Annotation Guidelines for Hindi

Microsoft Research Labs India Pvt. Ltd.
Bangalore

Contents

Table 1.1: Category and their Types

CATEGORY	Attributes
NOUN	Common
	Proper
	Verbal
	Spatio-temporal
VERB	Main
	Auxiliary
PRONOUN	Pronominal
	Reflexive
	Reciprocal
	Relative
	Wh
NOMINAL MODIFIER	Adjective
	Quantifier
DEMONSTRATIVE	Absolute
	Relative
	Wh
ADVERB	Manner
	Location
PARTICLE	Coordinating
	Subordinating
	Classifier
	Interjection
	Others
Punctuation	
RESIDUAL	Foreign word
	Symbol
	Others

Table 1.1: Category and their Types

Table 1.2: Attributes and their Values

ATTRIBUTE\ SYMBOL	Value\symbol				
GENDER\ GEN	Masculine\ mas	Feminine\ fem			
NUMBER\NUM	Singular\sg	Plural\pl			
PERSON\PER	First\1	Second\2	Third\3		
TENSE\TNS	Present\prs	Past\pst	Future\fut		
CASE\CS	Direct\dir	Oblique\obl			
CASE MARKER\CSM	Ergative\erg	Accusative\ acc	Instrumental ins	Dative\ dat	Genitive\gen
	Sociative\soc	Locative\ loc	Ablative\ abl	Benefactive\ bnf	Vocative\ voc
ASPECT\ASP	Simple\smp	Progressive\ prg	Perfect\pft		
MOOD\MOOD	Declarative\ dcl	Imperative\imp	Habitual\hab		
FINITENESS\FIN	Finite\fin	Non-finite\nfn	Infinite\ifn		
DISTRIBUTIVE\DSTB	Yes\y	No\n			
EMPHATIC\EMPH	Yes\y	No\n			
NEGATIVE\NEG	Yes\y	No\n			
HONORIFICITY\HON	Yes\y	No\n			
NUMERAL\NML	Ordinal\ord	Cardinal\crd	Non-numeral\ nnm		

Table 1.2: Attributes and their Values

Annotation Guidelines for Hindi

Introduction

This is a guideline for annotating Hindi text with Parts-of-Speech (POS) tags in accordance with the IL-POSTS (**I**ndian **L**anguage **POS T**ag **S**et); the hierarchical POS tagset framework developed by Microsoft Research India Private Limited in collaboration with linguists and NLP experts. ILPOSTS is the framework which can be used in POS tagging of all Indian languages. We will present the guidelines for Hindi after a short description of ILPOSTS which is the motivation behind this guideline.

The goal of IL-POSTS[1] is to provide a common POS-tagset framework for Indian Languages that offers flexibility, cross-linguistic compatibility and reusability across Indian languages. The framework allows language specific tagsets to be derived from it. An important consideration for its hierarchical structure and decomposable tags is that it should allow users to specify the morphosyntactic information applicable at the desired granularity according to the specific language and task. Thus, IL-POSTS offers broad guidelines for users to define their own tagset for a particular language and/ or a specific application.

Structure of the Framework (ILPOSTS)

The framework has three layers in the hierarchy – the top level having the universal lexical categories followed by types (of these categories) in the middle layer and features (attributes) carrying finer details placed in the last layer. The description of the same is as follows -

Categories:

Categories are the primary grammatical classes to which the words belong. 'Grammatical' means grossly the parts of speech through which each individual word is recognized, e.g., noun, verb, adjective etc. The Category level tags are determined on the basis of the categorization features of the word. It decides on the Parts of Speech the word that it belongs to.

Types:

Types are the subclasses or finer specification of the categories, which are determined on the basis of either form or function. E.g., Common Noun, Proper Noun etc. are the subcategory of the category 'Noun'. Types fine tune the classification of the categories. Each Type group words which are similar in terms of their characteristics, i.e., whether they form a class on the basis of their distribution, references etc. Thus all named entities are Proper Nouns under the category of Noun.

Attributes:

Attributes are the set of basic morpho-syntactic features of a type, like, gender, number, person etc. Each Type has fixed and exhaustive set of Attributes that accumulates the possible morpho-syntactic

[1] For detailed understanding of ILPOSTS, we highly recommend reading **Framework for a Common Parts-of-Speech Tagset for Indic Languages.** Draft. Microsoft Research Lab India. Bangalore.

features to be attached to each word. Some attributes are mandatory and some are optional. Mandatory attributes are those which contributes to the basic meaning of the word in its grammatical specification, e.g., Tense in Verb is mandatory, but Distributive[2] in pronouns since the information is derived from the syntactic structure of the word/ phrase. These attributes have a fixed set. In this framework we have a set of eighteen attributes distributed over the nominal and verbal morphology.

