectives are classified in inflectional and non-inflectional categories. Inflections result from gender, number and attachment of postpositions to the noun modified by such adjective. Table 2 shows a snapshot of inflectional rules. In the spellchecker, the root form is chosen as masculine form, from which other forms are generated.

Changing part in masculine form	Change		
	Feminine	Neuter	Oblique form
Aaa	[- I	e e	yaaya

Table 1. Adjective Morphology

When genitive case markers or some Shabdayogiavyays are attached to nouns, it produces adjectives. These forms are automatically covered in noun morphology.

1.5.6 Adverb, Conjunction and Interjections

This is an important class of part of speech, for which the rule-based approach proved to be appropriate. Attachment of postpositions to nouns, verbs and pronouns is one of the strategies of adverb formation. In addition, there are non-inflectional adverbs. The set of derived adverbs is automatically covered at the level of morphology of postpositions, nouns, verbs and pronouns. The list of all lexicalized adverbs is constructed. Similarly, all conjunctions and interjections are handled as a list since they are non-inflectional. When some

postpositions are attached to demonstrative pronouns, conjunctions are derived. These are handled at the level of rules for pronouns and postpositions.

Morphological analyzer forms the foundation for applications like information retrieval, POS tagging, chunking and ultimately the machine translation. Morphological analyzers for various languages have been studied and developed for years. But, this research is dominated by the morphological analyzers for agglutinative languages or for the languages like English that show low degree of inflection.

The proposed system, describes the morphemes, lexicon, type in English for Marathi word. It helps the person not knowing English language will be able to understand the English grammar in a better way.

2. RELATED WORK

Natural Language Processing [8]

The term "natural" languages refer to the languages that people speak, like English, Assamese and Hindi etc. The goal of the Natural Language Processing (NLP) group is
to design and build software that will analyze, understand, and generate languages that humans use naturally. The applications of Natural Language can be divided into two
classes

• Text based Applications: It involves the processing of written text, such as books,
newspapers, reports, manuals, e-mail messages etc. These are all reading based tasks.

• Dialogue based Applications: It involves human–machine communication like spoken language. Also includes interaction using keyboards. From an end-user's perspective, an application may require NLP for either processing natural language input or producing natural language output, or both. Also, for a particular application, only some of the tasks of NLP may be required, and depth of analysis at the various levels may vary. Achieving human like language processing capability is a difficult goal for a machine. The difficulties are:

