Morphological Analyzer forAffix Stacking Languages: A Case Study of Marathi

Raj Dabre[1] *Archana Amberkar*[1] *Pushpak Bhattacharyya*[1]
(1) Indian Institute of Technology Bombay, Mumbai-400076, India
prajdabre@gmail.com, amberkararchanaa@gmail.com,pb@cse.iitb.ac.in

ABSTRACT

In this paper we describe and evaluate a Finite State Machine (FSM) based Morphological Analyzer (MA) for Marathi, a highly inflectional language with agglutinative suffixes. Marathibelongsto the Indo-European familyand isconsiderably influencedby Dravidian languages. Adroit handling of participial constructions and other derived forms (*Krudanta*sand *Taddhita*s) in addition to inflected forms is crucial to NLP and MT of Marathi. We first describe Marathi morphological phenomena, detailing the complexities of inflectional and derivational morphology, and then go into the construction and working of the MA. The MA produces the *root word and the features*. A thorough evaluation against gold standard data establishes the efficacy of this MA.To the best of our knowledge, this work is the first of its kind on a systematic and exhaustive study of the Morphotactics of a suffix-stacking language, leading to high quality morph analyzer. The system forms part of a Marathi-Hindi transfer based machine translation system. The methodology delineated in the paper can be replicated for other languages showing similar suffix stacking behaviour as Marathi.

KEYWORDS: Marathi, Morphology, Derivational, Inflectional, Architecture, Finite State Transducer, Two-Level, Indian Language Technology.

1. Introduction

The number of Marathi speakers all over the world is close to 72 million[1]. Marathi uses agglutinative, inflectional and analytic forms. It displays abundant amount of both derivational (wherein attachment of suffixes to a word form changes its grammatical category) and inflectional morphology. About 15% of the word forms are participial forms known as *Krudanta*s, which resultfrom the influence of Dravidian languages. Traditional grammars of Marathi classify the derived forms in Marathi into two categories- *Krudanta*s and *Taddhita*s. *Krudanta*sare the adjectives, adverbs and nouns derived from verbs, while *Taddhita*s are nouns, adjectives and adverbs derived from words of any category other than verb. This is also accompanied by inflectional processes which help lend the words features of gender, number, person, case, tense, aspect and modality (the latter 3 for verbs only).

1.1. Related work

The first MA for Marathi used a very naïve suffix stripping approachpropounded by Eryiğit and Adalı, (2004). This neither had the ability to handle the stacking of suffixes which might involve orthographic changes at morpheme boundaries, nor could it indicate spelling mistakes and thus was discarded. The need for a mechanism to handle both inflectional and derivational morphology was felt, andwe adopted the Finite State Transducer (FST) based approach that allows specification of legal morpheme sequencesof both inflectional and derivational kind. We thus used a two level morphological analysis model (Oflazer, 1993; Kim *et al.*, 1994),including a Morphological Parser(Antworth, 1991).Dixit *et al.* (2006) implemented a Marathi spell-checker, which is an inherent part of our MA. Bapat*et al.* (2010) had developed aFST based MA which handled the derivational morphology of verbs, and Bhosale*et al.* (2011) showed that the inclusion of this MA helps improvethe translation quality. We extended the work of Bapat*et al.* to other grammatical categories, thereby increasing the coverage of Marathi morphological phenomena.

2. Morphological phenomena in Marathi

We first describe inflectional morphology. Nouns in Marathi are inflected for gender, number and case; adjectives areinflected for gender and number, pronouns for gender, number, case and person. The noun आंबा{*aambaa*}{*mango*}is masculine. Its direct singular and plural forms are: आंबाand आंबे{*aamb**e***}{*mangoes*}respectively. Its oblique singular and plural forms are: आंब्या{*aamb**yaa***}and आंब्यां{*aamb**yaan***}respectively. Verbs in Marathi are inflected for person, number and gender of the subject alone or that of both the subject the object of the verb and also for tense, aspect and mood.Marathi has three genders (*masculine*, *feminine* and *neuter*), two numbers (*singular* and *plural*), eightcases (*nominative, accusative, instrumental, dative, ablative, genitive, locative* and *vocative*) and three persons (*first*, *second* and*third*). Different linguists give different typologies for the tenses, aspects and moods in Marathi. We have followed the typology given by Damle, M.K. (1970).We have also followed the linguistic analyses in the book of Dhongde and Wali (2009).