Description of Hindi Tagset

ILPOSTS is designed in such a way so that it allows language specific tagsets to be derived from it. Any guideline therefore pertains to tagging and the description would be under this framework. It is always possible to derive a slightly different tagset for Hindi (say, a tagset ignoring some attributes or merging some types). Under such cases, the tagging guideline would a slightly different according to the tagset. But, by and large any tagset for Hindi from ILPOSTS should use very similar tagging guidelines.

The objective of the guideline is to provide complete and disambiguating guidelines for tagging Hindi sentences. This tagset consists of **categories, types**, and their **attributes** which are the three different levels of the hierarchy. **Categories** are the top-level part-of-speech classes like noun, adjective etc. and they are obligatory. **Types** are the main sub-classes of categories and may be included depending on whether or not those types exist in a particular Indian language. **Attributes** are morpho-syntactic features of Types. All attributes are optional, and are included in the language specific tagset depending on the morpho-syntactic realizations of the same.

The purpose of this guideline is to facilitate tagging Hindi sentences. Sentences are tagged using an Annotation Tool developed in Microsoft Research India. This Annotation Tool is a GUI for assignment of appropriate tags in to each word in a sentence. The tool can be adopted in any language given the language specific tagset in the proper format. This tool can be adopted for either flat or a hierarchical tagset. Annotation tool guideline[3] is complementary to the tagging guideline. The default values that are mentioned in this guideline are in accordance with the values set in the Annotation Tool. Reading the Annotation Tool Guideline is recommended before starting with the tagging process or using the tagset.

Some important points:

- While annotating the words we concentrate on the **forms** for the attributes, i.e., if the attributes are present in the word morpho-syntactically, we mark them accordingly. We focus on **function** while annotating the types. Those special cases would be dealt in Appendix A. However, this is only a general guiding principle and may vary depending on the context.

[2] Distributive is a function of Reduplication which distributes the event or the agent into parts, e.g., jo jo Aja jAyenge)

[3] Dandapat, S. MSRI Part-of-Speech Annotation Interface. Draft. Microsoft Research India Private Limited.

- The tags are saved or represented just after the words separated by back slash (\). Abbreviations of categories and their types are written without any symbol, but the values of the attributes are written being separated by DOTs (.).

- **Common value[4] for all the attributes** –
  - ***Not-applicable (0)***
    - When any value is not applicable to the category or the relevant morpho-syntactic feature is not available.
    - When the values of a particular Attribute are 'yes' and 'no' as in the case of Emphatic, Negative, Definiteness etc.; check whether the morphological attribute is present or not, tag '**y**' or '**n**' accordingly.
  - ***Undecided or doubtful (x)***
    - when the annotator is not sure about a possible attribute, instead of marking on the basis of doubt, tag it as '**x**', e.g., inherently ambiguous cases would be given priority of the contexts; but if they still remain disambiguated, annotate the attributes to be '**x**'.

Description of Tags

- **Categories and their Types: See table 1.1**
- **Attributes and their values: see Table 1.2**

1. NOUN (N)

A noun is generally inflected for gender, number, and case. The types and attributes of a NOUN are –

TYPE	ATTRIBUTES
Common (NC)	**Gender, Number, Case, (Case Marker)**
Proper noun (NP)	**(Gender), Number, Case, (Case Marker)**
Verbal (NV)	**Case, (Case marker)**
Spatio-temporal (NST)	**Case, (Case marker)**

1.1. COMMON NOUN (NC)

Common nouns in this tagset are the words that belong to the types of common noun (person, place or a thing), abstract noun (emotions, ideas etc), collective noun (group of things, animals, or persons), countable and non-countable nouns, and nouns in a complex verb etc.

[4] These common values are specifically for users of the annotation tool, read the annotation Tool guideline for better understanding of the structure of the tagset and the tool

Gender:

Annotate the gender according to the agreement with the verb. Hence the gender can be semantic (i.e. natural gender), as well as grammatical, that means चाय 'chAya' (tea) feminine, but पानी 'pAnI' (water) masculine etc. The natural gender and the grammatical gender coincide only in the case of animate nouns. Gender information might not be morphologically present in the word as, only common nouns are inflected for Gender. In these cases, annotate the words according to the grammatical gender, i.e., the way they show agreement with the verb. Usually the gender information would be present in the sentence, annotate them accordingly, and otherwise annotate them as \0\, 'not applicable'. .

