HinMA: Distributed Morphology based Hindi Morphological Analyzer

Ankit Bahuguna
TU Munich
ankitbahuguna@outlook.com

Lavita Talukdar
IIT Bombay
lavita.talukdar@gmail.com

Pushpak Bhattacharyya
IIT Bombay
pushpakbh@gmail.com

Smriti Singh
IIT Bombay
smritismriti@gmail.com

Abstract

Morphology plays a crucial role in the working of various NLP applications. Whenever we run a spell checker, provide a query term to a web search engine, explore translation or transliteration tools, use online dictionaries or thesauri, or try using text-to-speech or speech recognition applications, morphology works at the back of these applications. We present here a novel computational tool *HinMA*, or the Hindi Morphological Analyzer, based on the framework of Distributed Morphology (DM). We discuss the implementation of linguistically motivated analysis and later, we evaluate the accuracy of this tool. We find, that this rule based system exhibits extremely high accuracy and has a good overall coverage. The design of the tool is language independent and by changing few configuration files, one can use this framework for developing such a tool for other languages as well. The analysis of Hindi inflectional morphology based on the Distributed morphology framework, its implementation in the development of this tool and integration with NLP resources like Hindi Wordnet or Sense Marker Tool and possible development of a word generator are interesting aspects of this work.

1 Introduction

Natural Language Processing (NLP) systems aim to analyze and generate natural language sentences and are concerned with computational systems and their interaction with human language. Morphology accounts for the morphological properties of languages in a systematic manner, enabling us to understand how words are formed, what their constituents are, how they may be arranged to make larger units, what are the semantic and grammatical constraints involved and how morphological processes interact with syntactic and phonological ones. An analysis of the inflectional morphology of Hindi has been presented here in the theoretical framework of Distributed Morphology, as discussed by Halle and Marantz (1993, 1994); Harley and Noyer (1999). The theory has been used to develop the rules required to analyze and describe the various inflectional forms of Hindi words. Our tool takes an inflected word as input and outputs its set of roots along with its various morphological features using the output of the stemmer. The suffixes extracted by the stemmer are used to get the various morphological features of the word: *gender, number, person, case, tense, aspect and modality*. The tool consist of two parts – ***Stemmer***, which takes inflected word as input and stems it, to separate root and suffix and **Morphological Analyzer,** which takes <Root, Suffix> pair as input and outputs a set of features along with the set of roots.

Stemming aims to reduce morphologically related word forms to a single base form or stem. Stemmers use an affix-list and morphological rules that isolate the base form by stripping off possible affixes from a given word. The final stem is usually then looked up in the online language lexicon to verify its validity. **Morphological analysis** is provided by morphological analyzers that include morphological information for each morpheme − both stems and suffixes isolated by the stemmer. A Morphological Analyzer (MA), exploits only word level information and produces all possible roots and analyses for a given word. An MA should be able to produce all the possibilities if a word can be decomposed into two or more different ways to produce the roots of different Part of Speech (POS) categories. For such a word, the root and the morpheme analyses may be different in each case. For example, the Hindi word *khāte* in sentences **1** and

2 has two possible analyses: *khātā* 'ledger' as the root with suffix */-e/* and *khā 'eat'* as the root with suffixes */-t-/* and */-e/*. In Ex. 1, the word *khāte* has a noun root '*khātā*' and the suffix */-e/* appears to mark the plural number and the direct case. In Ex. 2, on the other hand, the word has a verb root *khā* 'eat' and the suffixes */-t/* and */-e/* appear to mark the features 'habitual aspect' and 'masculine-plural'. A morphological analyzer should typically provide both analyses for the word *khāte* unless some contextual information is used to resolve the categorical ambiguity. Examples:

1. *मेरे कई खाते हैं.*
   mere kəī khāte haĩ
   I-Poss many (bank) accounts be-pres-pl
   (I have many bank accounts)
2. *वे रोज़ चावल खाते हैं.*
   ve roz cāvəl khā-t-e haĩ
   They everyday rice eat-hab-pl be-pres,pl
   (They eat rice everyday)

Similarly, a word may also have multiple roots and multiple analyzes within the same POS category as shown in 3 below. The word *nālõ* can be analyzed in two ways: with *nāl* as the root or with *nālā* as the root. The suffix in both cases is same, *i.e., -õ* which represents the 'plural-oblique' case feature. Both are valid roots for the input word. Since an analyzer does not consider the contextual information of words to resolve POS ambiguities, it should be able to produce both outputs.

