1 Introduction

This paper primarily presents an analysis of nominal inflection in Hindi within the framework of Distributed Morphology (Halle & Marantz 1993, 1994 and Harley and Noyer 1999). Müller (2002, 2003, 2004) for German, Icelandic and Russian nouns respectively and Weisser (2006) for Croatian nouns have also used Distributed Morphology (henceforth DM) to analyze nominal inflectional morphology. This paper will discuss in detail the inflectional categories and inflectional classes, the morphological processes operating at syntax, the distribution of vocabulary items and the readjustment rules required to describe Hindi nominal inflection. Earlier studies on Hindi inflectional morphology (Guru 1920, Vajpeyi 1958, Upreti 1964, etc.) were greatly influenced by the Paninian tradition (classical Sanskrit model) and work with Paninian constructs such as root and stem. They only provide descriptive studies of Hindi nouns and verbs and their inflections without discussing the role or status of affixes that take part in inflection. The discussion on the mechanisms (morphological operations and rules) used to analyze or generate word forms are missing in these studies. In addition, these studies do not account for syntax-morphology or morphology-phonology mismatches that show up in word formation. One aim of this paper is to present an economical way of forming noun classes in Hindi as compared to other traditional methods, especially gender and stem ending based or paradigm based methods that give rise to a large number of inflectional paradigms. Using inflectional class information to analyse the various forms of Hindi nouns, we can reduce the number of affixes and word-generation and readjustment rules that are required to describe nominal inflection. The analysis also helps us in developing a morphological analyzer for Hindi. The small set of rules and fewer inflectional classes are of great help to lexicographers and system developers. To the best of our knowledge, the analysis of Hindi inflectional morphology based on DM and its implementation in a Hindi morphological analyzer has not been done before. The methods discussed here can be applied to other Indian languages for analysis as well as word generation.

1 Acknowledgements

We thank the anonymous conference reviewers for their reviews. In this revised submission, we have tried to incorporate their suggestions and answer their questions. We also thank P. Bhattacharyya and O. Damani (IIT Bombay) for their input and support and Nikhilesh Sharma who helped us implement the Hindi morphological analyzer.
2 Inflection in Hindi Nouns

Hindi nouns show morphological marking only for number and case. Number can be either singular or plural and can be represented as a binary valued feature [±pl]. Singular [-pl] is the default value for number which is morphologically unexpressed while plural or the non-default value [+pl] may be phonologically realized. Case marking on Hindi nouns is either direct or oblique. Marked (oblique) nouns show cumulative exp onence for case and number, e.g., -e in lāk-e (boy-oblique) and -ō in rājā-ō (kings-oblique) for singular-oblique and plural-oblique respectively. Gender, an inherent, lexical property of Hindi nouns(masculine or feminine) is not morphologically marked, but is realized via agreement with adjectives and verbs. We must point out that (1) a few nouns may be in either gender given the context, e.g., dost or mir (friend) and that (2) natural sex distinction in humans lākā-lākī (boy-girl), bācchā-bācchī (baby-boy and baby-girl), in a few animals ghūrā-ghoṛū (horse-mare) and some kinship terms dādā-dādi (paternal grandpa-grandma), māmā-māmī (maternal uncle-aunt) are marked using specific stem endings, i.e., feminine nouns tend to end in vowel /i/ while masculine nouns tend to end in /ā/. This is, however, not generally the case, for example pānī (water) is masculine and mālā (garland) is feminine.

In the following tables we show the inflections selected by Hindi nouns. Table 1 shows that Hindi feminine nouns of inflection Type 1 are marked null for all number-case values. Type 2 and Type 3 nouns inflect only in the plural for both case values. Table 2 shows the inflection for masculine Hindi nouns. Inflection is seen again in Type 2 and 3 nouns in the plural for both case values and in the singular for only Type 2 nouns in the oblique.

