Noun Group and Verb Group Identification for Hindi

Smriti Singh[1], Om P. Damani[2], Vaijayanthi M. Sarma[2]
(1) Insideview Technologies (India) Pvt. Ltd., Hyderabad
(2) Indian Institute of Technology Bombay, Mumbai, India
smriti.singh@insideview.com, damani@cse.iitb.ac.in,
vsarma@iitb.ac.in

ABSTRACT

We present algorithms for identifying Hindi Noun Groups and Verb Groups in a given text by using morphotactical constraints and sequencing that apply to the constituents of these groups. We provide a detailed repertoire of the grammatical categories and their markers and an account of their arrangement. The main motivation behind this work on word group identification is to improve the Hindi POS Tagger's performance by including strictly contextual rules. Our experiments show that the introduction of group identification rules results in improved accuracy of the tagger and in the resolution of several POS ambiguities. The analysis and implementation methods discussed here can be applied straightforwardly to other Indian languages. The linguistic features exploited here are drawn from a range of well-understood grammatical features and are not peculiar to Hindi alone.

KEYWORDS : POS tagging, chunking, noun group, verb group.

Proceedings of COLING 2012: Technical Papers, pages 2491–2506,
COLING 2012, Mumbai, December 2012.

1 Introduction

Chunking (local word grouping) is often employed to reduce the computational effort at the level of parsing by assigning partial structure to a sentence. A typical chunk, as defined by Abney (1994:257) consists of a single content word surrounded by a constellation of function words, matching a fixed template. Chunks, in computational terms are considered the truncated versions of typical phrase-structure grammar phrases that do not include arguments or adjuncts (Grover and Tobin 2006). For Abney, chunks are connected subgraphs of a sentence's parse tree. They are defined in terms of major heads and have their own syntactic structure that can be represented in the form of a tree. However a chunk does not include all the descendants of the root node that may be present in the parse-tree of the complete sentence. It only represents the root node (the head of the chunk) and its modifiers (auxiliaries in the case of verbs). Two heads of the same lexical category are not allowed inside a chunk. Consequently, 'Ram's son' in English and 'rām kā betā' in Hindi will have two chunks each [Ram's] [son] and [raam kā] [betā]. Similarly, verb complements are not grouped inside the verb chunk; they form separate chunks. The English sentence 'The bald man was sitting on his suitcase,' can be grouped into three chunks – [The bald man], [was sitting] and [on his suitcase]. The parallel sentence in Hindi 'ganjā ādmī apne sandūk pe baithā thā' will have the chunks [ganjā ādmī] [apne sandūk pe] [baithā thā] in that order.

The task of the chunker is to divide a sentence into chunks leaving out some words that are not grouped into any of the identified chunks. The output of the chunker is a shallow syntactic analysis employing simple, context sensitive grammars to detect the boundaries of syntactic groups such as a Noun Group (NG) or a Verb Group (VG). It identifies major constituents of a sentence without further identifying a hierarchical structure that connects and arranges the chunks (Abney 1991, Ramshaw and Mitchell 1995). The chunked structures (or groups, as we shall refer to them from here on) do not correspond straightforwardly to any structure in a typical phrase-structure analysis. A chunker makes use of the POS information provided by a tagger to form groups. A Noun Group consists of a head noun along with its qualifiers and modifiers (including particles). A Verb Group contains a single main verb and any auxiliaries, negation markers, and focus particles. The grammatical information of a word group depends on the order of its constituent morphemes and the information associated with those constituents. The group identification module makes use of the morphological information and the POS information provided by a morphological analyser and a POS tagger respectively. The group identification module can also be employed before POS Tagging in which case it works with possible POS tags of a given word given in a lexicon.

The focus of this work is Hindi word group identification. We performed a detailed corpus analysis and came up with word grouping rules. Local word grouping in Hindi was first discussed by Bharati et al. (1995). They built a Paninian Parser (using karaka or semantic case relations) that internally uses a morphological analyser as well as a Local word grouper (LWG). Following Bharati et al., Ray et al. (2003) attempted local word grouping using a list of regular expressions to form groups. From the list of ten possible modifier-modified structures discussed by Bharati et al., Ray et al. worked on the five structures that rely only on local modifier-modified relationships and do not need long distance dependencies. Hindi word grouping has also been attempted using statistical models including those by Baskaran (2006) using an HMM based approach and Singh A. et al. (2005) and Dalal et al. (2006) using Maximum Entropy Models. These systems rely very little on linguistic knowledge and instead use a large corpus for automatic learning. For Marathi, a close cousin of Hindi, verb group identification was deployed in a CRF based POS tagging system in Gune H. et. al. (2010). Limited noun phrase chunking has also been done for Turkish (Kutlu M. 2010) and Tamil(Vijay and Sobha 2010). While the focus of our work is on Hindi, the analysis and implementation methods discussed here can be applied straightforwardly to other Indian languages. The linguistic features exploited here are drawn from a range of well-understood grammatical features and are not peculiar to Hindi alone.