3. Input word form: *नालों* (*nālõ)*
   a. POS Category: Noun; Root 1: *nāl 'horse-shoe';* Suffix: *-õ;* Analysis: Plural, Oblique
   b. POS Category: Noun; Root 2: *nālā 'water channel/trough';*Suffix: *-õ;* Analysis: Plural, Oblique

An MA usually relies on its accompanying lexicon to match the extracted root and to provide the category information for a given word. However, the analyzer may fail to recognize certain word forms if the root formed by the stemmer after stripping off the suffix is absent in the lexicon. The analyzer may also fail to recognize spelling variants of the roots stored in the lexicon such as *क़ैदियों–कैदियों* (kædiyõ) 'prisoners', *हफ़्ते-हफ्ते* (həp^{h}te) 'weeks', etc. In the absence of the rules to handle spelling variations, the MA may not be able to analyse the spelling variants of a word. The remainder of this paper is organized as follows. We describe related work and background in section 2. Section 3 explains the concept of Distributed Morphology (DM). Implementation details are discussed in Section 4. Results are discussed in Section 5 and Error analysis in Section 6. Comparison with existing MA(s) is mentioned in Section 7. Section 8 discusses applications and Section 9 concludes the paper and points to future directions.

2 Related Work and Background

Several techniques have been utilized in building stemmers and morphological analyzers for Hindi. Some of them are morphology based, some statistical and some a hybrid of the two. The first ever reported work on Hindi stemming and morphological analysis was by Bharati *et al.* (2001). They present an algorithm that learns and predicts morphological patterns of Hindi using an existing Hindi morphological analyzer (MA). The paradigm-based MA uses a very low coverage lexicon. Roots are stored in a dictionary along with the paradigm information. Each paradigm stores information of the add-delete characters for a set of items for various inflectional categories (such as number and case for nouns). A representative root is chosen for each paradigm and is used as a label for paradigm assignment for the other roots in that paradigm. For each input word, the MA applies the add-delete strings and looks for a possible match in the root lexicon. If a match is found, it is considered to be the correct root and is the final output. If not, the next string is applied. Using this MA, Bharati *et al.* (2001) applied an automatic-learning algorithm to predict the stem of an inflected word using the frequency of occurrences of word forms in the raw (unannotated) corpus. The idea is to use the suffix to determine the set of possible stems and paradigms that may generate the input word form. Using the pairs of stems and paradigms, all possible word forms are generated. The frequency of these word forms is then obtained from the corpus and is stored in a vector. These vectors are compared for each 'guess' in order to select the most likely stem and the paradigm for the input word. This algorithm reportedly gave better coverage. Goyal and Lehal (2008) too developed a Hindi Morphological Analyzer that relies on a list of pos-

sible forms of the commonly used Hindi root words. Their approach promises to perform better than previous approaches, as the search time in a storage-based approach is very low. Another obvious advantage of storing all the forms in a list is that the system only needs to find a correct match in the system and output the corresponding root. In that sense, the user will always get accurate results. Ramanathan and Rao (2003) worked on 'light-weight stemming' for Hindi. They tried to build a computationally inexpensive and domain independent stemmer that extracts out the stem of a word by stripping off suffixes based on the 'longest match'. They created a list of 65 possible inflectional suffixes for Hindi nouns, adjectives, verbs and adverbs using McGregor's (1995) analysis of Hindi inflectional morphology. For an input word, the stemmer keeps stripping off suffixes using the suffix-list until it finds the longest match. But, the system may produce many incorrect stems since it has no way to identify whether or not a particular suffix is applicable to the identified stem. In addition, the stemmer does not output the root of the input word. Purely statistical methods were also tried out for Hindi stemming and morphological analysis. Larkey *et al.* (2003) worked on Hindi stemming, as it was needed in their Cross language information retrieval task. They used a list of 27 common suffixes supplied by a Hindi speaker that indicate nominalization, gender, number and tense features. In their system, the stemming was done to first extract out the longest possible suffix followed by smaller suffixes. But, the stemming process did not give them encouraging results. Since, the morphological analysis was not exhaustive, their system could not handle many word forms. They reported that stemming did not lead to any improvement in their retrieval task.