<table>
<thead>
<tr>
<th>Type 1</th>
<th>Type 2</th>
<th>Type 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td>Oblique</td>
<td>Direct</td>
</tr>
<tr>
<td>Singular</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>Examples</td>
<td>āg 'fire', pyās 'thirst'</td>
<td>āg, pyās</td>
</tr>
<tr>
<td>Plural</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>Examples</td>
<td>āg, pyās</td>
<td>āg, pyās</td>
</tr>
</tbody>
</table>
3 Noun Classification Systems for Hindi in the Literature

Traditional classification (from the Paninian perspective) of Hindi nouns is based on gender and stem endings. This system does not allow two nouns of different genders or different stem endings to be in one class. With two genders and around nine stem endings (ā, ī, ī, u, o, O/au, û and consonant), we get at least eighteen classes. In addition, nouns that have one of these stem endings but take null for all case-number values are put into different inflectional classes. This results in a large number of nominal classes (approximately thirty) that display similar inflectional behaviour. Many readjustment rules are also required to explain the phonological changes in the inflected forms. Table 3 provides one example of nouns placed in different classes because of different stem endings even though they take similar inflectional markers and belong to the same gender.

Kachru (2006) categorizes Hindi nouns into five declension types as given in Table 4 below. This classification is based on how Hindi nouns decline for gender, number and case. The classification criteria, however, are not clear.
Each class includes both masculine and feminine nouns. The last three declensions include nouns with identical stem endings i, ī, and consonant respectively while the first two do not, i.e., the masculine nouns in the first declension are ā ending while feminine nouns are ī ending and the second declension has ī ending masculine nouns and ā ending feminine nouns. Further, rules that describe affix insertion, stem alternation/modification are also missing from the discussion.

Table 4: Kachru’s Classification of Hindi Nouns (Kachru, 2006, p52-53)

<table>
<thead>
<tr>
<th>Class 1</th>
<th>Masc: ā, Fem: ī ending</th>
<th>[−pl, −obl]</th>
<th>[pl,+obl]</th>
<th>[+pl,-obl]</th>
<th>[+pl,+obl]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masc</td>
<td>ār̥kā ‘boy’</td>
<td>ār̥k-e</td>
<td>ār̥k-e</td>
<td>ār̥k-ō</td>
<td></td>
</tr>
<tr>
<td>Fem</td>
<td>ār̥kī ‘girl’</td>
<td>ār̥kī</td>
<td>ār̥kī-yā</td>
<td>ār̥kī-yā</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class 2</th>
<th>Masc: ī, Fem: ā ending</th>
<th>[−pl, −obl]</th>
<th>[pl,+obl]</th>
<th>[+pl,-obl]</th>
<th>[+pl,+obl]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masc</td>
<td>sār̥i ‘friend’</td>
<td>sār̥i</td>
<td>sār̥i</td>
<td>sār̥i-yō</td>
<td></td>
</tr>
<tr>
<td>Fem</td>
<td>kənyā ‘girl’</td>
<td>kənyā</td>
<td>kənyā-ê</td>
<td>kənyā-ō</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class 3</th>
<th>Masc: ā, Fem: ī ending</th>
<th>[−pl, −obl]</th>
<th>[pl,+obl]</th>
<th>[+pl,-obl]</th>
<th>[+pl,+obl]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masc</td>
<td>pāti ‘husband’</td>
<td>pāti</td>
<td>pāti</td>
<td>pāti-ŷō</td>
<td></td>
</tr>
<tr>
<td>Fem</td>
<td>siddī ‘success’</td>
<td>siddī</td>
<td>siddī-yā</td>
<td>siddī</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class 4</th>
<th>Masc: ā, Fem: ā ending</th>
<th>[−pl, −obl]</th>
<th>[pl,+obl]</th>
<th>[+pl,-obl]</th>
<th>[+pl,+obl]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masc</td>
<td>sār̥ū ‘co-brother’</td>
<td>sār̥ū</td>
<td>sār̥ū</td>
<td>sār̥ū-ō</td>
<td></td>
</tr>
<tr>
<td>Fem</td>
<td>bəhū ‘daughter-in-law’</td>
<td>bəhū</td>
<td>bəhū-ē</td>
<td>bəhū-ō</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class 5</th>
<th>Masc: ā, Fem: ī ending</th>
<th>[−pl, −obl]</th>
<th>[pl,+obl]</th>
<th>[+pl,-obl]</th>
<th>[+pl,+obl]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masc</td>
<td>sīyār ‘jackal’</td>
<td>sīyār</td>
<td>sīyār</td>
<td>sīyār-ō</td>
<td></td>
</tr>
<tr>
<td>Fem</td>
<td>cīl ‘eagle’</td>
<td>cīl</td>
<td>cīl-ē</td>
<td>cīl-ō</td>
<td></td>
</tr>
</tbody>
</table>