Number:

The default value set in the annotation tool is – singular. Select the value of the number attribute as 'plu' (plural) only when plural markers are present. Following are some clues for number marking -

1. -e (-A ending masculine), e.g., लड़का-लड़के *la.DkA – la.Dke,*
2. Zero plural (consonant ending and vowel other than –a ending masculine nouns), e.g., भाई *bhAi,* लौहार *lauhAra etc.*
3. -o.N (all oblique masculine forms and feminine nouns other than –I ending), e.g., लड़कों *la.Dko.N,* बहनों *bahano.N,* माताओं *mAtAo.* etc.
4. –I ending feminine nouns have –iyo.N, e.g. लड़कियों *la.Dkiyo.N*
5. -o (vocative masculine and feminine nouns), e.g., लड़कों la.Dko.N, बहनों bahano.N, माताओं mAtAo.N,
6. -iyo (for –I ending feminine nouns vocative case), e.g., लड़कियों *la.Dkiyo.N* etc.
7. -iyA.N (for –I ending feminine nouns direct case), e.g., लड़कियाँ *larkiyA.N*
8. -e.N (for feminine nouns other than –I ending, direct case), e.g., माताएँ *mAtAe.N,* बहनें *bahne.N*

- Most mass nouns would not be inflected for number; in that case, annotate as '0' (i.e., not applicable);
- If the word does not show any number agreement, but the noun is inherently plural, annotate the word as a plural noun.

e.g., *कुछ\JQ.0.pl.dir.n.nnm* ***राज्य\NC.0.pl.dir.0*** *लागत\JJ.0.sg.dir* *विभाजन\NC.0.sg.dir.0* *व्यवस्था\NC.0.sg.dir.0 के\PP.0.sg.gen संबंध*

*सम्पन्न\JJ.mas.sg.dir वर्ग\NC.mas.sg.dir.gen के\PP.neu.sg.gen लोग\NC.mas.***pl***.dir.0 ही उठाया करते थे*

Case:

In this framework, we have two types of cases: **direct** and **oblique**.

- Direct cases are unmarked and **not** followed by any post positions.
- Oblique cases are either marked by inflection or followed by post positions, or both. Select the value for case looking at the morphological form of the word and the post positions.

- Direct case is unmarked nominative case, and unmarked accusative case. It appears in nominative subjects and direct objects. E.g.,

  वह लड़का पेड़ काटता था

  Vah **_la.DakA_\NC.mas.sg.dir.0** **_pe.D_\ NC.mas.sg.dir.0** *kATatA thA*

  लड़की सेब खा रही है

  La.DakI seba khA rahI hai

- Oblique case on the noun or the pronoun is marked by inflections. Inflected noun or the pronoun must be followed by post positions. Usually oblique cases are object of a post position which serves as a case marker.

E.g., राम कल एक **लड़की से** लालबाग मे मिलने जा रहा है

rAma kal ek la.DakI se lAlabAga me milane jA rahA hai

राम कल कुछ **लड़कियों से** लालबाग मे मिलने जा रहा है

rAma kala kucha laRkiyo.N se lAlabAga me milne jA rahA hai

राम कल एक अच्छे **लड़के से** लालबाग मे मिलने जा रहा है

rAma kala eka acche la.Dake se lAlabAga me milne jA rahA hai.

राम कल कुछ अच्छे **लड़कों से** लालबाग मे मिलने जा रहा है

rAma kala kuCha acche la.Dake se lAlabAga me milne jA rahA hai.

- Though vocative case does not have an oblique case marker, we consider it to be oblique as it is morphologically marked. When the word is inflected for this case; it is not followed by post positions. Usually the marker is –'o' in plural nouns, irrespective of gender and number, except for –A ending masculine singular nouns where the marker is '-e' , e.g.,

ए लड़के...

e la.Dake…

भाईयों और बहनोँ

bhAIyo.N aur bahano.N…

- But sometimes, oblique cases are not always followed by post positions. E.g.,

  वह दो दिन (तक) घर (मेँ /पर) रहा

  vah do dina (taka) ghara (me.N/ para) rahA

  In the above case, annotate the case as 'dir' if the noun is not followed by a post position. If it is followed by a post position, it would be marked as oblique case.