3 Distributed Morphology

Distributed Morphology, a recent theory of the architecture of grammar, was proposed by Halle and Marantz (1993, 1994). The theory proposes that 'words' are structurally not different from other constituents such as phrases or sentences, and are formed and manipulated using syntactic rules. This suggests that word formation is primarily a syntactic operation, *i.e.*, the morphological structure of a word or a word form is generated using syntactic operations. It is syntax that provides features and the structures upon which morphology operates. This view is opposed to the one that believes that morphology operates in an entirely separate component that generates words or word forms outside syntax that later feed into syntactic structures. Unlike lexicalist approaches that assume all morphology to happen in the lexicon, DM believes that the constituent components of morphology are distributed among various levels in the architecture of grammar and work in close connection with syntax and phonology. Halle and Marantz postulate a separate level of representation called *Morphological Structure* (MS) that operates in between *Syntactic Structure* (SS) and *Phonological Form* (PF). This level receives hierarchical structures from syntax that contain 'abstract' morphemes as the terminal nodes; abstract, because at this level, these nodes only have morpho-syntactic and semantic features and lack any associated phonological features. The DM grammar is represented by Halle and Marantz (1993) as shown in Figure 1.

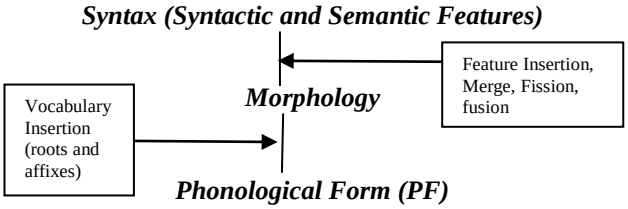

Figure 1. Architecture of grammar in DM.

4 Implementation of Distributed Morphology based Morphological Analyzer

The overall process can be summarised into *three distinct steps*: stemming, root formation and lexicon look-up and morphological analysis. For stemming, HinMA uses a set of ordered contextual rules to isolate and extract out suffixes from a given inflected word form. For implementation purposes, the vocabulary entries developed for nouns, adjectives, quantifiers, ordinals and verbs were converted into *if-then* rules arranged in order of specificity of inflectional and contextual features. The internal processes of HinMa is shown in Figure 2. The rules are applied from right to left iteratively until no suffixes remain and the base root is left. Readjustment rules apply wherever applicable to produce the correct root which is then matched

with the incorporated root-list to determine match (es). Then, the root is validated by performing a lexicon lookup. On successful validation, root(s) is obtained and it completes the second step. The information associated with the various rules and the lexicon is combined and provided as output of morphological analysis. A number of rules Singh S. et al. (2011) were constructed over a period of one year and later another one year was taken to develop and test the system with the help of a dedicated team of 4 linguists and two computer scientists. Due to space limitation, we are unable to present the individual rules here.

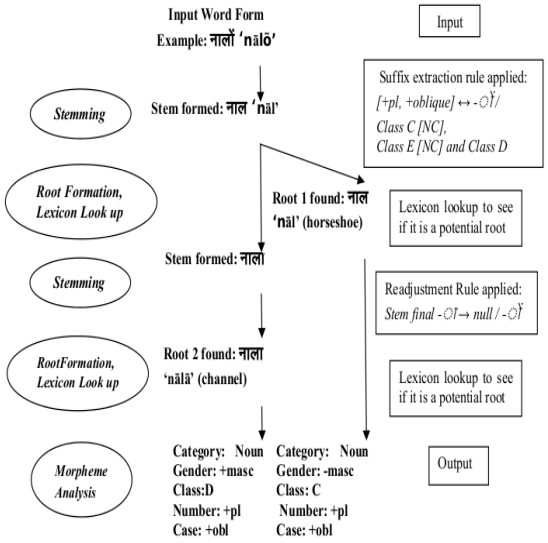

Figure 2. Steps show working of HinMa.

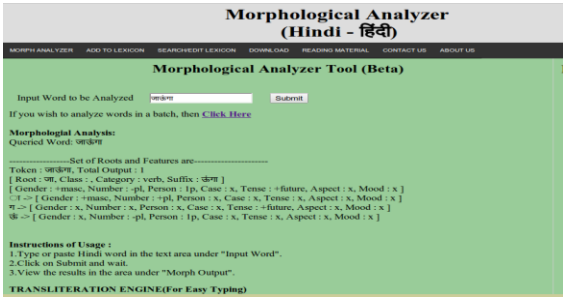

Figure 3: HinMA online implementation: Output of verb "जाऊंगा" (jAuMga ~ will go).

