A Paradigm-Based Finite State Morphological Analyzer for Marathi

Mugdha Bapat **Harshada Gune** **Pushpak Bhattacharyya**

Department of Computer Science and Engineering,
Indian Institute of Technology Bombay

{harshadag,mbapat,pb}@cse.iitb.ac.in

Abstract

A morphological analyzer forms the foundation for many NLP applications of Indian Languages. In this paper, we propose and evaluate the morphological analyzer for Marathi, an inflectional language. The morphological analyzer exploits the efficiency and flexibility offered by finite state machines in modeling the morphotactics while using the well devised system of paradigms to handle the stem alternations intelligently by exploiting the regularity in inflectional forms. 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.

1 Motivation and Problem Definition

A highly inflectional language has the capability of generating hundreds of words from a single root. Hence, morphological analysis is vital for high level applications to understand various words in the language. 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. Though agglutinative languages show high morpheme per word ratio and have complex morphotactic structures, the absence of fusion at morpheme boundaries makes the task of segmentation fluent once the model for implementation of morphotactics is ready. On this background, a morphological analyzer for highly inflectional language like Marathi which has the tendency to overlay the morphemes in a way that aggravates the task of segmentation presents an interesting case study.

Eryiğit and Adalı (2004) propose a suffix stripping approach for Turkish. The rule based and agglutinative nature of Turkish allows the language to be modeled using FSMs and does not need a lexicon. The morphological analyzer does not face the problem of the changes taking place at morpheme boundaries which is not the case with inflectional languages. Hence, although apprehensible this model is not sufficient for handling the morphology of Marathi.

Many morphological analyzers have been developed using the two-level morphological model (Koskenniemi, 1983) for morphological analysis. **(**Oflazer, 1993; Kim et al., 1994) have been developed using PC-Kimmo (Antworth, 1991), a morphological parser based on the two-level model. Conceptually, the model segments the word in its constituent parts, and accounts for phonological and orthographical changes within a word. While, the model proves to be very useful for developing the morphological analyzers for agglutinative languages or the languages with very less degree of inflection, it fails to explicitly capture the regularities within and between paradigms present in the inflectional languages. Marathi has a well defined paradigm-based system of inflection. Hence, we decided to develop our own model which works on the similar lines of PC-Kimmo (Antworth, 1991) but exploits the

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

usefulness of paradigm-based inflectional system.

Bharati et al. (2004) propose a paradigm based algorithm for morphological analysis of Hindi, an inflecting language. 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.

Dixit et al. (2006) developed a morphological analyzer with a purpose of using it for spell checking. Though their analyzer successfully analyzes the words with a single suffix, its scope is restricted to the handling of only first level suffixes.

1.1 Our Approach

In this paper, we present the morphological analyzer for Marathi which is official language of the state of Maharashtra (India). With 90 million fluent speakers worldwide, Marathi ranks as the 4th most spoken language in India and the 15th most in the world. The methodology is based on the use of paradigm-based inflectional system combined with finite state machines (FSMs) for modeling the morphotactics. To the best of our knowledge, such an approach has never been tried out for Marathi. The crux of the system lies in the detailed study of morphosyntactic phenomena, the morphotactic structure of the language and the use of paradigm-based inflectional system.

The approach can be used for other inflectional languages by developing the resources like language specific inflection rules and the FSM that models the morphotactics for the language.

1.2 Marathi Morphology

Marathi is a morphologically rich language. It is highly inflectional and also shows derivation to a high degree. Like other synthetic languages, Marathi morphological analysis faces some well-known challenges. Words contain multiple morphemes fused together in such a way that, it becomes difficult to segment them. A single morpheme contains a bunch of grammatical attributes associated with it which creates a challenge for morphological parsing. A single root is capable of generating hundreds of words by combining with the other morphemes.