2 Need for Group Identification in a POS Tagging System

On analysing the output a CRF (Conditional Random Fields) based POS Tagger, we discovered that most of the systems errors were due to its inability to disambiguate POS tags in the absence of large training corpora. The system needed a detailed group level analysis to resolve the ambiguities between adjective and noun, main verb and auxiliary verb or demonstrative and pronoun. In other words, a granular group level analysis was needed that made use of the morphotactical arrangement both within a word form and in between words. To motivate this further, we provide a detailed error analysis in the following.

a) Demonstrative-Personal Pronoun POS ambiguity: Data-driven learning may not help much in resolving the ambiguity because of a number of qualifiers that may appear between a demonstrative and the head noun. In most cases, a word with the demonstrative-personal pronoun ambiguity is assigned the tag PRON (pronoun) if it is not immediately followed by a noun. Hence, the tagger incorrectly tags some demonstratives (DEM) as pronouns as shown in 1a. In 1b, in contrast, *'us'*appears as a PRON and not as DEM.

1) a) *us* *kāl*-e ghod-e ko rok-o
that-obl black-obl horse-obl ACC stop-imp
'Stop that black horse'

b) us ko roko
He-obl ACC stop
'Stop him'

The first word in 1a should be tagged as Dem while the tagger incorrectly tags it as Pronoun because of a lack of representative training data. Similarly, sentences 2a and 2b are ambiguous for the system. In1a-b and 2a-b, *'us'* and '*ve'* are valid candidates for both DEM and PRON tags. The ambiguity arises for the system because it seeks to resolve it by looking at the words in the immediate vicinity. Note that these sentences are not ambiguous for a native speaker.

2) a) *vo* *kāl*-e ghod-e ko rok rəh-*ā* hai
he black-obl horse-obl ACC stop prog-masc,sg be-pres
'He is stopping the black horse'

b) vo *kālā* ghod*ā* so rəh-*ā* hai
that black horse sleep prog-masc,sg be-pres
'That black horse is sleeping'

The TAM (tense, aspect and modality) information of the verbs in the two sentences can also help in resolving the ambiguity but requires subject-object information in the sentence along with a syntactic analysis of the sentence.

b) Adjective-Noun ambiguity: An adjective may function as a head noun if the noun is dropped, and bears the same inflection as the nominal head, as may be seen in 3 and 4.

3) əcch-e k*ām kā nətijā əcchā* *nikə*l-t-*ā* *hai*
good-obl deed of result good turn-hab,masc,sg be-pres
'D*o good have good'*

4) *əcch-e* *kā nətijā əcchā nikə*l-t-*ā* *hai*
good-obl of result good turn-hab,masc,sg be-pres
'Do good have good'

If the case marker orpostposition immediately follows the adjective, it is treated as a nominal head. Since the occurrence of əcche (or any other adjective) as an adjective is more likely than its occurrence as a noun in any learning corpus, this will result in incorrect learning and consequently, in incorrect tagging. NG identification rules help in resolving such ambiguity by using the featural information of the NG constituents.

d) Noun-Verb ambiguity: Many nouns may appear as verbs (even when inflected) and vice versa in Hindi[1]. Verbs may appear as verbal nouns in their infinitival form and may function as nouns. Nevertheless, a verbal noun retains many of its verbal properties. While functioning as a noun, it appears only in the 'singular, oblique' and the 'singular, direct' cases and inflects like other */ā/* ending masculine nouns in the language.

5) tair-*nā* *bəhut lābhkārī hai*
swim-Inf very beneficial be-pres
'Swimming is very beneficial'

6) tair-n-e ke b*əhut lābh* ha*ĩ*
swim-Inf-obl Poss many benefits be-pres,pl
'Swimming has many benefits'

Infinitival verbs as either main verbs or verbal nouns have identical forms. The POS ambiguity is easy to resolve when the verbal noun is in the oblique and is followed by a postposition as shown in 6 above. More difficult are sentences where it appears in the direct form. In 7 the verb j*ānā* appears inside a VG and should be tagged as an infinitival verb. While in 8, it appears as a verbal noun and is also modified by a possessive pronoun *merā.* These two occurrences of *jānā* as a noun and an infinitival verb yields POS ambiguity (N or V). However, when an infinitival verb is immediately preceded by a possessive pronoun or a genitive postposition, as in 8, it should be tagged as a verbal noun, and the NG information can be successfully exploited by the POS tagger.

7) mujh-*e [jā-nā hai]*
I-DAT go-Inf be-pres
'I have/want to go'

8) *[merā jā-nā] zərūrī hai*
My go-Inf important be-pres
'For me to go is important '

3 Noun Groups in Hindi

Nominal groups are defined by Halliday (1977:7) as *"...nouns plus their determi*ners and any *other modifiers....".* Specifiers and modifiers/qualifiers are optional while the headword (noun) constitutes the obligatory element in the structure of an NG. Specifiers can be determiners, ordinals, and cardinals. Qualifiers include adjectives, prepositional, or postpositional groups or a relative clause. The idea of an NG is not far removed from that of a syntactic NP but it does not straightforwardly match the constituents of syntax. The constituents of a group always appear in a particular default order and this is subject to cross-linguistic variation. Hindi being a head-final language, the head of an NG is the rightmost constituent in the group. The head may be preceded

[1] For example, in the sentence 'merekə*ĩ*kh*ātehaĩ*'' the token'kh*āte'* isambiguous for NOUN (pl, direct) and VERB (habitual, pl, masc/fem).

