Hindi Noun Inflection and Distributed Morphology

Smriti Singh
Indian Institute of Technology Bombay

Vaijayanthi M Sarma
Indian Institute of Technology Bombay

Proceedings of the 17th International Conference on
Head-Driven Phrase Structure Grammar

Université Paris Diderot, Paris 7, France

Stefan Müller (Editor)

2010

CSLI Publications

pages 307–321

http://csli-publications.stanford.edu/HPSG/2010

Singh, Smriti, & Sarma, Vaijayanthi M. 2010. Hindi Noun Inflection and Distributed Morphology. In Müller, Stefan (Ed.), *Proceedings of the 17th International Conference on Head-Driven Phrase Structure Grammar, Université Paris Diderot, Paris 7, France*, 307–321. Stanford, CA: CSLI Publications.

1 Introduction

This paper[1] 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 exponence for case and number, e.g., *-e* in *ləɽk-e (boy-oblique)* and *-õ* in *rājā-õ (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 *mitr (friend)* and that (2) natural sex distinction in humans *ləɽkā-ləɽkī (boy-girl), bəccā-bəccī (baby-boy and baby-girl)*, in a few animals *ghoɽā-ghoɽī (horse-mare)* and some kinship terms *dādā-dādī (paternal grandpa-grandma), māmā-māmī (maternal uncle-aunt)* are marked using specific stem endings, i.e., feminine nouns tend to end in vowel */ī/* 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 1: Types of Inflections for Hindi Feminine Nouns

	Type 1		**Type 2**		**Type 3**	
	Direct	Oblique	Direct	Oblique	Direct	Oblique
Singular	***null***	***null***	***null***	***null***	***null***	***null***
Examples	*āg 'fire', pyās 'thirst'*	*āg, pyās*	*nədī 'river', šəkti 'power'*	*nədī, šəkti*	*lətā 'vine', rāt 'night'*	*lətā, rāt*
Plural	***null***	***null***	***-yə̃***	***-yõ***	***-ẽ***	***-õ***
Examples	*āg, pyās*	*āg, pyās*	*nədi-yə̃, šəkti-yə̃*	*nədi-yõ, šəkti-yõ*	*lətā-ẽ, rāt-ẽ*	*lətā-õ, rāt-õ*

Table 2: Types of Inflections for Hindi Masculine Nouns

	Type 1		Type 2		Type 3	
	Direct	Oblique	Direct	Oblique	Direct	Oblique
Singular	***null***	***null***	***null***	***-e***	***null***	***null***
Example	*krodh 'anger', pyār 'love'*	*krodh, pyār*	*ləɽkā 'boy', bəccā 'baby'*	*ləɽk-e, bəcc-e*	*ādmī 'man', ghər 'home'*	*ādmī, ghər*
Plural	***null***	***null***	***-e***	***-õ***	***null***	***-õ/-yõ***
Example	*krodh, pyār*	*krodh, pyār*	*ləɽk-e, bəcc-e*	*ləɽk-õ, bəcc-õ*	*ādmi ghər*	*ādmi-yõ, ghər-õ*

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, i, ū, u, o, O/au, yā* 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.

Table 3: Hindi Feminine Nouns Taking Similar Inflections

	consonant ending	***ā ending***	***ū ending***	***u ending***	***au ending***
Noun	***rāt 'night'***	***mātā 'mother'***	***bəhū 'daughter-in-law'***	***ritu 'season'***	***lau 'flame'***
Pl-dir	*rāt-ẽ*	*mātā-ẽ*	*bəhu-ẽ*	*ritu-ẽ*	*lau-ẽ*
Pl-obl	*rāt-õ*	*mātā-õ*	*bəhu-õ*	*ritu-õ*	*lau-õ*

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, ū* 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)