- Again, subjects can be oblique too, e.g.,

  मेरे /राम से यह काम नहीँ होता

  mere/ rAma se yaha kAma nahI.N hotA

  मुझे बहुत सरदर्द है

  mujhe bahut sardarda hai.

- Follow the rule to find out oblique case marking:

  -o.N (obq pl) for both masculine and feminine plural nouns, 0 for obq sg in masculine
  and feminine (except for -e in –A ending masculine nouns, larkA > larke)

  We are focusing on grammatical case; so, annotate the case just looking at the form of the noun (or pronoun).

Case Marker:

Case markers in Hindi are usually postpositions that refer to different cases, e.g., /*se*/ *'*Instrumental or Ablative*'* case. Case marker in our framework is **optional**, i.e., it is not necessary to annotate the values of case markers.

The list of case markers in the tagset comprises ergative, accusative, instrumental, dative, genitive, sociative, locative, ablative, benefactive, vocative etc. Nominative case is not listed since nominative is not marked morphologically. All the other case markers are marked by postpositions. Therefore, if the case is tagged as ‘direct’, the case marker would be tagged as ‘0’ (for both subject and direct object). If the case is tagged as ‘oblique’, the tag of the case marker would be indicated by the following postposition. This includes ergative and vocative as oblique cases. If the noun (or pronoun) is attached with the following post position, i.e., the post position is suffixed to the noun (host), case marking is **mandatory**, otherwise it would be optional. In those cases where the oblique cases are not followed by post positions or not marked by suffixation, annotate them as direct. E.g.,

रामने rAmane\NC.mas. sg.dir.erg

However, the case markers listed in the following–

Ergative, accusative, instrumental, dative, genetive, sociative, locative, ablative, benefective, vocative, purposive

And the morphological forms corresponding to the case markers are –

ne **ने**	ergative
ko **को**	accustative
se **से**	instrumental
ko **को**	dative
kA/ke/kI **का के की**	genitive
ke sAth **के साथ**	sociative
me.N/par **में पर**	locative
se **से**	ablative
ke lie **के लिए**	benefactive or purposive
Marked on the noun itself	vocative

- Some annotated words - *श्रम\NC.mas.sg.dir.0, धारा\NC.fem.sg.obl.loc में\PP.0.0.loc, बातों\NC.fem.pl.obl.gen की\PP.fem.sg.ge, भाभी\NC.fem.sg.obl.gen का\PP.mas.sg.gen चेहरा\NC.mas.sg.dir.0, शरीर\NC.mas.sg.obl.loc पर\PP.0.0.loc.*

As seen from the example given above, oblique cases are not always followed by a post position; the morphological form is the only clue to render the information.

- *के\PP.0.sg.gen साथ\PP.0.sg.soc* together acts as sociative case marker. In this case, the first of the two post position is marked as genitive case and the other one is marked as sociative case. This kind of cases forms a closed class and can be listed easily.
- In complex predicates the number attribute of the noun of a compound verb (NC/JJ +V) is annotated as '0', e.g.,

  कायम\NC.0.0.dir.0 करते\VM.0.0.0.0.0.0.ifn.n हुए\VAUX.0.sg.0.0.0.0.nfn.n,

  सालों-साल\NC वंचित\JJ.0.0.dir रही\VM हूं\VAUX

1.2. PROPER NOUNS (NP)

When the word denotes a specific name of a person, place, shop, institution, date, day, month, species, etc., or whatever is considered to be a name would be marked as proper noun. If the word is of some other category, but is used as a proper noun in a context; should be marked as proper noun.

Gender:

Gender is **optional** for proper nouns. If it is a name of a person, go by the natural gender. If it is name of a place, institution or otherwise, annotate the attribute only if it is necessary.

Number:

Number is usually not morphologically present on proper nouns, except for the cases like -

Class kI sAre kavitAyeN gAnA gAtI hE

Annotate number only if they are morphologically present.

Case:

Direct and oblique cases are marked depending upon the morphological form of the noun. Usually the default case is the direct case, oblique case is not marked in the singular nouns. Proper nouns in their plural forms are marked for the oblique case.

Case Marker:

Case Markers are **optional** for any proper noun; annotate them only if they are written together.

1.3. VERBAL NOUNS (NV)

Verbal nouns are nominal forms which are derived from verbs, they are may be participles and/or gerunds. In Hindi, verbal nouns are gerunds. These are derived forms of verbs which allow nominal inflections.

The attributes are - Case and Case markers.