We see in Table 4 that the feminine nouns in Classes 2, 4 and 5 show similar inflectional behaviour as they are marked with -ē and -ō in the plural, direct and the plural, oblique respectively. Similarly, the feminine nouns in Classes 1 and 3 take similar inflections. The masculine nouns in Classes 2, 3, 4 and 5 are marked with -ō or -vō in the plural, oblique and null for all other combinations of number and case values. Since many of these classes group together quite naturally they should be merged. This classification appears to be neither intuitive nor systematic.

4 Inflection-based Noun Classes for Hindi Nouns

We propose that nominal classes in Hindi should be formed based entirely on the inflectional behaviour of nominal forms. Consequently, all feminine
nouns in Table 3 can be put in a single class. The feminine nouns in Classes 2 and 4 in Kachru’s classification scheme given in Table 4 belong in this class. Class 1 and Class 3 feminine nouns in her classification may be merged to form another class. Masculine nouns in Classes 2, 3, 4 and 5 can be merged into one class, while the masculine nouns in Class 1 form a separate class. This classification is similar to that of Shapiro (2000), summarized in Table 5, who identifies four inflectional classes based on the inflectional behaviour of Hindi nouns, two each for masculine and feminine nouns. Shapiro, however, does not give any reasons for his classification strategy nor the rules to derive the inflectional forms.

Table 5: Shapiro’s Classification of Hindi Nouns (Shapiro, 2000, p31-33, 38-39)

<table>
<thead>
<tr>
<th>Feminine</th>
<th>Masculine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class I</td>
</tr>
<tr>
<td>Sg-dir</td>
<td>null</td>
</tr>
<tr>
<td>Sg-obl</td>
<td>null</td>
</tr>
<tr>
<td>Pl-dir</td>
<td>-yā</td>
</tr>
<tr>
<td>Pl-obl</td>
<td>-yō</td>
</tr>
</tbody>
</table>

Shapiro also does not discuss the behaviour of nouns that are marked null for all case-number pairs. We put these nouns in Class A along with Type 1 feminine and Type 1 masculine nouns seen in Tables 1 and 2 respectively. The five proposed nominal classes along with the exponents (leaving out vocative case inflections) are shown in Table 6 below.

Table 6: Inflectional Classes and Suffixes for Hindi Nouns

<table>
<thead>
<tr>
<th></th>
<th>Class A</th>
<th>Class B</th>
<th>Class C</th>
<th>Class D</th>
<th>Class E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sg-dir</td>
<td>null</td>
<td>null</td>
<td>null</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>Sg-obl</td>
<td>null</td>
<td>null</td>
<td>null</td>
<td>-e</td>
<td>null</td>
</tr>
<tr>
<td>Pl-dir</td>
<td>null</td>
<td>-yā</td>
<td>-ē</td>
<td>-e</td>
<td>null</td>
</tr>
<tr>
<td>Pl-obl</td>
<td>null</td>
<td>-yō</td>
<td>-ō</td>
<td>-ō</td>
<td>-yō/-ō</td>
</tr>
</tbody>
</table>

The inflection based nominal classification system, permits us to describe the inflectional behaviour of Hindi nouns using a very small set of affixes and readjustment rules. All nouns of one class display similar inflectional
behaviour for all case-number pairs. In the following we discuss briefly some identifiable properties of each class.

**Class A:** Includes those nouns (masculine and feminine) that take null for all case-number values such as pyār (love), krodh (anger), bhūkh (hunger), pyās (thirst), mithās (sweetness), etc. These nouns are typically abstract or uncountable.