[1] http://en.wikipedia.org/wiki/List_of_Indian_languages_by_total_speakers

2.1. Derivational Morphology:

In the derivational process, a derivational morpheme is affixed to the word stem (the form a root takes when a derivational morpheme is attached to it), in order to add meaning to it and thereby derive a new word. The resulting word may or may not be of the same grammatical category. For example, Marathi has a derivational morpheme-"पणा"{*panaa*}, which isattachedto adjectives like "मूर्ख"{*moorkh*} {*foolish*}, in order to derive nouns like"मूर्खपणा"{*moorkhapanaa*} {*foolishness*}. Marathi has many such derivational morphemes.

Another important feature of Marathi is its set of participles, which are derived by attaching derivational morphemes to verbs. These participles indicate tense, aspect, voice, mood in addition to gender and number features. For e.g. "येणारा"{*yenaaraa*} {*coming*} is a masculine, singular present participle form, while"गेलेला" {*gelelaa*} {*has gone*} is a masculine, singular past participle. Most of these participles, in addition toinfinitive forms are currently handled byour MA. It also handles the extraction of mostof the derivational morphemes that attach to verbs and a few that attach to nouns, adjectives and adverbs. Handling Derivational Morphology is important as it requires only base forms to be stored thereby reducing the lexicon size.We now describe some of the morphological complexities and methods of handlingthem.

2.2. Complexities in handling Inflectional Morphology

1. When a genitive case marker is attached to a noun or a pronoun, the resulting form holds the gender and number information of boththe base noun and the genitive case marker. Forexample, in the word "मुलींचा" {*muleenchaa*} {*of the girls*}, the stem "मुली" {*muleen*} has the features feminine, plural, while the suffix "चा"{*chaa*} has the features masculine, singular. Thus, the morphological analysis of this form should consist of two feature structures- one for the stem and the other for the genitive suffix. Currently, we obtain a selective combination of both.
2. Pronouns take all cases except the vocative. However in case of pronouns, all cases are not overtly marked.For example, the instrumental case is not overtly marked in case of first and second person pronouns ("मी"{*mee*} {*I/me*} and"तू" {*tu*} {*you*}). Marathi also has demonstrative pronouns which are same as third person pronouns. However, when these pronouns occur as demonstrative pronouns, they do not take case postpositions. Distinguishing between pronouns and demonstratives becomes difficult (a property of almost all Indo-Aryan languages). For instance, "त्यामुलाने"{*tyaamulaane*} {*that boy* (*did*): ergative form}. We handle these by special entries in the repository of inflected forms (REPO) (see next section).
3. Spatial and temporal adverbs like "आता" {*aata*} {*now*}, which act as nouns, can take some case markers like "चा"{*chaa*} {*of*}to give "आताचा"{*aataa-chaa*} {*now-of*}for which the Marathi MA uses the type NST (Noun of Space and Time). We create special paradigms (Bapat*et al.*, 2010) for NSTs.
4. There are morphemes that indicate a few features of the agent of the verb and a few features of the object of the verb. For example, the morpheme "लीस"{lees} in"खाल्लीस" {*khallees*} {*eaten*} in addition to indicating the perfective aspect, indicates that the agent of the verb is in singular and second person, while the object of the verb is feminine, singular and third

person. In such cases, the morphological output should ideally have two separate feature structures- one for the agent and the other for the object.