by the words that belong to pre-nominal categories. NGs are formed around a noun or a pronoun that acts as a nucleus in the group. Types of Hindi NGs include:

- N, e.g., mez (table)
- Dem Pron+N, e.g., vo mez (that table)
- Poss pronoun+N, e.g., *merā kəmrā* (my room)
- Adj+N, e.g., sundər l*əɽkī* (beautiful girl)
- Dem pron+Adj+N, e.g., vo sundər l*əɽkī* (that beautiful girl)
- Card+N, e.g., *cār* ghoɽe (four horses)
- Ord+N, e.g., *dūsrā ləɽkā* (second boy)
- Non-Spec Det+N, e.g., *kuch kitābẽ* (some books)
- Det+Adj+N, e.g., *kuch purānī kitābẽ* (some old books)
- Inten+Adj+N, e.g., b*əhut purānī kitābẽ* (very old books)
- Pron, e.g., ve (they), v*əh (he/it/she), tum (you), āp* (you-honorific)
- N or Proper N (postpositions fuse with Hindi Pronouns; they are not written as free words) followed by a simple postposition or a compound postposition, e.g. ləɽke **ke lie** (for the boy), kəmre **mẽ** (in the room), mez **pər** (on the table)
- Part/Discourse marker+N, e.g., l*əɽkī* **hī** (girl only), l*əɽkī bhī (girl too), pānī tə*k (water even)

The ordering of the constituent elements of a Hindi NG can be captured using morphotactical rules and a few additional constraints. For example, the end of a Noun Group may easily be marked when the group is Oblique, i.e. when a post-position appears immediately after the head noun. NGs where the head is not directly followed by a postposition require deeper analysis to mark the group boundary. In addition to consulting standard Hindi grammar texts like Kachru(2006), we performed a detailed corpus analysis to determine the word grouping rules. Candidate constituents of the Hindi NG may be placed in five sets as shown below. Optional elements are marked by parentheses. If two or more elements always appear together, they are shown as a single unit within parentheses. Curly brackets are used to show optionality between constituents competing for a single position.

Set 1 includes possessive demonstrative pronouns. Both are optional elements of a Hindi NG and may appear in any order with or without the other. For example, both *vo tumhārī mīthī bātẽ* (those your sweet word*s), tumhārī vo mīthī bātẽ* (your those sweet words) are possible. The possessive followed by a demonstrative pronoun is the canonical order, while the reversed order is a stylistic, poetic construction. The optionality and the order of Set 1 elements is shown in 9.

9) ((Demonstrative) (Possessive)) OR ((Possessive) (Demonstrative))

Thus, any of the following outputs are valid:

- Both items are optional – (*vo tumhārī) mīthī bātẽ*
- Both items may appear together - *vo tumhārī mīthī bātẽ* or *tumhārī vo mīthī bātẽ*
- One item appears without the other - *vo mīthī bātẽ* or *tumhārī mīthī bātẽ*

Set 2 includes intensifiers and numerals. A numeral may be of the type - approximate, fractional, universal quantifier, indefinite quantifier, multiplicative, aggregative, ordinal, cardinal and measure word. Kachru (2006:133) provides the ordering among Hindi quantifiers as in 10.

10) approximate-cardinal-collective-ordinal-multiplicative/fractional-measure

We modified this ordering to capture the arrangement of numerals in a more elaborated way (in Figure 1 below).

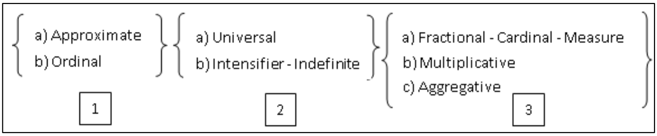

Figure 1: Ordering of quantifiers in a Hindi NG

The categories within curly braces are mutually exclusive while those separated by a hyphen '−' can appear one after another in a sequence. The ordering suggests that:

- Approximate quantifier and ordinal (e.g., *lǝgbh*ǝg dūsrā vyǝ*kti *'around second man'*), universal and indefinite quantifier (e.g., *sǝ*bhī kuch log 'all few people'*), cardinal and aggregative (e.g., *do donõ log), aggregative and multiplicative (e.g., **donõ dugunā*), and fractional and aggregative/multiplicative quantifier (e.g., **ādhā donõ, *ādhā dugunā*) are mutually exclusive
- An intensifier may precede an indefinite quantifier but not a universal quantifier (e.g., bǝhut k*ǝm log 'very few people' (intensifier*-indefinite quantifier), bǝhut s*ǝbhī log 'very all people'* (intensifier-universal quantifier))
- Fractionals do not appear with aggregative or multiplicative quantifier (e.g., **ādhā donõ, *ādhā dugunā*)

Set 3 includes adjectives (including imperfective or perfective verbal adjectives) and are optional in an NG. Many adjectives may appear inside an NG recursively. Examples include, *bhāgtā huā kālā* gho*ṛā (running blackhorse), bhāgtā huā ghoṛā (running horse), kālā ghoṛā (black horse),* th*ǝkā huā kālā ghoṛā (tired black horse), thǝkā huā ghoṛā (tired horse*).