The complexity involved in the formation of a polymorphemic word can be better illustrated using an example. Consider the word देवासारख्याला {devaasaarakhyaalaa} (to the one like the god). The nominal root 'देव' {deva} (god) gets inflected to the oblique case, singular form 'देवा' {devaa} which is then followed by the adjectival suffix 'सारखा' {saarakhaa} (alike). This derives the adjective 'देवासारखा' {devaasaarakhaa} (the one like the god) which then starts behaving like a noun. This noun on getting inflected to the oblique case, singular form देवासारख्या {devasaarakhyaa} is followed by the case marker ला {laa} (to). This gives the word देवासारख्याला {devaasaarakhyaalaa} (to the one like the god). Equation 1 illustrates this process.

$$(deva \text{ "}\rightarrow\text{" } devaa) + saarakhaa = devaasarakhaa$$
$$(devasaarakhaa \rightarrow devaasarakhyaa) + laa = devaasaarakhyaalaa$$

Equation 1. Formation of देवासारख्याला {devaasaarakhyaalaa} (to the one like the god)

This suggests that the process of formation of polymorphemic words is recursive in nature with inflection taking place at every level of recursion.

Section 2 discusses the design of the morphological analyzer which tries to overcome the problems discussed above with respect to Marathi language. Sections 3 and 4 discuss the linguistic resources and the processing of words belonging to various categories respectively. Sections 5 and discuss the classification of suffixes and development of automata based on this classification respectively. Section 7 briefs on the experimental setup and the results.

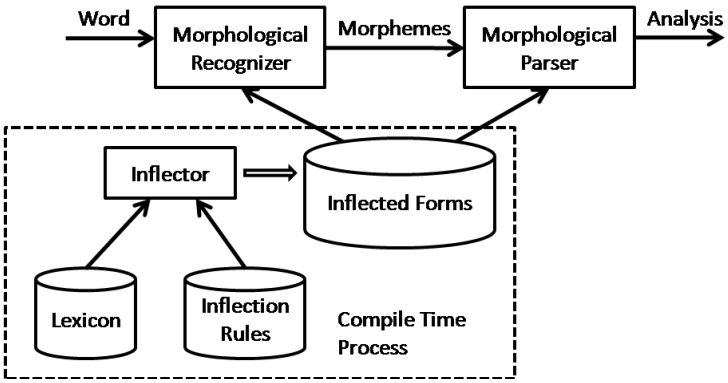

Figure 1. Architecture of Marathi Morphological Analyzer

2 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 i^{th} recursion only $(i-1)^{th}$ morpheme changes its form which can be handled by suitably modifying the FSM. The formation of the same word devaasaarakhyaalaa described above can be viewed as illustrated in Equation 2**.**

$$(deva \rightarrow devaa) + saarakhaa = devaasarakhaa$$
$$(deva \rightarrow devaa) + (saarakhaa \rightarrow sarakhyaa) + laa = devaasaarakhyaalaa$$

Equation 2. Simulating the formation of देवासारख्याला {devaasaarakhyaalaa} (to the one like the god)

Generalizing the word formation process we arrived at the formulation specified by Equation 3.

$$Polymorphemic\ word = (inflected\_morpheme1) + inflected\_morpheme2 + ..$$

Equation 3. Formulation of Polymorphemic Word Formation

This requires a morphotactic FSM which is aware of the inflected forms of morphemes in addition to the actual morphemes to handle the above recursive process of word formation. We use the paradigm-based system to generate the inflected form of the morphemes and feed them to the FSM. Figure 1 shows the architecture of the morphological analyzer based on this philosophy.

Inflector inflects all morphemes in the lexicon using the inflection rules associated with the paradigms to which they belong.

Given a word, Morphological Recognizer recognizes the constituent morphemes in their inflected forms using finite state machine that models the morphotactics. For example, the output of the Morphological Recognizer for the word devaasaarakhyaalaa is devaa + saarakhyaa + laa. Morphological Parser outputs per morpheme analysis of the word using the morphemes recognized by the Morphological Recognizer.