Output of the System: A detailed morpheme analysis is given as output for each word, with information such as root, grammatical category, inflection class and feature values. The system also produces a detailed morphological analysis for each morpheme that constitutes the word form. The output format is:

Input Token: XXXXX
Possible Root 1: class: category: suffix: morphemes (morpheme 1 … etc.): Morpheme Analysis (morpheme 1, morpheme 2, etc.)
Possible Root 2: …

The morpheme analysis of each suffix is produced in a seven field with values for the features *gender, number, person, case, tense, aspect,* and *mood.* Our system offers the analysis of words which could yield more than one root from with added capability of handling compound words. We provide demo output of online system[1] in Figure 3 and actual outputs categorised *w.r.t.,* various morphological phenomena below:

1. **Multiple roots within the same category:** The input word नालों 'nālõ' may have two possible noun roots which are नाल 'nāl' (horseshoe) and नाला 'nālā' (trough/channel). The two roots belong to different inflection classes. The system is able to output both analysis.

Token: नालों, Total Output: 2
Root: नाल, Class: C, Category: noun, Suffix: ोंो
Gender: -masc, Number: +pl, Person: x, Case: +oblique, Tense: x, Aspect: x, Mood: x
Root: नाला, Class: D, Category: noun, Suffix: ोंो
Gender: +masc, Number: +pl, Person: x, Case: +oblique, Tense: x, Aspect: x, Mood: x

2. Multiple roots across POS categories: The input word खाते 'khāte' may have two roots of different POS categories. It may be analyzed as a noun with the root खाता 'khātā' (ledger) and suffix -े 'e'. As a verb, its root is खा 'khā' (eat) with suffix -ते 'te'. Our MA is able to produce both outputs and their analysis, shown below:
Token: खाते, Total Output: 2
Root: खाता, Class: D, Category: noun, Suffix: े
Gender: +masc, Number: -pl, Person: x, Case: +oblique, Tense: x, Aspect: x, Mood: x
Root: खा, Class:, Category: verb, Suffix: ते
Gender: +masc, Number: +-pl, Person: x, Case: x, Tense: , Aspect: +conditional, Mood: x]
े -> [Gender: +masc, Number: +-pl, Person: x, Case: x, Tense: , Aspect: x, Mood: x
त -> [Gender: x, Number: x, Person: x, Case: x,

[1] http://www.cfilt.iitb.ac.in/~ankitb/ma/

Tense: x, Aspect: +conditional, Mood: x]
Gender: x, Number: x, Person: x, Case: x, Tense: x, Aspect: (-perfect: +habitual), Mood: x

3. Multiple morphological analyzes for a word form: A word may have multiple analyzes possible for the same suffix and root. The token साए 'sāe' (shadows) may represent the features 'singular-oblique' or 'plural-direct'.
Token: साए, Total Output: 2
Root: सा, Class:, Category: particle, Suffix: ए
Gender: , Number: , Person: , Case: , Tense: , Aspect: , Mood: x
Root: साया, Class: D, Category: noun, Suffix: ए
Gender: +masc, Number: -pl, Person: x, Case: +oblique, Tense: x, Aspect: x, Mood: x

4. Irregular forms: The system is able to yield the roots of irregular forms using the set of rules specific to irregular verbs. Ex. For the inflected word "गए", we have:
Token: गए, Total Output: 1
Root: जा, Class:, Category: verb, Suffix: ए
Gender: +masc, Number: +pl, Person: x, Case: x, Tense: x, Aspect: +perfect, Mood: x

5. Stem modifications: The system is able to do phonological readjustment on the stem after affix stripping such as vowel lengthening (i-ī in ताइ-ताई 'tāi-tāī' and पि-पी 'pi-pī', u-ū in बहु-बहू 'bəhu-bəhū' and छु-छू 'chu-chu'), vowel addition at the end (द-दो 'd-do') etc. For Example, 'taiyan'
Token: ताइयाँ, Total Output: 1
Root: ताई, Class: B, Category: noun, Suffix: याँ
Gender: -masc, Number: +pl, Person: x, Case: -oblique, Tense: x, Aspect: x, Mood: x