		[-pl, -obl]	**[pl,+obl]**	**[+pl,-obl]**	**[+pl,+obl]**
Class 1 *Masc: ā, Fem: ī ending*	**Masc**	*ləɽkā 'boy'*	*ləɽk-e*	*ləɽk-e*	*ləɽk-õ*
	Fem	*ləɽkī 'girl'*	*ləɽkī*	*ləɽki-yə̃*	*ləɽki-yə̃*
Class 2 *Masc: ī, Fem: ā ending*	**Masc**	*sāt^{h}ī 'friend'*	*sāt^{h}ī*	*sāt^{h}ī*	*sāt^{h}i-yõ*
	Fem	*kənyā 'girl'*	*kənyā*	*kənyā-ẽ*	*kənyā-õ*
Class 3 *i ending*	**Masc**	*pəti 'husband'*	*pəti*	*pəti*	*pəti-yõ*
	Fem	*sidd^{h}i 'success'*	*sidd^{h}i*	*sidd^{h}i-yə̃*	*sidd^{h}i*
Class 4 *ū ending*	**Masc**	*sāɽū'co-brother'*	*sāɽū*	*sāɽū*	*sāɽu-õ*
	Fem	*bəhū 'daughter-in-law'*	*bəhū*	*bəhū-ẽ*	*bəhū-õ*
Class 5 *consonant ending*	**Masc**	*siyār 'jackal'*	*siyār*	*siyār*	*siyār-õ*
	Fem	*cīl 'eagle'*	*cīl*	*cīl-ẽ*	*cīl-õ*

We see in Table 4 that the feminine nouns in Classes 2, 4 and 5 show similar inflectional behaviour as they are marked with *-ẽ* and *-õ* 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 *-õ* or *-yõ* 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)

	Feminine		**Masculine**	
	Class I	**Class II**	**Class III**	**Class IV**
Sg-dir	*null*	*null*	*null*	*null*
Sg-obl	*null*	*null*	*-e*	*null*
Pl-dir	*-yɑ̃*	*-ẽ*	*-e*	*null*
Pl-obl	*-yõ*	*-õ*	*-õ*	*-yõ/-õ*

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

	Class A	**Class B**	**Class C**	**Class D**	**Class E**
Sg-dir	*null*	*null*	*null*	*null*	*null*
Sg-obl	*null*	*null*	*null*	*-e*	*null*
Pl-dir	*null*	*-yɑ̃*	*-ẽ*	*-e*	*null*
Pl-obl	*null*	*-yõ*	*-õ*	*-õ*	*-yõ/-õ*

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[2].

Class B: Includes */ī/, /i/* or */yā/* ending feminine nouns that take *-yə̃* for the features [+pl, -oblique] and *-yõ* for [+pl, +oblique] such as *ləɽkī (girl), šəkti (power)* and *dibiyā (small box), guɽiyā (doll),* etc.

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

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

Class E: Includes masculine nouns that inflect only for the features [+pl, +oblique]. The nouns in this class end with /ū/, /u/, /ī/, /i/ or a consonant. Examples are *ālū (potato), sādhū (saint), mālī (gardener), kəvi (poet), ghər (home), khet (farm),* etc. The */ā/* ending *tatsam* masculine nouns borrowed from Sanskrit such as *rājā (king), pitā (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 *bhāvavācək (abstract) or guŋavācək (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., -õ and -yõ* 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 *-yõ* while all other vowel and consonantal ending nouns take *-õ.* 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 *ləɽkā (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 *moʈər (motor)* refer to *'gāɽī' (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.