**Class B:** Includes /iː/, /i/ or /yā/ ending feminine nouns that take -yā for the features [+pl, -oblique] and -yō for [+pl, +oblique] such as lōrkī (girl), śakti (power) and dībiyā (small box), guṛīyā (doll), etc.

**Class C:** Includes feminine nouns that take -ē for the feature [+pl] and -ō for [+pl, +oblique] such as rāt (night), mālā (garland), bēshū (daughter-in-law), ritu (season), lō (flame), etc.

**Class D:** Includes masculine nouns that end in /ā/ or /yā/ such as lōrkā (boy), dhāgā (thread), lohā (iron), kuā (water well), etc. A few kinship terms such as bhatījā (paternal nephew), bhaṭā (maternal nephew), sālā (borther-in-law) (Guru, 1920) are also a part of this class. Nouns borrowed directly from Sanskrit such as rājā (king), pītā (father), yuvā (youngster), devtā (God), kārtā (doer), etc. also belong to this class.

**Class E:** Includes masculine nouns that inflect only for the features [+pl, +oblique]. The nouns in this class end with /ā/ or /yā/ or a consonant. Examples are ālū (potato), sādhū (saint), mālī (gardener), kōvī (poet), ghor (home), khet (farm), etc. The /ā/ ending tātsam masculine nouns borrowed from Sanskrit such as rājā (king), pītā (father), yuvā (youngster), etc. also belong to this class.

There are significant advantages to forming inflection based noun classes. First the classification is based on the choice of inflectional markers for four case-number pairs rather than on the stem endings or gender property of nouns which do not uniquely describe the inflectional behaviour of nominals in Hindi. Gender or stem endings are stored as lexical features of the nouns. Second, this approach yields fewer nominal classes, and this economy is coupled with greater generalization of nominal inflectional behaviour. Many stem alternation patterns are properly left to the domain of phonology.

---

2 According to classical Hindi Grammar, these nouns are bhāvācak (abstract) or guṇavācak (qualitative) nouns (Guru, 1920).
5 Syncretism and Allomorphy in Hindi Nouns

In DM, syncretism is defined as the realization of a single vocabulary item (affix) that is matched with more than one set of features on a terminal node. Intra-class syncretism in Hindi is exhibited by suffix -e of Class D that consists of /a/ ending masculine nouns. This suffix marks nouns of the same class for two different set of morphological features [+pl, -oblique] as well as [-pl, +oblique]. Some of the nominal suffixes are also allomorphic. The two suffixes, i.e., -ô and -yô which realize the features [+pl, +oblique] for Classes B and C are phonologically conditioned allomorphs selected based on the phonological form of the stem. Nouns that end in the vowels /ī/, /i/, or /yā/ take the suffix -yô while all other vowel and consonantal ending nouns take -ô. Allomorphy in Hindi is also driven by etymological origins of the words. Masculine tatsam nouns such as rājā (king) and pitā (father) do not behave like non-tatsam /ā/ ending words such as lākā (boy) and dhāgā (thread). All /ā/ ending Hindi nouns take -e for the features [-pl, +oblique] and [+pl, - oblique] (except those in Class A). But, tatsam nouns do not inflect for these features in the language of origin, Sanskrit, and appear to retain the same behaviour in Hindi as well.