• Ambiguity
• Interpreting partial information
• Many inputs can mean same thing

3. PROPOSED SYSTEM

Architecture design for analyzer

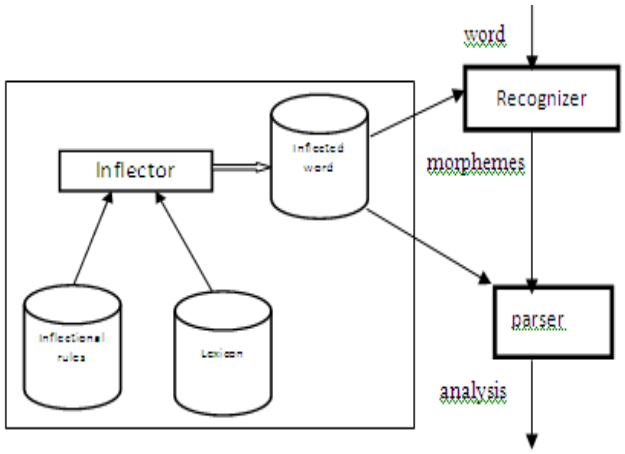


Fig1: architecture for Marathi morphological analyzer

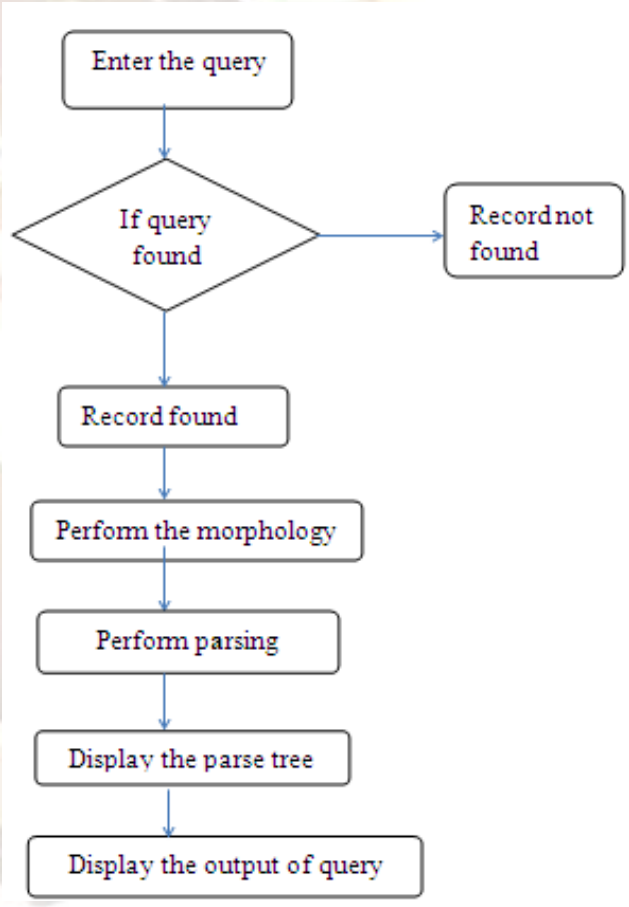


Fig.2: flow of query inflector

Step1: Enter the query or the word which morphemes are needed to be found. But we need to enter the word written into English language
Step2: Decision making
2.1If query found follow up the next steps.
2.2If query enter is not return to first step.
Step3: If record found display the records.
Step4: Perform the morphological analysis.

Step5: Perform the parsing.
Step5: Display the parsed tree. Display the morphemes found.
Step6: Stop.

4. EXPERIMENTAL RESULTS

In this paper, we present the morphological analyzer for Marathi which is official language of the state of Maharashtra (India). Marathi is the language spoken by the native people of Maharashtra. Marathi belongs to the group of Indo-Aryan languages which are a part of the larger of group of Indo-European languages, all of which can be traced back to a common root. Among the Indo-Aryan languages, Marathi is the southern-most language. All of the Indo-Aryan languages originated from Sanskrit.
The morphological analyzer takes the Marathi input then converts it into the English language. We gets its type, lexicon, morphemes.

4.1 Morphological Analyzer for Marathi

The formation of polymorphemic words leads to complexities which need to be handled during
the analysis process. FSMs prove to be elegant and computationally efficient tools for modeling
the suffix ordering in such words. However, the recursive process of word formation in Marathi
involves inflection at the time of attachment of every new suffix. The FSMs need to be capable
of handling them. Koskenniemi (1983) suggests the use of separate FSMs to model the orthographic changes. But, Marathi has a well devised system of paradigms to handle them. One of our observations led us to a solution that combines paradigm-based inflectional system with FSM for modeling. The observation was that, during the ith recursion only (i-1)th morpheme changes its form which can be handled by suitably modifying he FSM.

A. Example for morphological analysis:
1. Consider the word “Darashejari” (besides the door)
Equation (1) illustrates this process

dara ->daraa + shejari = darashejari
The parse tree is as follows:

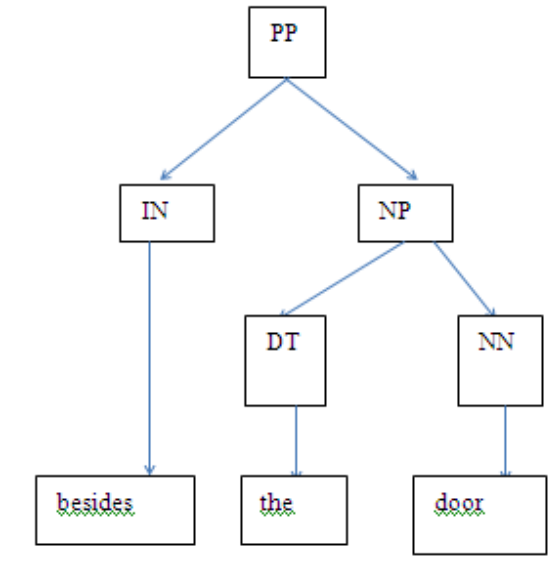


Fig3. Example1
IN: Preposition or subordinate
DT: Determiner
NN: Noun singular
2. consider the word “aayushyabharasathi” (for the life time)
Equation (2) illustrates this process
aayushya ->aayushya +bhara+sathi =aayushyabharasathi
The parse tree is as follows:

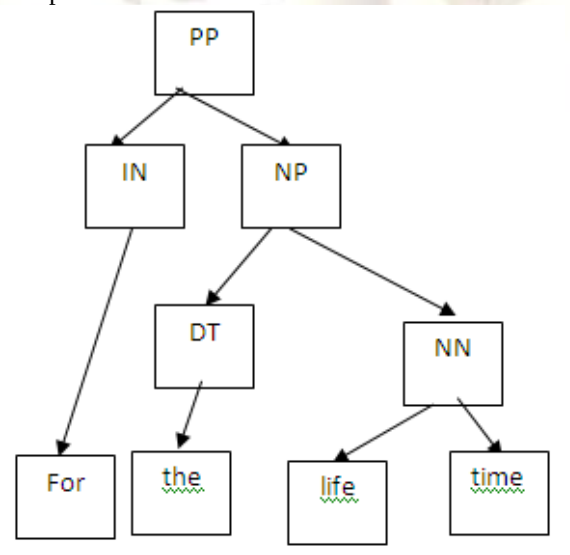


Fig4: Example2

5. CONCLUSION

We presented a high accuracy morphological analyzer for Marathi that exploits the regularity
in the inflectional paradigms while employing he Finite State Systems for modeling the language
in an elegant way. We gave detailed description of the morphological phenomena present in Marathi. The classification of postpositions and the

development of morphotactic FSA is one of the important contributions since Marathi has complex morphotactics. As a next step the morphological analyzer can be further extended to handle thederivation morphology and compound words. We plan todevelop a hybrid system using methodsto handle unknown words and to improve the overall accuracy of the system. In the meantime, more analysis will be addedto the system to cover aspects which might have eluded us so far. Even consideration of sentence also taken into account to develop the further models using this kind of analysis.

REFERENCES

[1] "*An improvised Morphological Analyzer for Tamil:A case of implementing the open source platformApertium"*. Parameswari K. Unpublished M.Phil. Thesis. Hyderabad:University of Hyderabad. 2009.

[2] *" Design and Implementation of a Morphology-based Spellchecker forMarathi, an Indian Language"*. Dixit, Veena, Satish Dethe, and Rushikesh K. Joshi. 2006.

[3] *"James Allen. Natural Language Understanding".* Pearson Education, Singapur, second edition, 2004.

[4] "*Natural Language Processing: A Paninian Perspective".* Bharati, Akshar, Vineet Chaitanya, and Rajeev Sanghal 1995.

[5] *" Natural Language Processing – A Paninian Perspective".* R. S. Akshar Bharati and V. Chaitanya,1995.

[6] *"Two-level Morphology: a general computational model for word-form recognition and production".* Koskenniemi, Kimmo 1983.

[7] *"A Paradigm-Based Finite State Morphological Analyzer for Marathi",* Pushpak Bhattacharyya.

[8] *"Natural Language Processing".* Utpal Sharma. Department of Computer Science and Information Technology, Tezpur University, Tezpur-784028, Assam ,India.