11) ((Verbal Adjective) (Adjective))

The adjectives are internally ordered based on the adjective type. Those that denote shape, color, size or the origin of a noun are known as fact adjectives. Those that refer to a noun's quality or those that denote a speaker's opinion appear before fact adjectives. The order followed by different kinds of adjectives is quality-size-age-shape-color-origin material, as in l*ǝmbī kālī reshmī bǝnārǝsī sā*dī *(long black silk* banarasi saree), n*ǝyā khushhāl bhārtiyǝ* s*ǝmudāyǝ* (new happy Indian community), etc.

Set 4 includes nouns and pronouns (except the demonstrative) which are obligatory members (Heads) of an NG. They are the right most element of the group with the exception of particles and postpositions that appear after them.

Set 5 includes postpositions that form oblique NGs. Postpositions may be primary (such as ne, ko, ke, etc.) or compound (such as *'ke bā*d*', 'ke sā*t', etc.) and are optional.

Particles (focus, emphatic, etc.) may appear at many places (even at the end) within an NG and we have not put them in any set. The particles to and *bhī* may appear only at the end of an NG.

The complete ordering of constituents within a Hindi NG is given in the expression in 12 where () show optionality, [] represents a set and * stands for zero or more repetition.

12) NG = (Set 1) (Set 2) (Set 3)* Set 4 (Set 5) (Particle)

3.1 Procedure for Noun Group Identification

The Noun Group identification module attempts to isolate the basic non-recursive NG that includes only one head and its specifiers and modifiers. The input to the algorithm is the output of a morphological analyser. For each word, the morphological analyser gives the stem,andthe set of suffixes along with the associated morphological properties. A look-up is performed in a lexicon to retrieve the set of possible POS tags for each stem. The NG is built from right to left in a given sentence. As discussed in the previous section, we formulated five sets of constituents that contain different lexical categories that combine in various ways to form NGs in Hindi. Set 4, the head marks the right end of an NG (neglecting any postpositions and particles).

Sets 1, 2 and 3 contain categories which mark the left end of an NG. Processing from right to left, once the system encounters a Set 4 element, it starts to look for Set 3, Set 2 and Set 1 elements appearing to the left of the head in that order. By 'finding a Set X element', we mean 'finding a stem whose potential POS tag list in the lexicon contains a POS tag belonging to Set X.' The potential candidates are considered to be members of the NG. As soon as any word of a lexical category other than those mentioned in Sets 1, 2 and 3 is encountered, the NG is considered closed. The previous word marks the left end of the NG in such a case. The number, gender and case information for nouns, demonstratives and pronouns are required at each step to select or reject a potential POS tag. This information is extracted from the output of the morphological analyser. The pseudo code for NG identification is given below.

Steps for NG Identification

```
1. For all tokens, processing goes from right to left
        1a. Look for a post-position or a Set 4 element to start an NG
        1b. If Set 5 member, i.e., a postposition is found
                1b (i) Oblique NG has started
        1c. If Set 4 element is found
                1c (i) Direct NG has started
        1d. If a Demonstrative pronoun is found
                1d (i) Consider it as a Pronoun (head)
2. If oblique NG has just started with a Set 5 element, i.e., with a postposition
        2a. Look for a Set 4 element
        2b. If Set 4 element is not found; find the list of possible POS tags for the
                current word
        2c. If a POS Tag appears in the possible POS Tags' list and also in Set 4
                2c (i) Assign the tag which is common to both.
        2d. If there is no common element in the list and Set 4s
                2d (i) Assign the tag other than PP to the next word using
               the list of possible tags for it.
3. If any NG has started
        3a. Look for a Set 3 and/or Set 2 and/or Set 1 element
        3b. If Set 3, 2 and 1 elements are found
                3b (i) The NG includes the current word
        3c. If set 3, 2 and/or 1 elements are not found
                3c (i) The NG has already ended with the previous word
4. If any NG is completely identified
        4a. Apply rules to check the agreement between modifiers/qualifiers and their head and
        do corrections if necessary
5. Start looking for the next NG
```

In what follows, we give an example of how the NGI helps correct a POS Tag error.

13) *ve pūr-e māml*-e ko *suljhā-nā cāh-te haĩ*
They whole-obl matter-obl ACC solve-Inf want be-pres-pl
‘they want to solve the whole matter’

For 13 the tagger produces the output as [DEM ADJ NN PP VM VAUX]. *‘ve’* is tagged as a DEM instead of PRON. Scanning right to left, the NG identified is (*ve pūre māmle ko*). Now the computational rules are applied to make any POS corrections required. By the first rule for oblique NG,reading the rule from right to left, we find a PP ‘ko*’* followed by a noun ‘*māmle’* in the oblique case. ‘*pūre’* is allowed in the NG as its category and features warrant its being a Set 2 member. ‘*ve’*may be a Set 1 member and may mark the left end of the NG and may be a demonstrative or a pronoun. The tagger may tag ‘*ve’*as DEM ‘demonstrative’ but, as a demonstrative, it does not concord with the head noun for the relevant case feature. Thus, the tag DEM is rejected and PRON is selected.