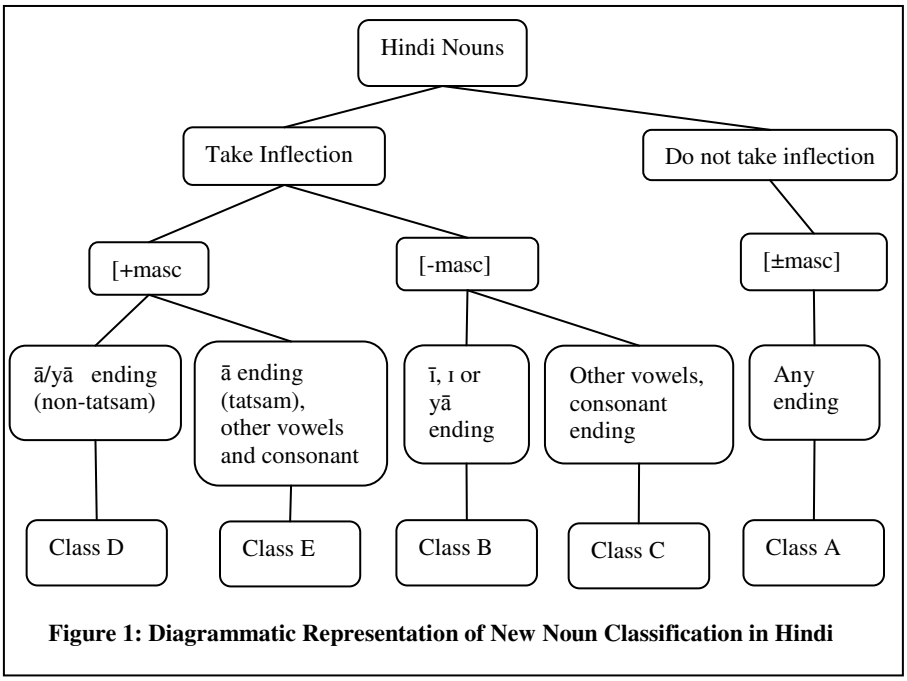

Figure 1: Diagrammatic Representation of New Noun Classification in Hindi

7 Morphological Operations and Hindi Nouns

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-õ (kings-pl-oblique), ləɽki-yə̃ (girls-pl-direct), māli-yõ (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**

[±pl, ±oblique] ↔ null / Class A ---------- *1*
[+pl, +oblique] ↔ -yõ / Class B & E (Stem ending ī or yā) ---------- *2*
[+pl, +oblique] ↔ -õ ---------- *3*
[+pl] ↔ -yə̃ / Class B ---------- *4*
[+pl] ↔ -ẽ / Class C ---------- *5*
[+pl] or [-pl, +oblique] ↔ -e / Class D ---------- *6*
[±pl] ↔ null ---------- *7*
(elsewhere rule)

Rule 1 applies when a noun root is specifically marked for Class A. It inserts *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 *-õ* 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 *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] ↔ null* and *[+pl, -oblique] ↔ null* can be replaced by a single rule, *i.e., [±pl] ↔ null.*

(2) **Impoverishment Rule**
[-oblique] → null / [±pl]

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

(3) **Readjustment Rules**

Stem final /ā/ → null / Class D with -e or -õ ---------- *8*
Stem final /ū/ → u / -ẽ or -õ ---------- *9*
Stem final /ī/ → i / -yə̃ or -õ ---------- *10*

The first readjustment rule (rule 8) deletes the stem final vowel of Class D nouns that take either *-e* or *-õ*, e.g., *ləɽkā-e, ləɽkā-õ* and *sāyā-õ* and create *ləɽke, ləɽkõ* and *sāyõ* respectively. Rules 9 and 10 are not class specific and result in final vowel shortening in nouns (masculine or feminine) that end in either /ū/ or /ī/. Thus, *bəhū-ẽ* and *bahū-õ* become *bəhuẽ* and *bəhuõ* while *ləɽkī-yə̃* and *ləɽkī-yõ* become *ləɽkiyə̃* and *ləɽkiyõ* 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: *ləɽkiyə̃ (girls)*
- Rule (vocabulary item) applicable: *[+pl]* ↔ *-yə̃ / Class B (rule 4)*
  Output after extracting out the suffix → Stem: *ləɽki*, Suffix: *yə̃*
- Readjustment Rule applied: *Stem final /ī/→i/ -yə̃ 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 *ləɽkī (girl)*. If not, try applying another applicable rule.

The actual output of the system for the input words शहरों (šəhərõ) 'cities' and मौक़े (mauke) 'chances' is given below.

(4) Token: शहरों, Total Output: 1

[Root: शहर, Class: E, Category: noun, Suffix: ◌ो◌]
[Gender: +masc, Number: +pl, Case: +oblique]

(5) Token: मौक़े, Total Output : 1

[Root: मौक़ा, Class: D, Category: noun, Suffix: े]
[Gender: +masc, Number: -pl, Case: +oblique]
[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/, /ẽ/, /e/, /ũ/* and /ĩ/ 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)**