3 Linguistic Resources

The linguistic resources required by the morphological analyzer include a lexicon and inflection rules for all paradigms.

3.1 Lexicon

An entry in lexicon consists of a tuple <*root, paradigm, category*>. The *category* specifies the grammatical category of the *root* and the *paradigm* helps in retrieving the inflection rules associated with it. Our lexicon contains in all 24035 roots belonging to different categories.

3.2 Inflection Rules

Inflection rules 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: 1^{st}, $2^{nd,}$ 3^{rd}
5) Tense: Past, Present, Future
6) Aspect: Perfective, Completive, Frequentative, Habitual, Durative, Inceptive, Stative
7) Mood: Imperative, Probabilitive, Subjunctive, Conditional, Deontic, Abiltive, Permissive

The *label* specifies the morphotactic class to which the inflected form (generated by applying the inflection rule) belongs. It is used by the Morphological Recognizer.

4 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.

4.1 Noun Morphology

Marathi nouns inflect for number and case. Postpositions get attached to the oblique forms of the nouns (known as stems) to show their relationship with other words in the sentence. A single stem is used for the attachment of all postpositions which makes nominal morphology absolute economic in nature. For example various forms of the word दार {daara} (door) are दारास {daaraasa} (to the door), दाराने {daaraane} (by the door), दाराशेजारी {daaraashejarii} (besides the door). Please note that the same stem दारा {daaraa} is used for the attachment of various postpositions.

Depending upon their ending, gender and the inflectional patterns, the nouns in Marathi can be classified into various paradigms. A paradigm is a complete set of related inflectional forms associated with a given root. All words that share the similar inflectional forms fall in the same paradigm. Table 1 presents the paradigm दार {daara} (door).

		Case	
		Direct	Oblique
Number	Singular	दार {daara}	दारा {daaraa}
	Plural	दारे {daare}	दारां {daaraaN}

Table 1. Paradigm Table for दार {daara} (door)

कापड {kaapaDa} (cloth), पान {paana} (leaf), पुस्तक {pustaka} (book), कपाट {kapaaTa} (cupboard) are the few nouns that fall into this paradigm.

Every paradigm has a set of inflection rules associated with it one corresponding to every inflectional form of the word. A noun has four inflectional forms each one corresponding to a case-number pair. Hence, every paradigm has four inflectional rules associated with it.

An inflectional rule for Marathi consists of a tuple specifying the inflectional suffixes that should be inserted and deleted from ultimate and penultimate position of the root. Table 2 lists the inflectional suffixes that collectively form an inflectional rule.

The procedure to obtain the inflected form of the given root R belonging to paradigm P by applying the inflectional rule I <UD, UI, PUD, PUI> is as follows:

i. R =R - PUD
ii. R = R + PUI
iii. R = R – UD
iv. R = R + UI

Suffix	**Description**
Ultimate Deletion	Suffix to be deleted from the ultimate position of the root

(UD)	
Ultimate Insertion (UI)	Suffix to be inserted at the ultimate position of the root
Penultimate Deletion (PUD)	Suffix to be deleted from the penultimate position of the root
Penultimate Insertion (PUI)	Suffix to be inserted at the ultimate position of the root

Table 2. Suffixes in an Inflectional Rule

For a given word, even if a single rule out of the four is different from the set of available paradigms, a new paradigm needs to be created. Table 3 shows the paradigm भक्त {bhakta} (devotee). Note that, the only difference between the two paradigm tables is in the direct case plural form.

		Case	
		Direct	Oblique
Number	Singular	भक्त {bhakta}	भक्ता {bhaktaa}
	Plural	भक्त {bhakta}	भक्तां {bhaktaaN}

Table 3. Paradigm Table for भक्त {bhakta} (devotee)

In this way, our lexicon contains 16448 nouns categorized into 76 paradigms. Out of the 76 paradigms, 30 correspond to feminine gender, 29 to masculine and 17 to neuter gender. This set of paradigms includes three null paradigms, one corresponding to each gender. In modern Marathi, the stem of the proper nouns or foreign words transliterated in Marathi is same as the root. In short, postpositions can be directly attached to these roots without any modification. Such nouns belong to the null paradigm.