4 Verb Groups (VGs) in Hindi

A Hindi VG includes a single main verb root followed by a sequence of inflectional suffixes and/or auxiliary verb sequences. The group contains various verbal morphemes that centre on a single event. The verbal morphemes occur in a fixed order and are subject to several grammatical and semantic constraints. Some examples of Hindi VGs are *[khā-yā]* (eat-past) and *[khā-yā* gə-*yā* hai] (eat-perf passive-perf,sg be-pres).While analysing Hindi VGs, we have not considered complex predicates such as conjunct verbs (as *ārəmbhkər ‘start’ (literally ‘start do’)*) or compound verbs (*‘*kəɽd*āl’ (‘somehow finish’)* (Chakrabarti et al.2007, Chakrabarti et al.2008, Begum et al. 2011). Here, a main verb is a single verb root that appears with associated inflectional morphemes.

4.1 Identifying Verb Group Boundaries

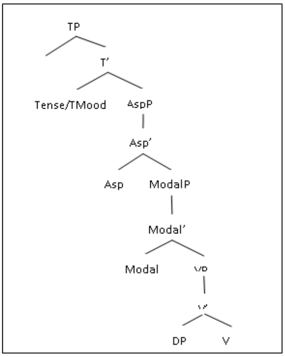

Figure 2: Order of Verbal Elements in a Hindi Verb Group

A VG boundary is marked using the order in which Hindi verbal elementsarrange themselves . The linear order of the major grammatical categories within a Hindi VG is Verb-Aspect-Tense/Mood as shown in Figure 2. The grammatical properties for which Hindi verbs inflect are tense, aspect, mood, modality, gender, number, person, honoroficity, voice and finiteness. These properties are realised analytically or periphrastically (either as suffixes or as auxiliaries). A

Hindi verb group must always begin with a main verb root with or without a suffix. Once the main verb is identified, the verb group is assumed to have begun. Scanning from left to right, the main verb may be followed by a string of intermediate verbal suffixes and auxiliaries until a must-end VG marker is encountered. These elements broadly follow the linear order in 14, though with co-occurrence constraints that are listed towards the end of this section.

14) *Verb Root−*Infinitive*/Passive−Modal Auxiliary−Aspect−Tense−*Mood

The three kinds of morphemes are called Start markers, Intermediate markers and Must-end Markers and are shown in Figure 3. Particles and negation markers are also allowed to appear inside a VG.

Start Marker	Intermediate Markers		Must End Markers
	Possible End Markers	Must-Continue Markers	
Main Verb (Root)	Necessity	Ability/Probability,	Present Tense
	Perfective-gen-num	Obligation/Permission	Past Tense
		Habitual/Progressive	Future+gen-num

Figure 3: Boundary Markers for Hindi VG

a) Start markers

A Hindi VG 'start' marker is always a verb root whether inflected or uninflected. All verbal auxiliaries may also be considered as start markers. Since the identification begins from left to right, the first instance of a free verbal morpheme is always the root or main verb. In rare cases (poetic constructions), verbal auxiliaries appear before the main verb in a Hindi VG, as for examplein 15, where the tense auxiliary precedes the main verb and starts a VG. This scrambling is usually seen only with tense auxiliaries. Such reordering with aspectual auxiliaries is even rarer (see 16). We exclude here the identification of these rare VGs. If such constructions are encountered, the system will identify them as two separate VGs, albeit incorrectly.

15) vo roz mujh-se [hai mil-*tā]*
he everyday I-DAT be-pres meet-hab
'He meets me everyday'

16) vo mujh-se [rəh-*ā* *hai* *mil]*
He I-DAT prog-masc,sg be-pres meet
'he is meeting me'

b) Intermediate markers

These markers include two kinds of morphemes, 1) possible-end markers and 2) 'must continue' markers. Possible end markers are those which may end a VG such as the perfective marker or the modal auxiliary for necessity (preceded by an infinitive-gender, number sequence). These morphemes, however, may be followed by other morphemes to further extend the VG. For example, the perfective marker may be followed by the past or the present tense auxiliary as in vo *āyā* 'he came', *vo āyā* hai 'he has come' and vo *āyā thā* 'he had come'. Similarly, the modal auxiliary for necessity may be followed by the past tense auxiliary, such as *usko ānā cāhiye (thā)* 'he should (have) come' and the subjunctive marker may be followed a future-person, number marker, such as *khā*-*ũ*-g-*ā* *'eat*-subjunctive-will-*person, number'*. The 'must-continue' markers, however, must be followed by other verbal morphemes in order to complete the VG. Details of such markers are given below in Table 1 along with their inflections.

Possible End-Markers	
Modal Auxiliary	चाहिए *(cāhie)* 'should'
Aspect: Perf+gen-num	-या *(-yā),* -◌ा *(-ā),* -आ *(-ā),* -◌ी *(-ī),* -◌े (-e), -ए (-e), -ई *(-ī),* -◌ी◌ं (-ĩ), -ईं *(-ĩ)*
Subjunctive	-◌ुँ (-ũ), -ऊँ (-ũ), -◌े (-e), -ए (-e), -◌ं *(-ŋ),* -◌े◌ं (-ẽ), -एँ (-ẽ), -◌ो (-o) ,-ओ (-o)
Must-Continue Markers	
Aspect: Habitual	-त (-t), Progressive रह (rəh), Completive चुक (cuk)
Modal Auxiliaries: Ability/probability	सक (sək), ability: पा *(pā), o*bligation: पड़ (pəɽ), permission: दे (de)
Passive	या (-y*ā)*/ यी (-y*ī)* / ये (-ye)/ जा (-j*ā*)
Must-End Markers	
Future+gen-num	-गा *(-gā),* -गी *(-gī),* -गे (-ge)
Mood:Imperative	null, -◌ो (-o) ,-ओ (-o),◌िए (-ie), इए (-ie), जिए (-jie), –ना *(-nā)*
Tense Auxiliary: Present	है (hai), हैं *(haĩ),* Past: था *(thā),* थे (the), थी *(thī),* थीं *(thĩ)*
Mood:Conditional	-त- (-t-)