6 Predicting Inflectional Class for New Lexemes

Using the inflection based nominal classification system, let us see how a new noun lexeme entering the language could be assigned gender and how we could predict its inflectional class. Gender can be assigned in two ways to a new lexeme 1) by virtue of its phonological form and 2) by semantically mapping the noun to an existing noun in Hindi. In Hindi, most of the masculine nouns end in ā while feminine nouns end in ī. If the new lexeme ends in one of these vowels, it is relatively easy to assign it gender. Certain new words such as kār (car) or mōtār (motor) refer to ‘gārī’ (vehicle) in Hindi which is feminine. Both borrowed words are assigned feminine gender. After gender is lexically assigned to the new lexeme, its inflectional class can be predicted using the procedure outlined in Figure 1. A masculine noun may or may not be inflected - based on its semantic property. If it is an abstract noun or a mass noun it will fall into the non-inflecting Class A irrespective of its phonological form. On the other hand, a countable lexeme will fall into one of the two masculine classes based on its phonological form. For example, zirauks (xerox) and pepər (paper) are both consonant final nouns that fall into the second masculine class, Class E. Similar procedures apply to feminine nouns as well.
In DM, before vocabulary insertion, the terminal nodes available in the syntactic structure undergo morphological operations such as merger, fusion, fission, and impoverishment (Halle & Marantz 1993, Harley & Noyer 1999). The operations account for the mismatches between the syntactic and morphological structures of word forms. In Hindi, where number and case inflections are marked cumulatively on a noun, a terminal node with case-number features accompanies the N node for all nouns in the syntactic tree. The noun node raises up the tree by head movement and merges with the case-number node (after fusion of case and number node). Thus, even though syntax provides insertion nodes for root, case and number, only two remain available for final insertion after morphological operations are applied. This results in a structure where two kinds of morphemes (root and an affix) are inserted in the two nodes. The final surface form is realized as a single word with two morpheme pieces such as rājā-ā (kings-pl-oblique), lāṛki-ā (girls-pl-direct), māligi-ō (gardeners-pl-oblique), etc.

After syntax and the application of morphological operations, vocabulary items are inserted into terminal nodes to provide connections between phonological and grammatical features. This is called vocabulary insertion
in DM. These vocabulary items are underspecified and compete for insertion at the terminal nodes. The items are arranged in order of specificity (highly specified followed by less specified ones) and feature hierarchy (plural entries followed by those for singular). The more specific entries succeed over less specified items. The vocabulary items for Hindi nouns are given below in (1).

(1) Vocabulary Insertion Rules

\[
\begin{align*}
&\{\pm \text{pl}, \pm \text{oblique}\} \leftrightarrow \text{null} / \text{Class A} \quad \text{----------- 1} \\
&\{+\text{pl}, +\text{oblique}\} \leftrightarrow -\text{yõ} / \text{Class B & E (Stem ending ī or yā)} \quad \text{----------- 2} \\
&\{+\text{pl}, +\text{oblique}\} \leftrightarrow -\text{ō} \quad \text{----------- 3} \\
&\{+\text{pl}\} \leftrightarrow -\text{yā} / \text{Class B} \quad \text{----------- 4} \\
&\{+\text{pl}\} \leftrightarrow -\text{ē} / \text{Class C} \quad \text{----------- 5} \\
&\{+\text{pl}\} \text{ or } [-\text{pl}, +\text{oblique}] \leftrightarrow -\text{e} / \text{Class D} \quad \text{----------- 6} \\
&\{\pm \text{pl}\} \leftrightarrow \text{null} \quad \text{----------- 7} \\
\text{(elsewhere rule)}
\end{align*}
\]

Rule 1 applies when a noun root is specifically marked for Class A. It inserts \textit{null} for all case-number values. Rule 2 is for those /ī/ and /yā/ ending nouns that take -yõ for the features [+pl, +oblique]. Rule 3 inserts -ō for the features [+pl, +oblique] for all other nouns. Rule 4 and 5 are specific for plurals of Class B and Class C respectively. Rule 6 applies to Class D nouns in [+pl] and [-pl, +oblique]. Rule 7 is the elsewhere rule that entails \textit{null} insertion for the remaining plural and singular noun forms.

We also propose an impoverishment rule in (2) that deletes [-oblique] when the number feature is present. This means that the entries specified for number (singular or plural) need not be specified for [-oblique] feature (or for direct case). Thus the rules \{-pl, -oblique\} \leftrightarrow \textit{null} and \{+pl, -oblique\} \leftrightarrow \textit{null} can be replaced by a single rule, i.e., \{±pl\} \leftrightarrow \textit{null}.