[±pl, ±oblique] ↔ null /Class A ---------- *1*
[+pl, +oblique] ↔ -यों/ Class B and E (Stem ending ī, i or yā) ---------- *2*
[+pl, +oblique] ↔ -ों/ Class C and E [NC], Class D ----------*3a*
[+pl, +oblique] ↔ -ओं ---------- *3b*
[+pl] ↔ -याँ/ Class B ---------- *4*
[+pl] ↔ -ें/ Class C [NC] ----------*5a*
[+pl, -oblique] ↔ -एँ / Class C ----------*5b*
[+pl] or [-pl, +oblique] ↔ -े/ Class D [Nā] ----------*6a*
[+pl] or [-pl, +oblique] ↔ -ए/ Class D ----------*6b*
[±pl] ↔ null ---------- *7*

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

(7) **Readjustment Rules (revised)**

Stem final -ा or -आ→ ϕ / Class D [Nā] with -े or -ों ---------- *8*
Stem final -ू → -ु/ -एँor -ओं ----------*9a*
Stem final -ऊ → -उ/ -एँor -ओं ----------*9b*
Stem final -ी→ -ि/ -याँor -यों ----------*10a*
Stem final -ई → -इ/ -याँor -यों ----------*10b*

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

Testing Results	
Total Number of Words in the Testing Corpus	14480
Number of words correctly analyzed	12784
Total number of unidentified words	1696
Total number of unique unidentified words	900

Table 8: Types of Unidentified Words and their Counts

Unique unidentified/unknown words (900)	
Words with incorrect or variant spelling	200
Hyphenated words	350
Missing root entry in the lexicon	350

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.:

क़ैदियों/कैदियों 'prisoners', हफ़्ते/हफ्ते 'weeks', क्रान्तिकारी/क्रांतिकारी 'revolutionists', कम्पनियों/कंपनियों companies', स्तम्भ /स्तंभ 'pillar'

- Hyphenated words:
  दाह-संस्कार 'cremation', वर्ण-भेद 'casteism'
- Incorrect spelling:
  भैंसों (correct spelling: भैंसों) 'buffaloes', कीर्ती (correct spelling: कीर्ति) 'fame', कर्ज (correct spelling: क़र्ज़) 'debt'
- Adjectives/qualifiers functioning as nouns:
  सैंकड़ों 'thousands', तीनों 'all three'

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 *kār-ẽ (car-s), moʈər-õ (motor-s), pepərõ (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

Guru, K. P. (1920). *Hindi Vyakaran.* Kashi: Lakshmi Narayan Press. (1962 edition).

Halle, M. and A. Marantz (1993). 'Distributed Morphology and the pieces of inflection'. In K. Hale and S. J. Keyser (eds.), *The View from Building 20*. Cambridge, MA: MIT Press, 111–176.

Halle, M. and A. Marantz (1994). 'Some Key Features of Distributed Morphology'. In A. Carnie, et al. (eds.), *Papers in Phonology and Morphology*, 275-288. MITWPL 21.

Harley, H. and R. Noyer (1999). 'Distributed Morphology'. *GLOT International 4:4*: 3-9.

Lieber, R. (1992). *Deconstructing Morphology*. Chicago: University of Chicago Press.

Kachru, Y. (2006). *Hindi.* Amsterdam and Philadelphia: John Benjamins, 2006.

Müller, G. (2002). Remarks on Nominal Inflection in German. In: I. Kaufmann & B. Stiebels, eds., *More than words: A Festschrift for Dieter Wunderlich.* Akademie Verlag, Berlin, 113-145.
Müller, G. (2003). Syncretism and Iconicity in Icelandic Noun Declensions: A Distributed Morphology Approach. Ms., IDS Mannheim.
Müller, G. (2004). A Distributed Morphology Approach to Syncretism in Russian Noun Inflection. *Proceedings of Formal Approaches to Slavic Linguistics 12*, 353-373.
Shapiro, M. (2000). *A Primer of modern standard Hindi Grammar*. Motilal Banarsidass Publications.
Upreti, M. L. (1964). *Hindi me Pratayay Vichar.* Agra: Vinod Pustak Mandir.
Vajpeyi, K. (1958). *Hindi Shabdanushasan.* Kashi: Nagri Pracharni Sabha.
Weisser, P. (2006). A distributed morphology analysis of Croatian noun inflection. In G. Müller, & J. Trommer (Eds.), *Subanalysis of argument encoding in distributed morphology. Linguistische Arbeitsberichte* (Vol. 84, pp. 131–142). Universität Leipzig.