5. Stacking of two or more suffixes is very common. Consider the example, "जाणाऱ्यानेसुद्धा"{*jaanaaryanesuddhaa*} {*the one going also*(*instrumental*)} {जा + णारा + ने + सुद्धा}.The root is the verb "जा"{*jaa*} {*go*} attached with three suffixes "णाऱ्या" {*naarya*}, ने{*ne*}and "सुद्धा"{*suddha*} {*also*} respectively. Here "णाऱ्या"has "ने" as suffix which in turn has "सुद्धा" as suffix. The finite state approach (next section) for morphological analysis helps in solving this.
6. There are a few pairs of morphemes that have similar orthographical shape, and the stems to which these morphemes are attached are orthographically similar too. Thus, the resulting inflected/derived forms are orthographically similar, but have two different meanings. For example, there are two morphemes represented by the letter "त" {*ta*}, one of which denotes habitual past and the other,imperfective aspect. Thus, attached to a verbal root like "फिर" {*fir*} {*to wander*}, these two suffixes produce two similar forms- "फिरत"{*phirat*} {(*they*) *used to wander*} and "फिरत"{*phirat*} {*wandering*}. In such cases, the Morphological Analyzer should be able to produce both the analyses. Once again, the finite state approach helps.

2.3. Complexities in handling Derivational Morphology

1. Base roots may have multiple forms (called stems)depending on which derivational morpheme is attached to them. For example, the cardinal "पन्नास" {*pannaas*} {*fifty*}, when attached with the derivational morpheme "वा", takes the stem "पन्नासा" {*pannaasaa*}. However, when attached with the derivational morpheme "दा" {*da*}, the same cardinal takes the stem "पन्नास" {*pannaas*}. In such cases, we need separate Suffix Replacement Rules (SRRs) (Bapat*et al.*, 2010)for each derivationalmorpheme.
2. Some of the derivational morphemes like "पणा" {*panaa*}, "दा" {*daa*} are highly productive, as they are attached to all members of a particular grammatical category like nouns. However, some derivational morphemes are attached to only some particular semantic classes within a grammatical category. For instance "भर" {*bhar*} is attached to only nouns, and to only those nouns which indicate places or containers- "देश"{*desh*} {*country- a place*},"वाटी"{*vaati*}{*bowl- a container*}. The resultant form for "देश" is "देशभर"{*deshbhar*} {*throughoutthecountry*}.For such nouns, we need to create special paradigms.

3. Architecture and Working of the Morphological Analyzer

The Marathi Morphological Analyzer is fully rule-based and thus relies on string manipulation and file lookup. It requires two main resources, namely,a FST (Finite State Transducer) and a REPO (Repository of Inflected Forms), generated using an Inflector and SFST[2] (Stuttgart Finite State Transducer) compiler, which are explained below. These are in turn generated by the basic resources; namely, the monolingual lexicon, the suffix replacement rules (SRRs),the special word forms repository,the verb suffix (for *Krudanta*s) list(Bapat*et al.*, 2010)and morphology rules (Morphotactics).

[2]http://www.ims.uni-stuttgart.de/projekte/gramotron/SOFTWARE/SFST.html

3.1. Tools and Resources

3.1.1. FST (Finite State Transducer)

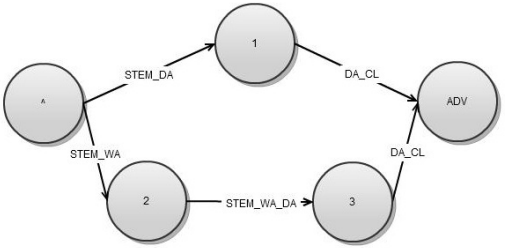

Figure 1 - FST for deriving Adverbs from Cardinal

The rules which specify the legal sequences of word-forming morphemesin Marathi arecalled Morphotactics.These rules constitute a Finite State Transducer. This helps identify incorrectly written words efficiently and allow for easy word segmentation. An example of a rule would be: "*ADJ = ADJ_OF SSY?*" This means that an adjective (ADJ) can be formed by a sequence of oblique form adjective (ADJ_OF) and an optional suffix (SSY). The question mark indicates optionality. This is a FSM rule.We thus work with parts instead of wholes.Here ADJ_OF and SSY are inflectional types. To understand this better consider the FST in figure 1above.