(2) Impoverishment Rule

\[-\text{oblique} \rightarrow \textit{null} / \{±\text{pl}\}\]

Affixation also yields some phonological changes. We propose the following Readjustment rules for Hindi:

(3) Readjustment Rules

\[
\begin{align*}
\text{Stem final /ā/} & \rightarrow \textit{null} / \text{Class D with -e or -ō} \quad \text{----------- 8} \\
\text{Stem final /ū/} & \rightarrow \textit{u} / -\text{ē} \text{ or -ō} \quad \text{----------- 9} \\
\text{Stem final /ī/} & \rightarrow \textit{i} / -\text{yā} \text{ or -ō} \quad \text{----------- 10}
\end{align*}
\]

316
The first readjustment rule (rule 8) deletes the stem final vowel of Class D nouns that take either -e or -õ, e.g., laṛkā-e, laṛkā-õ and sāyā-õ and create laṛke, laṛkō and sāyō respectively. Rules 9 and 10 are not class specific and result in final vowel shortening in nouns (masculine or feminine) that end in either ē or ē. Thus, bōhū-ē and bōhū-õ become bōhuē and bōhuõ while laṛkī-ŷā and laṛkī-ŷō become laṛkiyā and laṛkiyō respectively.

8 DM Based Hindi Morphological Analyzer

A morphological analyzer aims to recover from an inflected word its base form (stem) by stripping off possible affixes. To this base, phonological (readjustment) rules are applied to generate the root. A search is made for this root in the lexicon to determine if there is a match. This process can also yield multiple roots belonging to multiple lexical categories. Morphological information for roots and suffixes is also provided. In order to develop such a system, a root lexicon, affixal entries and phonological rules are needed. We developed a Hindi lexicon with forty thousand noun root entries. These roots were manually categorized into five classes and were then marked with information about the inflectional class, lexical category, gender and stem ending. Vocabulary items or affixal rule entries were developed that provide information about the context(s) in which affixes appear. Since these rules are bidirectional, these can be used to analyze as well as generate nominal forms. We provide an example below of the analysis of a noun using the DM based morphological analyzer.

- Input noun form: laṛkiyā (girls)
- Rule (vocabulary item) applicable: [+pl] ↔ -yā / Class B (rule 4)
- Output after extracting out the suffix → Stem: laṛkī, Suffix: yā
- Readjustment Rule applied: Stem final ō → ō -yā or -õ (rule10)
- Apply the rule in the reverse direction to get the root and look for it in the lexicon.
- If found, output the root which is laṛkī (girl). If not, try applying another applicable rule.

The actual output of the system for the input words शहरों (शहरों) ‘cities’ and मौके (mauke) ‘chances’ is given below.

(4) Token: शहरों. Total Output: 1

[Root: शहर, Class: E, Category: noun, Suffix: ऊं]
[Gender: +masc, Number: +pl, Case: +oblique]
It may be noted that we require a few more affixal rules to implement the morphological analyzer since the analyzer works on Hindi data in the devanagri script, the new set of rules is given below in (6). Rules 3, 5, 6, 9 and 10 have been split into a and b to account for different devanagri characters for the phonemes /õ/, /œ/, /ẹ/, /û/ and /ī/ respectively. We have also made some modification to our previous list of Stem Readjustment rules (rules 8-10 in (3)) for the same reason.

(6) Vocabulary Insertion Rules (revised)

\[
\begin{align*}
[±\text{pl, }±\text{oblique}] & \leftrightarrow \text{null} /\text{Class A} \quad \text{----------- 1} \\
[+\text{pl, }+\text{oblique}] & \leftrightarrow -\text{ॉू} /\text{Class B and E (Stem ending ō, i or yā)} \quad \text{----------- 2} \\
[+\text{pl, }+\text{oblique}] & \leftrightarrow -\text{ॉै} /\text{Class C and E [NC], Class D} \quad \text{----------- 3a} \\
[+\text{pl, }+\text{oblique}] & \leftrightarrow -\text{अू} \quad \text{----------- 3b} \\
[+\text{pl}] & \leftrightarrow -\text{ॉृ} /\text{Class B} \quad \text{----------- 4} \\
[+\text{pl}] & \leftrightarrow -\text{ॉृ} /\text{Class C [NC]} \quad \text{----------- 5a} \\
[+\text{pl, }-\text{oblique}] & \leftrightarrow -\text{एू} /\text{Class C} \quad \text{----------- 5b} \\
[+\text{pl}] \text{ or } [-\text{pl, }+\text{oblique}] & \leftrightarrow -\text{ॉू} /\text{Class D [Nā]} \quad \text{----------- 6a} \\
[+\text{pl}] \text{ or } [-\text{pl, }+\text{oblique}] & \leftrightarrow -\text{डू} /\text{Class D} \quad \text{----------- 6b} \\
[±\text{pl}] & \leftrightarrow \text{null} \quad \text{----------- 7}
\end{align*}
\]