4.2 Postposition Morphology

Postpositions follow the stems of nouns and pronouns. Postpositions in Marathi can be broadly classified into case markers and shabdayogi avyayas. Shabdayogi avyayas show the relationship of nouns and pronouns with the other words in the sentence while deriving the adjectives or adverbs in most of the cases. Depending upon the category of the word derived by them they are classified as adjectival and adverbial suffixes respectively. We have 142 postpositions listed in our lexicon.

4.3 Classification of Postpositions

The first step towards defining the morphotactics of a language is the classification of various suffixes into classes depending upon the morphemes they can follow and the morphemes that can follow them. Given the list of 142 postpositions, we carefully examined each one to understand its morphotactic behavior and came up with the classification of Marathi postpositions as presented in the Table 4.

Class	**Ordering Rules**	**Example**
Case markers	They can follow any oblique form. No other suffixes can follow them.	ने {ne} (by)
Adjectival Suffixes	They can follow an oblique form of a root. Since they derive an adjective, they can be followed by any other suffixes.	सारखा {saarakhaa} (alike
Possessive case marker	It can follow any oblique form. It can be followed by any other suffixes.	चा {chaa} (the one belonging to something)
Closing suffixes	They can follow any oblique form. No other suffixes can follow them.	पेक्षा {pekshaa} (instead of)
चा {chaa} adjectival suffix	It can follow Peculiar NSTs and Nearly closing postpositions. It can be followed by case markers.	चा {chaa} (the one)
Peculiar	They can follow any ob-	जवळ

NSTs	lique form. They can be followed only by Exclusive postpositions and चा {chaa} adjectival suffix.	{jawaLa} (near)
Exclusive postpositions	They can follow peculiar NSTs. They close the word.	ई {ii} (inside)
Nearly closing postpositions	They can follow oblique forms of nouns and pronouns. They can be followed by चा {chaa} adjectival suffix.	पर्यंत {paryanta} (uptil)
Shuddha-shabdayogi avyayas	They can follow almost any morpheme except oblique forms of nouns. They can be followed by some postpositions. But, this behavior is quite irregular and needs more investigation. In most of the cases, these suffixes close the word. Hence, we consider them to be occurring only at the end of the word.	च {cha} (only)

Table 4. Classification of Postpositions

4.4 Verbs

The verbs inflect for gender, number and person of the subject and the direct object in a sentence. They also inflect for tense and aspect of the action as well as mood of the speaker in an illocutionary act. They may even undergo derivation. Further discussion on verbal morphology will be based on Aakhyaata theory (inflection) and Krudanta theory (derivation) (Damale, 1970). Our lexicon contains 1160 verb roots classified into 22 paradigms.

Aakhyaata Theory forms the basis of verbal inflection in Marathi. *Aakhyaata* (*आख्यात*) refers to tense, aspect, and mood. Aakhyaata is realized through an aakhyaata suffix which is a closing suffix, attached to the verb root. There are 8 types of aakhyaatas named after the phonemic shape of the aakhyaata suffix. Associated with every aakhyaata are various aakhyaata-arthas which indicate the features: tense, aspect and mood. An aakhyaata may or may not agree with gender. There are around 80 Aakhyaata suffixes in Marathi.

Krudanta Theory forms the basis of verbal derivation in Marathi. Krudanta refers to the word ending in a krut-pratyaya (a suffix which refers to an action). Krut-pratyayas are attached at the end of verbs to form non-infinitive verb forms. These forms usually belong to one of the categories: noun, adverb or adjective. They contribute to the aspect of the verb from which they are derived. We cover only the krudanta forms which are regular in behavior.

Irregular Verbs: Some verbs in Marathi have different behavior as compared to the other verbs (regular verbs). These verbs are present in some inflected forms for which no definite stem exists.