6. Compound words: The system is able to yield the roots of compound words of the template [A-B] using the set of rules, which capture inflection on one or either both the words. We have introduced specific categories as compound-noun, compound-adj, compound-adv and compound-verb.
Example: For an inflected compound word "वर्ण-भेदों", 'varn-bhedon' we get the following output:
Token: वर्ण-भेदों, Total Output: 1
Root: वर्ण-भेद, Class: A, Category: noun, Suffix: ोों;
Gender: +masc, Number: +pl, Person: x, Case: +oblique, Tense: x, Aspect: x, Mood: x

5 Results

We tested HinMA on a corpus of around 66,000 words (annotated and manually cross-checked) to check its performance. We would like to emphasize that there was no instance of failure at analysis of an inflectional form as long as its root was available in the lexicon. In a few cases, the root of a given word is present in the root-list but under a different spelling. Since, the lexicon does not store variants of the same root word, many roots are left unidentified by the system. However, if we enrich the lexicon by adding more entries and include certain variations in spelling such as Urdu-Hindi letter alternations (*क़ैदियों/कैदियों* 'kædiyõ' (prisoners), *हफ़्ते/हफ्ते* 'həphte' (weeks)) and nasal vs. nasalization (*क्रान्तिकारी/क्रांतिकारी* 'krāntikārī' (revolutionists)), we ought to get better coverage. Below we discuss, results and error analysis for each POS category. **Nouns:** We tested the Morphological Analyzer on 14475 Hindi noun forms extracted from the corpus and the results were verified manually. The system could correctly identify the roots and provide the morphological analysis for 13523 nouns (more than half of which require multiple analysis). A total of 1022 nouns remain unidentified, with 643 unique noun forms (rest repeated entries). **Verbs:** We tested the analyzer on 13160 Hindi verb forms and manually verified the results. The system was able to correctly analyze most of the regular and irregular forms. The system fails again with cases of incorrect spelling, hyphenated word forms, missing roots or where in the analyzed text there were extra/incorrect characters in the word form. The performance of the system on Hindi verbs is very impressive. The system fails to identify only 116 verbal forms.

6 Error Analysis

We performed error analysis based on a variety of different parameters with respect to the part of speech under consideration. The most error causing cases were that of Nouns and Verbs and hence we present their results here. We present them, specific to the observed parameter and the respective examples as follows:

- **Nouns:** Incorrect spelling: भैसों (correct spelling: भैंसों 'bhaĩsõ' (buffaloes))**;** Spelling Variations: क़ैदियों/कैदियों 'kædiyõ' (prisoners)**;** Missing root entries in the lexicon: दोहराव 'dohrāv' (repetition)**;** Borrowed nouns from foreign languages (foreign words): इंटरनेट 'intərnet (internet)**;** Adjectives/qualifiers functioning as nouns: सैंकड़ों 'sænkəɖõ' (thousands).
- **Verbs:** With missing roots in the lexicon: पदा 'pədā' (make somebody run); Hyphenated verbs: आने-जाने 'āne-jāne'**;** Verbs with incorrect or variant spelling: रक्खा (correct spelling: रखा 'rəkhā' (kept))**;** Verbs with extra characters due to faulty tokenization: देखने 'dekhne'.

7 Evaluation

Currently, for Hindi, there is only one state of the art Morphological Analyzer which is under **active development** and provided **constant updates.** It is developed by IIIT Hyderabad[2]. Thus, to evaluate, we executed our system against 200 words chosen randomly from the BBC news corpus[3] and then manually checked the accuracy of results on both HinMa and IIITH-MA. This methodology was adopted, since there is no publicly available gold data for this task. The low number of the evaluation corpus was to provide ease to the verifying linguist. But, as the data is chosen in random order and only unique words are considered, this brings some integrity to the evaluation methodology.

MA Systems	***HinMa***	***IIITH - MA***
Correct Results	186	181
Wrong/Unknown Words	14	19
Accuracy (%)	93	90.5

Table 1: Accuracy figures for evaluation of HinMA results with that of IIIT-H MA.