Table 1: Intermediate Markers

As shown in 14, the verbal elements appear in a specific order. This ordering issubject to a number of constraints as listed below:

Specific Constraints within a Hindi VG

a) The modal auxiliary *chāhie* must be preceded by an infinitive (with gender-number) marker, such as *khā-nā cāhie (खा-ना चाहिए*). It may be followed neither by an aspect marker (17) nor by a present tense or future tense marker (18). It may only be followed by a past tense auxiliary (19).

17) **chāhie rəh/cuk* (aspect)

18) **chāhie hai/ chāhie-gā* (pres, future)

19) *chāhie thā* (past)

b) In the absence of a modal auxiliary, an infinitive must be followed by a mood or a tense marker (as shown in the examples below in 20 and 21). The expression in 21 where the mood or tense marker is optional is ungrammatical, as in 22.

20) *khā*-ne de-*tā hai*

21) *khā-nā pəɽ-tā hai*

22) **khā-nā (hai/thā/hogā/hotā)*

c) The modal auxiliary *sək* cannot be followed by the perfective marker, the progressive auxiliary *rə*h and the completive auxiliary cuk as shown below in 23-25. It can only be followed by a habitual aspect marker or by a subjunctive marker as in 26 and 27.

23) **khā sək rəhā hai* (progressive)

24) **khā sək-ā hai (perfective)*
25) **khā sək* cuk*ā hai (completive)*
26) *khā sək-tā hai (habitual)*
27) *khā sək*-e (subjunctive)

d) No modal auxiliary may precede the completive auxiliary cuk

28) **khā pā cukā hai*
29) **khānā pəɽ cukā hai*

e) Infinitive marker –*न*-(n), all aspectual markers, past tense auxiliary, conditional mood marker -*त*-(t) and future marker -*ग*-(g) must be followed a gender-number marker as in ता (tā), ती (tī), ते (te), रहा (rəhā), रही (rəhī), रहे (rəhe), चुका (cukā), चुकी (cukī), चुके (cuke), ना (nā), नी (nī), ने (ne), गा (gā), गी (gī), गे (ge).

4.2 Procedure for VG Identification

A VG is identified by scanning the sentence from left to right. The expression given in 30 below is used to detect VGs in a given sentence.

30) Start Marker (Intermediate marker)* Must-end marker

Thus, the start-marker and the must-end markers are obligatory to form a VG while intermediate markers are optional and may recurse (*). The three types of markers were shown in Figure 4 in the previous section. Particles and negation markers may also appear inside a VG. The VG identifier uses the root, suffixes and the morphological features supplied by the morphological analyser and the POS tags assigned by the POS tagger. A VG begins as soon as a verb is scanned and the following morphemes are marked as its suffixes or auxiliaries. The identified verb root may be locally POS ambiguous, i.e., noun or verb (*khānā*'food'and'to eat'), or main verb or auxiliary verb (rəh'live'and 'progressive auxiliary'). The appropriate tag is selected by applying the regular expression on the verbal morphemes. If the sequence of the markers is allowed by the expression, they are included in the VG. The identification continues until a must-end marker is encountered. Once the end of the VG is marked, the group members are assigned fresh, disambiguated POS tags. The head of the VG is assigned VM while the auxiliaries are assigned VAUX along with the TAM features that they express. Some examples are given next.

<u>Types of major POS ambiguity:</u>

a. Main Verb or Auxiliary Verb
b. Main Verb or Noun
c. Main Verb or Postposition

31) [rəh *rəh-ā hai]*
live prog-masc,sg be-pres
'is living'

32) [kər cuk-*ā thā]*
do comp-masc,sg be-past
'had done'

33) kər [cuk-*ā* *de*-g-*ā]*
tax pay-masc,sg give-fut-masc,sg
'will pay the tax'

In 31, *rəh* appears as the progressive aspectual auxiliary as well as a main verb ('live'). Often a POS tagger is unable to resolve this ambiguity in the absence of contextual information. In32, *kər* is ambiguous between being a verb and a noun. As a main verb, it means 'do' and as a noun, it means 'tax'. In order to resolve this POS ambiguity, the system requires the information that when cuk appears as an auxiliary and is followed by a tense auxiliary, it requires preceding main verb. This information rules out the possible tag Noun and leaves Main Verb as the correct one. This information may yield a faulty analysis for the expression in 33. The system will consider *kər*to be a part of the VG and will output the VG as *kərcukādegā*. We require a morphotactical constraint that prevents the completive aspectual auxiliary cuk from being followed by the modal auxiliary de,. We must note that these constraints are ad-hoc and may not always produce correct POS tags.