The FST describes the derivation of adverbs from cardinals. The adverb "वीसदा "{*veesdaa*} {*twenty times*} is derived by suffixing "दा "{*da*} {*time*(*s*)} (which comes under DA_CL) to the cardinal "वीस "{*vees*} {*twenty*} (which comes under STEM_DA), while the adverb "विसाव्यांदा " {*visaavyaandaa*} {*twentieth time*} is derived by suffixing "दा "{*da*} {*time*(*s*)}to the stem"विसाव्यां"{*visaavyaan*}ofthe ordinal "विसावा" {*visaavaa*} {*twentieth*} where "वा" {*vaa*} becomes "व्यां" {*vyaan*} (which comes under STEM_WA_DA). Here the ordinal "विसावा " {*visaavaa*} {*twentieth*} is derived by suffixing "वा "to the cardinal "वीस "{*vees*} {*twenty*} (which comes under STEM_WA). DA_CL represents the derivational suffix "दा" which cannot be followed by any other suffix. STEM_DA is the cardinal stem and STEM_WA_DA is the ordinal stem deriving suffix "व्यां"to which the suffix "दा" is attached. STEM_WA is the cardinal stem to which the suffix "वा" is attached. There are close to a 100 rules. We add more rules to handle more complex forms.

We use Stuttgart University's SFST (Stuttgart Finite State Transducer) compiler which takes the categorised inflected forms (in files) and the Morphotactics to give the transducer file, an augmented Finite Automaton (FA) transition table, called the Morphotact file.We chose SFST as it enjoys the ease of specifying Morphotactics. Alternatives like HFST (Helsinki Finite State Transducer) and FOMA also exist.

3.1.2. Repository of Inflected Forms (REPO)

After undergoing inflection, using an Inflector, which applies SRR'sto the words in the lexicon, all inflectional forms with their root words and features (gender, number, etc.) are stored in a single flat file called as the REPO file. Separate files for each inflectional type containing the inflected morphemes of that type are also created which are used for the generation of the FST.The format of this file is: <inflectional type>;<inflected word>; <root word-1, feature list-

1#root word-2, feature list-2#...#root word-n, feature list-n>. An example for "**महाबळेश्वर**" {*mahabaleshwar*}{*the god of great strength*}is<*DF*>;<**महाबळेश्वर**,*n,n,sg,,,,,d*#**महाबळेश्वर**,*n,m,sg,,,,,d*#**महाबळेश्वर**,*n,m,pl,,,,,d*>.

3.2. Morphological Processing

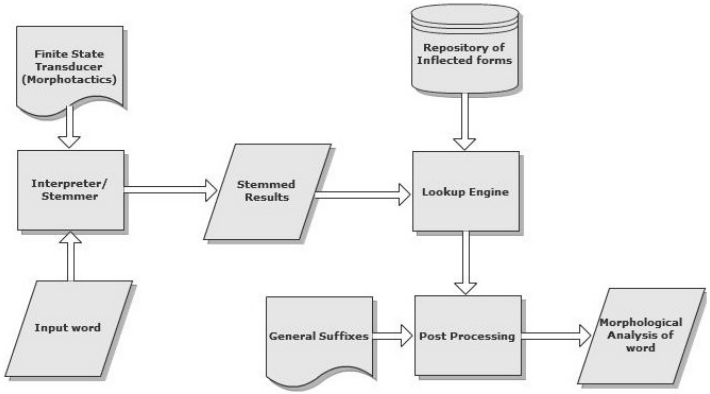

Figure 2 - Morphological Processing Flow

The flow of processing is in figure 2 above.There are 3 main components:the FST interpreter/Stemmer (level 1), a lookup engineand a post processing unit (level 2).An auxiliary support list of suffixes is also used.