8 Applications

We have integrated HinMa with Hindi Wordnet and Sense Marker tool, they are described below:

1. **Integration with Hindi Wordnet:** The work was inspired by English Wordnet[4] developed at Princeton, Miller (1995); Fellbaum (1998) which gives results based on the stem of the query words consisting of inflection. For example, if we search for the word "लड़कियाँ" (girls) in Hindi Wordnet integrated with HinMa, the result is same as for word "लड़की" (girl). "लड़की" (girl) is the root form of the inflected word "लड़कियाँ" (girls). Thus. such an integration increases the coverage of results.
2. **Integration with Sense Marker Tool:** The sense marker tool (Chatterjee et al.) is used for marking the correct sense of the word from a given set of senses. This allows one to create a corpora of manually tagged words and this is extremely useful in NLP problem areas like word sense disambiguation. We have integrated HinMa with the sense marker tool thereby providing a better coverage and accuracy in terms of returned result(s) whenever an inflected word needs to be sense marked.

9 Conclusion and Future Work

In our paper, we have described the Hindi Morphological Analyzer (*HinMA*) which handles the Inflectional Morphology in the framework of Distributed Morphology (DM). Our approach first analyses the formation of inflectional forms of Hindi through the application of suffix insertion rules and then apply phonological readjustment rules. It was found that it works quite well for the words that are present in the lexicon. Using the basic concepts of DM, our analysis of Hindi nouns and verbs is able to generate the inflectional forms using a very small set of rules and an inflection-based classification of nouns and adjectives. We showed that the DM-based Hindi morphological analyzer is quite accurate and reliable, capable of both analysis and generation. Future work involves developing a *Word Generator for Hindi*. The linguistic resources used in the DM-based MA namely, the vocabulary items (suffixal entries) and the readjustment rules need to be applied in the reverse direction to produce fully inflected words using the root entries from the root-list and combining them with the affixal entries to generate surface forms. We encourage using this framework to develop

[2]http://sampark.iiit.ac.in/hindimorph/web/restapi.php/indic/morphclient

[3] http://www.bbc.co.uk/hindi/

[4] http://wordnetweb.princeton.edu/perl/webwn

morphological analyzers for other languages as well.

Acknowledgements
The authors would like to thank our team of linguists, Mrs. Jaya Jha, Mrs. Laxmi Kashyap, Mrs. Nootan Verma and Mrs. Rajita Shukla for their valuable inputs and their work on manually developing lexicon for this task

10 References

A. Ramanathan, and D. D. Rao. 2003. *A Lightweight Stemmer for Hindi*, Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics, 2003.

Bharati, A., R. Sangal, S. M. Bendre, M. N. S. S. K. Pavan Kumar and K. R. Aishwarya. 2001. *Unsupervised Improvement of Morphological Analyzer for Inflectionally Rich Languages*. In the Proceedings of the 6th NLP Pacific Rim Symposium, 685-692. Tokyo, Japan, November.

Chatterjee Arindam, Joshi Salil Rajeev, Khapra Mitesh M. and Bhattacharyya Pushpak, 2010. *Introduction to Tools for IndoWordnet and Word Sense Disambiguation*, The 3rd IndoWordnet Workshop, Eighth International Conference on Natural Language Processing (ICON 2010), IIT Kharagpur, India.

Christiane Fellbaum (1998, edition) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Halle, M., and A. Marantz. 1993. *Distributed Morphology and the Pieces of Inflection*. In The View from Building 20: Essays in Linguistics in Honour of Sylvain Bromberger, eds. K.

Harley, H. and R. Noyer. 1999. *Distributed Morphology* In GLOT International 4.4:3-9.

George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

Goyal, V. and Lehal G. S. 2008. *Hindi Morphological Analyzer and Generator*. In the Proceedings of the First International Conference on Emerging Trends in Engineering and Technology, 1156-1159. Nagpur, IEEE Computer Society Press, California, USA.

Leah S. Larkey, Margaret E. Connell, Nasreen Abduljaleel. 2003 *Hindi CLIR in thirty days* ,ACM Transactions on Asian Language Information Processing (TALIP),Volume 2 Issue 2, pages 130 - 14, ACM New York, NY, USA, June 2003.

McGregor, R.S. 1995. *Outline of Hindi grammar*. Oxford: Oxford University Press.

Singh, Smriti 2011. *Hindi Inflectional Morphology and its implementation in Language Processing Tools: A distributed Morphology Approach*, PhD Thesis, IIT Bombay, Mumbai, India.

.