Annotate the **case** as direct or oblique, depending upon the morphological form of the noun.

Case markers are **optional.** Annotate them only if they are attached to the word itself.

e.g.,

tairnA\NV.dir.0 achhA hotA hai

rAm tairne\NV.obl.gen ke lie gayA

पूरा*JJ.mas.sg.dir* ***होने****NV.obl.gen* ***की****PP.fem.sg.gen* ***उम्मीद****NC.fem.sg.dir.0* ***है****VM.0.sg.3.prs.sim.dcl.fin.n*

1.4. SPATIO-TEMPORAL NOUNS (NST)

Spatio-temporal nouns are the nouns which denote time and location. Essentially spatio-temporal nouns form a specific class of nouns which can be used as post-position and nouns of time, space, direction etc. Since, they are interchangeable with the post positions; they will be tagged as NST irrespective of their places of occurrence. Hence, in this case we are going by the forms, e.g., age, nice, upar etc.

The attributes are - **case** and **case markers.**

Annotate the case as direct or oblique, depending upon the morphological form of the noun.

Case markers are **optional**. Mark them only if they are attached with the word itself.

For example –

बीच\NST.dir.0, बाहर\NST.dir.0 etc.

The list comprises of - पीछे, पहले, अगले, पिछले/पिछली, आगे, आसपास, बाद, निकट, बाहर, ऊपर, अंदर, समक्ष, भीतर, सम्मुख, निचले, आगे etc.

- Adjectives with wAlA [5]construction which are not followed by any noun are tagged as common nouns, similar to *kapRAwAlA\NC.sg.dir.0, dudhwAlA\NC.sg.dir.0,* e.g.,

  acChAwAlA\NC.sg.dir.0 etc.

- While tagging compound nouns (N+N- compounds), e.g., mukhya pratibedaka or nouns in a compound verb construction (N+V compounds, e.g., pAlana karte huye; the number attribute of the first noun should be '0'.

  e.g.,

  saMvAda\NC.0.dir.0   *pratibedaka\NC.sg.dir.0*

  pAlana\NC.0.dir.0   *karate\VM.0.0.0.0.0.0.ifn.n*   *huye*\VAUX

  Spatial adverbs should not be confused with Spatio-temporal nouns, for example -

- आज, कल, रात, रोज, सुबह, अभी, पिछले साल\NC would always be marked as adverbs and NOT NSTs.

2. VERBS (V)

The types and attributes of verbs are –

TYPE	ATTRIBUTES
Main verb (VM)	**Gender, number, person, tense, aspect, mood, finiteness, honorificity**
Auxiliary verb (VAUX)	**Gender, number, person, tense, aspect, mood, finiteness, honorificity**

2.1. MAIN VERBS (VM)

Typically a sentence in Hindi contains more than one verb, be it main verbs or a combination of main verbs and auxiliary verbs. If the sentence contains a single main verb, the main verb carries the main event information of the proposition or the sentence. It usually inflects in agreement with the subject (object for ergative cases). If the sentence contains a combination of main and auxiliary verbs, the main verb in the verb group does not carry verbal inflections; instead, the last auxiliary verb of the verb group inflects in verbal agreement.

However, the attributes a main verb might carry are-

Gender, number, person, tense, aspect, mood, finiteness, Honorificity

[5] See APPENDIX A for detailed description

Among the attributes listed above some pertains to the nominal domain and the rest to the verbal domain. Gender, number, person and honoroficity agree with the subject, hence they are nominal and the rest are completely verbal in the sense it is independent of the subject of the sentence. In a simple predicate (i.e., containing only one main verb) all the nominal inflections are marked on the verb. But in a complex predicate, the main verb and the following auxiliaries (except for the last auxiliary verb) of the in the verb group carry the nominal inflections where only the last auxiliary verb carries verbal infection. If the verb is in its root form, (e.g., **कर** रहे हैं) it does not inflect.

In a complex predicate (e.g., MV + VAUX$_1$ + VAUX$_2$), MV will inflect for gender, number, person, and Honorificity (unless it is not in the root form, e.g., **कर** रहे हैं). VAUX$_1$ would inflect for the mentioned nominal inflections as well the aspectual or modal information if the auxiliary verb is an aspectual or a modal auxiliary (e.g., rahA or cukA). VAUX$_2$ (usually 'be' verb) would be inflected for tense. It might not carry nominal information; attribute value for those VAUXes which do not show any morpho-syntactic suffixation for gender, number, person etc. would be '**0**'. In this combination, the tense attribute of VAUX$_1$ would be valued with '0', where the same of the VAUX$_2$ would be valued as per the temporal information of the predicate. However, the diverse nature of predicate argument structure in Hindi might present some structures that might not be mentioned in the guidelines. Notification of those structures would help to improve the accuracy of the guidelines.