(Note: NC: noun stem ending in a consonant, Nā: Noun stem ending in ā)

(7) Readjustment Rules (revised)

\[
\begin{align*}
\text{Stem final -ॉर or -अर} & \rightarrow \text{Ø} /\text{Class D [Nā] with -ॉ or -अर} \quad \text{----------- 8} \\
\text{Stem final -ू} & \rightarrow -\text{ूू} / -\text{ूू} \quad \text{----------- 9a} \\
\text{Stem final -ै} & \rightarrow -\text{ैू} / -\text{ैू} \quad \text{----------- 9b} \\
\text{Stem final -ौ} & \rightarrow -\text{ौू} / -\text{ौू} \quad \text{----------- 10a} \\
\text{Stem final -ै} & \rightarrow -\text{ुू} / -\text{ुू} \quad \text{----------- 10b}
\end{align*}
\]
9 Evaluation, Results and Future Directions

We performed the test on 14480 Hindi noun forms extracted from news items sourced from the website www.bbc.co.uk/Hindi and carried out manual evaluation to verify the results. The system was able to identify and produce correct root and morphological analysis for 12784 nouns (more than half of which had more than one possible stem) while 1696 remain unidentified. Out of these 1696 noun forms, about 900 were unique forms. Analysis showed that many of these words (two hundred) were left unidentified because of either incorrect or variant spelling. Hyphenated compound nouns (350) too remain unidentified. A large number of the remaining unrecognized entries were uninflected nouns for which the lexicon lacked entries. The current system does not produce any output for these uninflected nouns. The types of unidentified words with their counts are given in Table 7 and Table 8 below.

Table 7: Results of DM Based Hindi Morphological Analyzer

<table>
<thead>
<tr>
<th>Testing Results</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Number of Words in the Testing Corpus</td>
<td>14480</td>
</tr>
<tr>
<td>Number of words correctly analyzed</td>
<td>12784</td>
</tr>
<tr>
<td>Total number of unidentified words</td>
<td>1696</td>
</tr>
<tr>
<td>Total number of unique unidentified words</td>
<td>900</td>
</tr>
</tbody>
</table>

Table 8: Types of Unidentified Words and their Counts

<table>
<thead>
<tr>
<th>Unique unidentified/unknown words (900)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Words with incorrect or variant spelling</td>
<td>200</td>
</tr>
<tr>
<td>Hyphenated words</td>
<td>350</td>
</tr>
<tr>
<td>Missing root entry in the lexicon</td>
<td>350</td>
</tr>
</tbody>
</table>

Below are various types of errors faced by the system and the examples of each error type.

- Roots not available in the lexicon:
  इंटरनेट ‘internet’, मेमरी ‘memory’, टॉयलेट ‘toilet’
- Spelling variants, Urdu-Hindi letter alternations, nasal vs. nasalization etc.:
We would like to emphasize that there was no instance of failure at analysis of a nominal form as long as the root was available in the lexicon. In addition, roots for a number of forms including borrowed words from English taking Hindi nominal inflections such as अकर (car-s), मोटर (motor-s), पेपर (paper-s) for which roots are missing in the dictionary are also, interestingly, suggested by the system. This is done by applying a rule that is applicable for the given form (i.e., if there was a match between the suffix in the word form and in the rule). Thus, the morphological analysis that is discussed here finds reliable, natural extension in other Natural Language Processing systems and tools such as Part-of-Speech Taggers and Parsers.

References