Secondly, suffixes too may be ambiguous. For example*, '-t'* attached to the stem *ā(*come) may indicate either the habitual aspect or the conditional mood. This ambiguity may be resolved by using the regular expression and by looking at the next morpheme. For example, in 34, the suffix -t- is rejected as being a conditional mood marker as it belongs to the category of must-end markers and cannot be followed by any other verb morpheme (except for the gender-number marker). On the other hand, the habitual -t- may be followed by a tense auxiliary.

34) *bādəl roz [ā*-**te** the]
Clouds everyday come-hab be-past
'Clouds used to form everyday'

During the process of VG identification, feature agreement among elements of the group is also checked. Many invalid sequences are rejected using feature combination rules. For example, *'bhāīthā'* in 36 unlike in 35) cannot be a verb group. It is instead a noun-verb sequence since the masculine gender of the tense auxiliary *thā* does not agree in gender-number with main verb (*bhā 'like'*) marked for feminine gender using *-ī*. On the other hand, *bhāī* (brother) may be a noun with which the gender of the verb (masculine) agrees. The VG identifier thus rejects the Verb tag for the word *bhāī* and retags it as a Noun.

35) *'vo merā bhāī thā'*
he my brother be-past-masc
'He was my brother*'*

36) **'bhāī thā'*
like-past-fem be-past-masc
'was liked*'*

Another example where feature checking resolves the POS ambiguity is given in 37 below where the word *'liye'* is POS-ambiguous between a Verb (take-past-pl) and a Postposition (for). If a verb, the form liye should be plural but the tense auxiliary does not agree with it for number(singular, in this case). Thus, it cannot form a VG. By discarding the Verb tag, it is instead assigned the Postposition tag. The VG is formed only with hai 'is'. For the sentence in 38, the word *pāī* may belong to one of two POS categories - Noun (penny) or Verb (found-fem, sg). According to the VG identification rules, negation may appear inside a VG but the perfective must be followed by either a tense marker or a mood marker. The given sequence does not conform to the rule and thus the tag Verb for *pāī*is rejected and a tag Noun is assigned instead.

37) un-*kī* *yojnā shāntipūrnə* uddeshy-õ ke **liye** hai
their plan peaceful aims-obl for be-pres
‘Their plan is for peaceful aims’

38) ve ek ***pāī*** n*əhĩ* *le*-te the
they one penny not take-hab be-past
‘They would not take a single penny’

5 Performance Evaluation

We use a CRF based POS Tagger. Without NGI/VGI, the features used for the POS-Tagger include (a) Tag ambiguity scheme from the dictionary, (b) suffix given by the stemmer, (c) prefix and suffix character streams of size one and two, (d) previous word’s suffix and (e) tag ambiguity scheme for previous and next word. We tried NG and VG identification at two different places, before and after CRF. When the NGI/VGI module is run before CRF, its output is used as features supplied to CRF and the tags assigned by CRF are considered final. The tag ambiguity scheme of the NG/VG members is simply replaced by the tags given by NGI/VGI modules. On the other hand, when NGI/VGI follows CRF, then NGI/VGI overwrites the tags assigned by CRF.

The Hindi POS Tagger was tested on a corpus of 66,990 words, which is a subset of the BBC Hindi news corpus (downloaded from http://www.bbc.co.uk/hindi) and the IIIT Hyderabad corpus. We partitioned the corpus into four testing folds. The accuracy of the CRF based POS Tag system using Verb Group and Noun Group Identification rules for the four folds are as follows:

Experiment	**Average Accuracy of 4 folds**
CRF	95.18%
CRF + NGI after	95.67%
CRF + VGI after	95.73%
CRF + NGI after + VGI before	95.87%
CRF + NGI after + VGI after	95.26%

Table 2: Experimental Results

We find that while both NGI and VGI help improve accuracy, the best performance is obtained when VGI is applied before CRF and NGI is applied after CRF. It is interesting that applying both NGI and VGI after CRF does not help very much since the errors from one module result in multiple, cascading errors in the second module as the tags given by the first module are considered final. When we apply one module before CRF, then the CRF still gets a chance to overwrite the wrong tags as CRF treats VGI tags as features rather than as final tags.

While a 15% error reduction (from 4.72% to 4.1%) may not appear much numerically large, it should be noted that removing the final 5% of errors is an uphill task with corpus inaccuracies, annotator disagreement, and long distance dependencies dominating. Some of the challenging examples are:

39) mæṭʃ 48-48 ovərõ k*ā* kər diy*ā* gəy*ā* hai
match 48-48 overs of do has been be-pres
‘Match has been made of 48-48 overs’

Here, the verb group is identified as (diyā gəyā hai). (kər) is marked as a verb whereas in 40, it appears as a noun.

40) mæt∫ kā kər diyā gəyā hai
match of tax give-past has been
'Tax has been given/paid for the match'

Another example is that of long distance dependency of the possessive marker:

41) unkā yeh bhī kehnā hai ki
They-ACC this also saying be-pres that
'They also said that'

The verb group is identified as (kehnāhai), whereas (kehnā) is a noun which should co-occur with the preceding possessive. But the possessive pronoun (unkā) is not adjacent to (kehnā).

42) tīm ne spænish līg lā līg kā khitāb jītā
Team-ERG Spanish League Lā Liga of prize win-past
'The team won the Spanish League La Liga title*'*

In 42 (La Liga) is a proper name but as per the morphological analysis (lā) only qualifies to be a verb.