3.2.1. FST interpreter / Stemmer / Segmenter

The interpreter (our Java equivalent of SFST interpreter) takes the input word and gives the morphemes it contains. As such this is a Stemmer orSegmenter. It uses the transition table of the FST and gives the output in the form: <input word>: morpheme-1 <category-1> morpheme-2 <category-2> ….. morpheme-n<category-n>. The first morpheme is the Stem.There is a possibility that a word may be stemmed in more than one way because it could bea direct form of a word or a morphologically complex word witha root andsuffix(es). Anexample for"हलवा" {*halawaa*}:

1. हलवा: An inflected form withthe imperative suffix "ा" is attached to the verbal root "हलव" {*halaw*} {*to shake*} .
2. हलवा: A direct form of a noun referring to a dessert.

3.2.2. Lookup engine / Parsing

This unit accepts stemmed results to give intermediate morphological analyses. This and the next stage constitute Morphological Parsing (MP). We currently perform MP for all inflectional morphemes and for the derivational morphemes that attach to verbs.First a hash table of the REPO file, by using the inflected form and the word form category as the joint index, is constructed. The first morpheme and its category are then used on this hash structure to obtain its root form and its features. This is followed by the most crucial-*Krudanta* processing. If the stemmer detects a *Krudanta* suffix, the lookup gets it from the hash table and modifies the

features of the feature list using those of the *Krudanta*suffix. Otherwise the following suffixes (if any) are either case markers or postpositions or non*Krudanta*derivational morphemes which we append to the feature list.

Verbshave additional features namely "*Krudanta* Type" and "*Krudanta* Case Marker/Suffix". An example for "**धावणारा**" {*dhaavnara*}{*runner*}, an adjective, would be:**धावणारा**<*fsaf(feature structure abbreviated form*)=**'धाव**,*v,m,sg,,d,***णारा**,**णारा**' *tense='' aspect='' mood='' kridanta_type='nara' kridanta_cm=***'णारा**'>.

3.2.3. Post Processing

A word can be stemmed in multiple ways and hence the resulting duplication of features that happens is eliminated in this unit. Some Marathi specific cases which cannot be handled by rules are also handled here. The final part of this is handling unrecognised words. A word will not be stemmed if either it was not entered in the lexicon or there is a spelling mistake or there are no rules to handle it. It is important to identify the suffix as it shows relations between words and must be translated even if the word it is attached to is unknown or unidentifiable. This is mostly for foreign words. The word is matched against the list of suffixes and the one identified is extracted. There will be no linguistic features associated with it.

Builders of Morphological Analyzers, especially, for Indian (and other similar) languages can use our framework effectively. Our Java based stemmer can completely stem/segment and parse around 50000 tokens in 8-10 seconds.The end result of all this processing is the minimally sufficient morphological analysis of the input word. In the next section we present the methods for evaluation of our MA and the results.

4. Evaluation

We have two measures of quality, namely, accuracy and usability. We prepared Gold Standard Data of 101 sentences with a total of 1341 tokens/words. We comparedthe outputs of our MA with the gold standard data.For analysis, each word is put into one of 6 different categories.Table1belowdescribes these categories and also gives the results of our evaluation.

Analysis number	Analysis category	Number of words	Percentage
1	Same analysis: Identical to gold	968	72.18
2	Spurious analysis: Extra analyses along with gold	161	12.00
3	Missing analysis: Missing analyses from gold	66	4.92
4	Missing and spurious analysis: Missing and extra analyses from gold	70	5.21
5	Completely spurious analysis: Totally incorrect	2	0.14
6	No output: No analysis given	74	5.51
	Total	1341	

Table 1- Distribution of Analysis types

Our formula for accuracy is:

$$Accuracy = (Number\ of\ type\ 1\ analyses)/(Total\ number\ of\ words)$$

This gives us an accuracy of 72.18%.This proportion of wordsis perfectly analyzed and their analyses were correct, complete and useful in terms of the *root and features* information. The analyses of words under type 2, 3 and 4 give at least one usable analysis (feature list including root and suffix), which is mostly sufficient for NLP applications. Analysis belonging to categories 5 and 6 are totally useless.