4.5 Adjectives

Marathi adjectives can be classified into two categories: ones that do not inflect and others that inflect for gender, number and case where such an inflection agrees with the gender and number of the noun modified by them. The inflectional forms of the adjectives are generated using similar procedure as that of nouns.

4.6 Pronouns

There are nine types of pronouns in Marathi. Pronouns possess very irregular behavior resulting into a large number of suppletive forms. In addition to these forms every pronoun has a specific oblique form (one each for singular and plural) to which shabdayogi avyayas can be attached.

4.7 Indeclinable Words

Adverbs, conjunctions and interjections are the indeclinable words. Some adverbs can be followed by a subset of postpositions.

5 Morphotactics and Automata

Along with the postpositions mentioned in the Table 4 the complete set of morphemes in Marathi includes the roots and their inflectional forms. Every morpheme is labeled according to the class it belongs to. These labels are used to define the 'Morphotactic FSM' that models Ma-

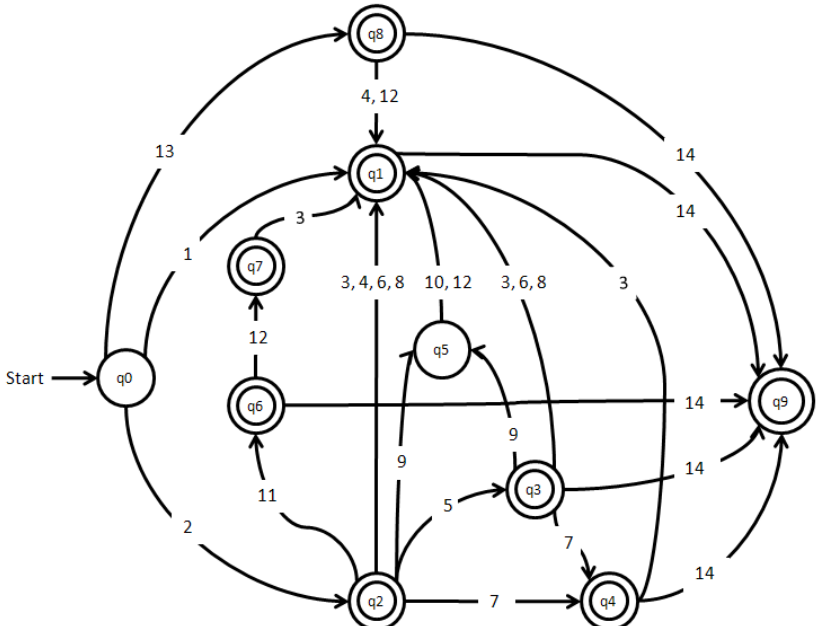

Figure 2. Morphotactic FSM

rathi language. Table 5 enlists various labels used in the Morphotactic FSM.

Type of Suffix		**Label**
Nouns, pronouns, nominal or adjectival krudantas	DF	1
	OF	2
Case markers		3
Adjectival postpositions	DF	4
	OF	5
Possessive case marker	DF	6
	OF	7
Closing postpositions		8
Peculiar NSTs		9
Exclusive postpositions		10
Nearly closing postpositions		11
चा {chaa} adjectival suffix		12
Adjective		1
Aakhyaatas		1
Adverbial krudantas		1
Adverbs-1		1
Adverbs-2		13
Shuddhashabdayogi avyayas		14

Table 5. Morphotactic Labels of Morphemes
DF: Direct form of a root or a suffix
OF: Oblique form of a root or a suffix
Adverb-1: The adverbs those cannot be followed by any postpositions
Adverb-2: The adverbs those can be followed by some postpositions

Note that, the *label* field mentioned in the inflection rules refers to the corresponding labels of the morphemes mentioned in Table 5.