However, following is the list of attributes for Main verb which is applicable to the auxiliary verb too. Emphasis is given on the presence of morpho-syntactic suffixes during annotation as well as valuing of the attributes.

Gender:

Gender is inflected on the verb. It agrees with the subject of the sentence, otherwise, in case of perfective cases, it agrees with the object, e.g.,

vah ***la.DkA*** *gAnA* ***gAtA\VM.mas…*** ***hai\VAUX.0***

vah ***la.DkI*** *gAnA* ***gAtI\VM.fem*** ***hai\VAUX.0***

Number:

Number is inflected on the verb in agreement with the subject of the sentence. e.g.,

vah ***la.DkA*** *gAnA* ***gAtA\VM.ma.sg…*** ***hai\VAUX.0***

vah ***la.DkI*** *gAnA* ***gAtI\VM.fem.sg*** ***hai\VAUX.0***

सदस्यों द्वारा सुने\VM.mas.pl.3.prs.pft.dcl.fin.n

जाते\VAUX.***mas.pl***.3.prs.pft.dcl.fin.n हैं\VAUX.0.pl.3.prs.sim.dcl.fin.n

us ***la.Dke*** *ne kala gAnA gAyA thA* *ve la.Dke kala gAnA gAye.nge*

Person: as the verb agrees with the subject, the person of the subject would decide the value of the person attribute on the verb.

Tense:

Tense situates the sentence in relation to the time of utterance. The values are present, past, and future. The verb '*ho*' (to be) carries the tense and nominal agreement information (gender, number, person etc.). Future tense is expressed by a suffix (-gA/ge/gi etc.) attached to the main verb.

e.g.,

rAma chAwala khAtA\VM hai\VA.0.sg.3.prs.sim.dcl.fin.n

rAma chAwala khAtA\VM thA\VA.0.sg.3.pst.sim.dcl.fin.n

rAma chAwala khAyegA\VM.mas.sg.3.fut.sim.dcl.fin.n

पूरा होने की उम्मीद है\VM.0.sg.3.prs.sim.dcl.fin.n

The simple present tense is used to denote the habitual mood too. In both simple present and past tenses, the main verb form a participial stem (e.g., khAtA)

e.g., *पिछड़े क्षेत्रों के मार्ग खोलने को प्रोत्साहन* ***मिलता\VM****.mas.0.0.0.0.0.ifn.n* *है**\VAUX.**0.sg.3.prs.sim.dcl.fin.n*

छोटे बन्दरगाहों के साथ संयोजन कायम ***\NC*** *करते**\VM.**mas.0.0.0.0.0.ifn.n* *थे**\VAUX**.0.sg.3.pst.sim.dcl.fin.n*

कुछ राज्य लागत विभाजन व्यवस्था के संबंध में पहले ही ***सहमत\NC*** ***हो\VM****.0.0.0.0.0.0.0.ifn.n* ***गए\VAUX.****mas.sg.3.prs.sim.dcl.fin.n* ***हैं\VAUX****.0.sg.3.prs.sim.dcl.fin.y /\PU*

*परियोजनाएं\NC.0.pl.dir.0 चल**\VM**.0.0.0.0.0.0.0.ifn.n रही**\VAUX**.fem.sg.3.prs.prg.dcl.fin.n* *हैं**\VAUX.**0.pl.3.prs.sim.dcl.fin.n /\PU*

Aspect:

Aspect is the grammatical device that deals with state of the event (completed or incomplete action, iterative action etc.) expressed by the verb in a sentence. Aspect is morphologically marked on the verb. The values are simple, progressive, perfect. Simple and perfect aspects are morphologically marked on the main verb itself. A simple aspect indicates an imperfective action.

e.g.,

*ritA Ama khAti\VM.fem.sg.3.0.**simp.**dcl.fin.n   hai/VAUX.fem.sg.3.prs.pft.simp.fin.n*

*ritA Ama khAti\VM.fem.sg.3.0.simp.**hab**.fin.n   thI/VAUX.fem.sg.3.pst.simp.dcl.fin.n*

This indicates the habitual mood too.