The formula for usability is:

$$Usability = Number\ of\ type\ 1,2,3\ and\ 4\ analyses/Total\ number\ of\ words$$

This brings our usability score to 94.33%. Out of 273 derivational morphemes,265 (97.06%) were correctly segmented and 237 (86.81%) of them were correctly parsed. Our parsing of derivational morphemes needs more work, as only 237 (89.43%) out of 265 recognised are correctly parsed.

4.1. Error Analysis

Errors found in the MA output are of two types- errors of commission (false positives) and errors of omission (false negatives). Errors of commission which occur due to wrong entries and overgenerating rules in the lexicon grammatical rules list, respectively, are solved by modifying the entries and rules.Errors of omission which occur when necessary entries and rules are not made in the lexicon and grammatical rules list, respectively, are solved by adding the missing entries and rules.

Conclusions and Future work

We described the construction of a morphology analyzer for Marathi, which can be adapted for other languages that do suffix stacking. The Morphotactics have to be carefully captured- all generalities and exceptions included, after which standard FSM type tools can be harnessed to perform the analysis. The lexicon needs to be exhaustive and rich in morphosyntactic information. Our MA for Marathi has the ability to handle inflectional and derivationalmorphologyforalmost all of the grammatical categories.In future work, the parsing of derivational morphemes for categories other than verbs needs to be handled.We also need to adopt the suffix stripping approach where the FST approach fails, thereby leading to a hybrid MA.In the context of translation, the influence of derivational morphology needs to be investigated.Multiword and compounds form another area of investigation.

References

Damale, M. K. (1970). *ShastriyaMarathiiVyaakarana*.Deshmukh and Company, Pune, India.

Koskenniemi, Kimmo(1983). *Two-level Morphology: a general computational model for word-form recognition and production*. University of Helsinki, Helsinki.

Antworth, E. L. (1990). *PC-KIMMO: A Two level Processor for Morphological Analysis.*Occasional Publications in Academic Computing.Summer Institute of Linguistics, Dallas, Texas.

Deok-Bong, Kim., Sung-Jin, Lee., Key-Sun, Choi and Gil-Chang, Kim(1994). *A two levelMorphological Analysis of Korean.* In Conference on Computational Linguistics (COLING), pages 535–539.

Bharati, Akshar.,Chaitanya,Vineet and Sanghal, Rajeev(1995). *Natural Language Processing: A Paninian Perspective*. Prentice Hall, India.

Eryiğit, Gülşen and Adalı,Eşref(2004).*An Affix Stripping Morphological Analyzer for Turkish.*In IASTED International Multi-Conference on Artificial Intelligence and Applications. Innsbruck, Austria, pages 299–304.

Dixit,Veena.,Dethe,Satish and Joshi, Rushikesh K. (2006). *Design and Implementation of a Morphology-based Spellchecker for Marathi, an Indian Language.* In Special issue on Human Language Technologies as a challenge for Computer Science and Linguistics. Part I. 15, pages 309–316. Archives of Control Sciences.

Dhongde and Wali(2009).*Marathi.*JohnBenjamins Publishing Company, Amsterdam, Netherlands.

Bapat,Mugdha.,Gune,Harshada andBhattacharyya,Pushpak(2010). *A Paradigm-Based Finite State Morphological Analyzer for Marathi.*Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), pages 26–34,the 23rd International Conference on Computational Linguistics (COLING), Beijing, August 2010

Bhosale, Ganesh.,Kembhavi,Subodh.,Amberkar,Archana.,Mhatre,Supriya.,Popale,Lataand Bhattacharyya,Pushpak(2011).*Processing of Participle* (*Krudanta*) *in Marathi.* In International Conference on Natural Language Processing (ICON 2011), Chennai, December, 2011.