In summary, even with detailed rules for NG and VG identification, there is little improvement in the accuracy of the tagger as (1) our Morphological Analyzer is not able to analyze Compounds (both Verbs and Nouns) and Conjunct verbs as single units unless they are stored in the lexicon,(2) because some of the tags show real ambiguity in a given sentence and 3) because the MA fails to recognize and analyse unknown or foreign words that are not listed in the lexicon.
Our results compare favourably with the 93.45% accuracy reported in Singhet al. (2006) for a CN2 based tagger forthe Hindi BBC news corpus. Guneet al. (2010) report 94% accuracy for CRF on Marathi using a corpus of size 20K. They did not implement NGI but only VGI. They found that use of VGI did not improve the accuracy since not much VM-VAUX ambiguity (their main focus) remained after applying CRF.

6 Conclusions

We have presented algorithms to identify Hindi Noun and Verb Groups by using morphotactical information and the constraints that apply to the constituents of these groups. We also provided the list of grammatical categories and their markers that may appear inside a group and discussed ways in which these markers may be arranged. Group Identification enabled the resolution of major POS ambiguities. The identified groups may also be used at a later stage, i.e., in parsing or in language generation. We cannot handle all the POS ambiguous cases (that involve scrambling or those that are structurally ambiguous) where immediate contextual rules do not help. However, using the ordering among the major categories and their possible combinations, we have tried to present ways that can be applied to other languages equally well. The methods are especially beneficial for languages with meagre corpora or other NLP resources. Since a system will not be able to learn patterns that might be absent in small training corpora, with the useof morphological patterns that govern the ordering of the elements inside a group, a large number of ambiguities and errors may be avoided at a first pass.

Acknowledgements: We would like to thank Nikhilesh Sharma and Neha Gupta for implementing the group identification system discussed here.

References

Abney, S. (1994). “Parsing by Chunks.” In Principle-Based Parsing, eds. B. Berwick, S. Abney, and C. Tenny, 257-278. Dordrecht: Kluwer Academic Publishers.

Baskaran S. (2006). “Hindi POS Tagging and Chunking.” In the Proceedings of NLPAI Machine Learning Contest. Mumbai, India, June.

Begum, R., Jindal K., Jain A., Husain S., Sharma D. (2011). “Identification of Conjunct verbs in Hindi and its effect on Parsing Accuracy.” 12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing).

Bharati, A., Chaitanya V. and Sangal R. (1995). “Natural Language Processing: A Paninian Perspective.” New Delhi: Prentice-Hall of India.

Chakrabarti, D., Mandalia H., Priya R., Sarma V. and Bhattacharyya P. (2008). “Hindi Compound Verbs and their Automatic Extraction”, In Proc. of Computational Linguistics Conference (COLING), Manchester, UK.

Chakrabarty, D., Sarma V. and Bhattacharyya P. (2007). Complex Predicates in Indian Language Wordnets, Lexical Resources and Evaluation Journal, 40 (3-4).

Dalal, A., Nagaraj K., Sawant U. and Shelke S. (2006). “Hindi Part-of-Speech Tagging and Chunking: A Maximum Entropy Approach.” In the Proceedings of the NLPAI Machine Learning Workshop on Part Of Speech and Chunking for Indian Languages. Mumbai, India.

Grover, C. and Tobin R. (2006). “Rule-based Chunking and Reusability.” In the Proceedings of LREC 2006, 873-878. Genoa, Italy.

Gune H., Bapat M., Khapra M. and Bhattacharyya P. (2010). “Verbs are where all the Action Lies: Experiences of Shallow Parsing of a Morphologically Rich Language”, Computational Linguistics Conference (COLING), Beijing, China.

Halliday, M. A. K. (1977). “Text as Semantic Choice in Social Contexts.” In Grammar and Descriptions (Studies in Text Theory and Text Analysis), eds. T. A. van Dijk and J. Petofi, 176-225. New York: Walter de Gruyter.

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

Kutlu M. (2010). “Noun Phrase Chunker for Turkish using Dependency Parser”, MS Thesis. Department of Computer Engineering, Bilkent University.

Ramshaw, L. A. and Mitchell, P. M. (1995). “Text Chunking Using Transformation-Based Learning.” In the Proceedings of the Third ACL Workshop on Very Large Corpora, 82-94. Cambridge, MA, USA.

Ray, P. R., Harish V., Basu A. and Sarkar S. (2003). “Part of Speech Tagging and Local Word Grouping Techniques for Natural Language Parsing in Hindi.” In the Proceedings of (ICON). Mysore, India.

Singh, A., Bendre S. M. and Sangal R. (2005). “HMM Based Chunker for Hindi.” In the Proceedings of International Joint Conference on NLP.

Singh, S., Gupta K., Shrivastava M. and Bhattacharyya P. (2006). “Morphological Richness Offsets Resource Demand – Experiences in Constructing a POS Tagger for Hindi.” In the Proceedings of the COLING/ACL-2006, 779-786. Sydney, Australia, July.

Vijay S. and Sobha D. (2010), "Noun Phrase Chunker Using Finite State Automata for an Agglutinative Language", In the Proceedings of the Tamil Internet Conference.