Perfect is indicated by a vowel suffix that is attached to the main verb, e.g.,

ritA ne roTI khAyI\VM.fem.sg.3.0.pft.dcl.fin.n   hai/thi\VAUX

rAma ne Ama khAyA\VM.mas.sg.3.0.pft.dcl.fin.n   hai/thi\VAUX

The progressive aspect is indicated by an auxiliary verb 'rah' (to stay). The main verb remains in its root form and the auxiliary verb inflects for gender, number and person, while temporal information is carried out by the other auxiliary verb (i.e. rahA/rahI/rahe).

The main verb in progressive aspect bears no morphological attribute, except for the gender; hence all the relevant attributes will be tagged as '0' or not applicable.

e.g.,

परियोजनाएं चल\VM.0.0.0.0.0.0.0.ifn.n रही\VAUX.fem.sg.3.0.prg.dcl.fin.n   हैं\VAUX.0.pl.3.prs.sim.dcl.fin.n

Mood:

Mood expresses the frame of mind of the speaker. Moods that we annotate in the framework for Hindi are Indicative, Imperative, and Habitual. These are morphologically present in Hindi, e.g.,

ritA	Ama	khAti	hai	declarative
ritA,	Ama imperative	khAo*\VM.0.sg.2.prs.simp.**imp**.fin.n*		

*ritA Ama   khAti\VM.fem.sg.3.0.simp.**hab**.fin.n   thI/VAUX.fem.sg.3.pst.simp..fin.n*

Some moods in Hindi are expressed through additional auxiliary verbs. We do not annotate other moods since semantic information plays least role in the framework.

Finiteness:

Finiteness is the attribute that indicates whether the action denoted by the verb is finite or not. A finite verb is morphologically inflected for Gender, Number, Person, Tense, Aspect, and Mood. An infinitive form is the base form of the verb and is not inflected for any of the inflectional categories mentioned above, except for gender, e.g.,

a. **Infinitive: _kartA_** *huyA*: inflected for gender only
b. **Non-finite: _kara_** *chuka,* ***bol*** *rahA hai,* ***khAne*** *jA rahA hai*

When the VM or VA is non-finite or infinite, all the attributes, except the gender would be tagged as ‘0’, e.g.,

प्रतिफल\NC.0.sg.obl.0 को\PP.0.sg.acc ***देखते\VM.0.0.0.0.0.0.ifn.n हुए\VAUX.0.0.0.0.0.0.nfn.n***

परियोजनाएं\NC.0.pl.dir.0 ***चल\VM.0.0.0.0.0.0.ifn.n*** *रही\VAUX.fem.sg.3.prs.prg.dcl.fin.n*

हैं\VAUX.0.pl.3.prs.sim.dcl.fin.n

In the above case, both infinite and non-finite cases have the attributes as ‘0’. Gender is not marked here since both the verbs are in their oblique forms.

Honorificity:

Honorificity is the inflectional feature that is marked on the verb in agreement with the honorific status of the subject. If relevant morphological attribute is present in the word, annotate them accordingly.

- If there is a complex predicate (i.e., NC/JJ + V, e.g, शामिल होना, निवृत्ति ले लेना, स्थापना करना etc.) the verb following the noun or the adjective would be annotated as MAIN VERB and NOT an Auxiliary Verb.

2.2. AUXILIARY VERBS (VAUX)

Auxiliary verbs are the helping verbs which along with the main verb completes the temporal or eventual information expressed by the predicate. There should be ideally one main verb in a simple sentence but a verb group forming a complex predicate (V +V) is also found. Auxiliary verbs comprise a closed class of verbs which comes after a main verb. Usually the last auxiliary verb carries the verbal inflections. According to Kachru (2006)[6] , auxiliary verbs and light verbs are different. The main difference is the following - auxiliary verbs are of two types: tense auxiliary (honA, rahnA) and modal auxiliaries. Modal auxiliaries are – saknA, pAnA, honA, paRnA, denA, cuknA, and cAhiye. 14 light verbs are – A, jA, le, de, uTh, bETh, paR, DAl, rakh, chor, mAr, nikal, dhamak, pahuNc. We do not distinguish between auxiliary verbs and light verbs or vector verbs in the framework. If the second verb in a verb complex (i.e., compound verb [noun + verb], conjunct verb [verb + verb-main/ auxiliary]) belongs to the list of verbs (light verb or auxiliary verb mentioned above), annotate it as auxiliary verb. Usually the first verb will be tagged as a VM and others will be tagged as VAUX for a verbal complex. If the second verb does not belong to the Auxiliary list given above, consider the meaning of the verb group (i.e., whether the verb group denotes one action or multiple / sequential actions) to decide whether it is an auxiliary verb or a main verb. Auxiliary verbs comprise a closed class of verbs universally which comes after a main verb.