Figure 2 shows the FSM for morphological recognition of Marathi. The input symbols are the labels of the morphemes as mentioned in the Table 5. The classification of the suffixes as specified in Table 5 explains the construction of FSM. We use SFST [1] (Stuttgart Finite State Transducer) for implementing the FSM.

6 Experiments

Morphological analysis caters to the needs of variety of application like machine translation, information retrieval, spell-checking. Different applications are interested in different bit of information provided by the analyzer like the stem, the root, the suffixes or the morphosyntactic features. Hence, the performance evaluation of a morphological analyzer has to be observed in terms of its impact on the performance of the applications that use it. Hence, we carry out the evaluation in two parts: In **direct evaluation** we directly measure the accuracy of morphological analyzer on the given data. In **indirect evaluation**, we observe the improvement in the performances of statistical pos tagger and chunker

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

by using the morphological analyzer to generate the morphological features that help in boosting their accuracies. We used the corpora in TOURISM and NEWS domain for all our experiments.

6.1 Direct Evaluation

We used Marathi Morphological Analyzer for the analysis of 21096 unique words. We manually measured the accuracy of the morphological analyzer by counting the number of correctly analyzed words out of the total number of words. In the cases where a word has multiple analyses, the word was counted as correctly analyzed only when all of the correct analyses are present. Note that, in order to emphasize more on the usefulness of our *approach* towards morphological analysis of Marathi, we added most of the roots used in the corpus to the lexicon before starting the experiments. For a language like Marathi, it is required to build a very rich lexicon which can be done over a larger period of time.

Out of the 21096 unique words, 20503 (97.18%) were found to be correctly analyzed. Of the remaining 593 words, 394 words could not be recognized by Morphological Recognizer and 199 words were assigned the incorrect or insufficient analyses.

By taking a closer look at the 394 words which were not recognized (segmented) we could come up with the causes of recognition failure as listed in Table 6.

Cause	**Number of Words**
Lexicon Coverage	82 (20.81%)
Absence of Rules	69 (17.51%)
Acronyms	66 (16.75%)
Compound words	55 (13.96%)
Irregular forms needing further investigation	47 (11.92%)
Transliterated words which are uncommon	25 (6.34%)
Unidentified words	20 (5.08%)
Dialect words/ words used in spoken language	20 (5.08%)
Use of common nouns as proper nouns	5 (1.27%)
Missing Paradigm	3 (0.76%)
Fusion (Sandhii)	2 (0.51%)

Table 6. Causes of Recognition Failure

6.2 Indirect Evaluation

CRF based sequence labelers (pos tagger and chunker) were trained using morphological features and the other elementary features like (contextual words and bigram tags). The morphological features include ambiguity scheme (set of all possible categories of a word) and the suffixes for the pos tagger whereas just the suffixes in case of chunker.

To throw the light of role played by morphological analyzer in improving the accuracies of the sequence labelers, we performed the experiments using two sets of features: The Learning Based (LB) labeler was trained using only elementary features whereas Morphologically Driven Learning Based (MDLB) labeler used the morphological features along with the elementary features. The results were obtained by performing 4-fold cross validation over the corpora. The average accuracy of MDLB Pos tagger turned out to be 95.03 as compared to 85% of LB. The average accuracy of MDLB chunker was found to be 97.87% whereas that of LB was found to be 96.91%. .

7 Conclusion and Future Work

We presented a high accuracy morphological analyzer for Marathi that exploits the regularity in the inflectional paradigms while employing the Finite State Systems for modeling the language in an elegant way. The accuracy figures as high as 97.18% in direct evaluation and the performance improvement in shallow parsing speak about the performance of the morphological analyzer. 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 the derivation morphology and compound words.

References

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

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

Damale, M. K. 1970. *Shastriya Marathii Vyaakarana*. Deshmukh and Company, Pune, India.

Dixit, Veena, Satish Dethe, and Rushikesh K. Joshi. 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.

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.

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

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

Oflazer, Kemal 1993. *Two-level Description of Turkish Morphology.* In The European Chapter of the ACL (EACL).