[6] Kachru, Y. 2006. *Hindi.* John Benjamins: Amsterdam.

e.g.,

परियोजनाएं चल\VM.0.0.0.0.0.0.ifn.n ***रही\VAUX.fem.sg.3.0.prg.dcl.fin.n हैं\VAUX.0.sg.3.prs.sim.dcl.fin.n***

All the verbs in their default agreement patterns will be marked as third person agreement pattern, if the language manifests third person agreement as default agreement pattern.

Following here are some annotated examples –

होती\VM.fem.sg.3.pst.pft.dcl.fin.n थी\VAUX.fem.sg.3.pst.sim.dcl.fin.n

मिलता\VAUX.mas.sg.3.pst.pft.dcl.fin.n था\VAUX.mas.sg.3.pst.sim.dcl.fin.n

कर\VAUX.0.0.0.0.0.0.0.nfn.0 सकें\VAUX.mas.pl.3.prs.pft.dcl.fin.n

समझ लिया\VM.mas.sg.2.0.pft.dcl.fin.n है\VAUX.0.sg.3.prs.sim.dcl.fin.n

ख़चाल\NC.mas.sg.dir.0 आया\VM.mas.sg.3.pst.pft.dcl.fin.n

घर\NC.mas.sg.dir.0 जा\VM.0.0.0.0.0.0.0.nfn.0 रहे\VAUX.mas.sg.3.0.prg.dcl.fin.n

- Hindi specific tagset in this framework does not have participle as a separate category, hence participles are marked as non-finite or infinite verbs.

3. PRONOUNS (P)

The types and attributes of pronouns are –

TYPES	ATTRIBUTES
Pronominal (PPR)	**Gender, number, Person, Case, (Case marker), (Distributive), Emphatic, Distance[7], Honoroficity**
Reflexive (PRF)	**Gender, number, Case, (Case marker), Emphatic**
Reciprocal (PRC)	**Case**
Relative (PRL)	**Number, Case, (Case Marker), (Distributive), Emphatic, Honorificity**
Wh-pronoun (PWH)	**Number, Case, (Case Marker), (Distributive), Emphatic, Honorificity**

3.1. PRONOMINALS (PPR)

Pronominal include all personal pronouns, demonstrative pronouns, inclusive pronouns and indefinite pronouns. In the case of indefinite pronouns, person attribute should be annotated as [0] i.e., not

[7] Distance is an optional feature throughout the framework. One can add it if feels necessary.

applicable. Some traditional classifications of pronouns are listed below. We do not follow this classification, but include/ assemble them all under the type 'Pronominal' (PPR).

- personal pronouns: ***mai.N, tum, tu, vah*** etc.
- Demonstrative pronouns*:* ***ye, wah, is*** etc.
- Inclusive Pronouns: ***saba, dono*** etc.
- Indefinite pronouns: ***koi, kisa, kisa ka, kuCha*** etc.
- Others: **hara, aura** etc. which are used as quantifiers if followed by a noun.

Many pronouns from the above list can be used as a demonstrative. The main difference in the distribution between a pronoun and a demonstrative is that a demonstrative is followed by a noun or adjective, but a pronoun has the distribution of a noun, and need not be followed by another noun or other parts of speech. All the **demonstrative pronouns** (listed above) are considered as pronominal in the tagset.

Gender:

Gender information is not morphologically encoded in the pronouns.

Number:

Number is morphologically marked in most of the cases in pronouns. Annotate it as plural 'pl' if number is morphologically present in the word. In case of inclusive pronouns, the number attribute should be annotated as '0', as no singular-plural distinction is found morphologically in those cases.

Person:

Person attribute would be annotated only in personal pronouns. All the demonstrative pronouns would be annotated as '3'. Annotate the person as '0' (not applicable) in inclusive, indefinite, and other pronouns.

Case*:*

Usual distinction between the direct and the oblique cases are found, e.g.,

Mai.N-mujh-merA/mere/merI; tu- tujh- terA/ terI/ tere; koi-kisi-kisikA/kinake/kinakI;

Case-Marker:

This is an optional attribute. The default value of the case-marker is '0' when it is not marked or not applicable. Annotate the Case-Markers only if the morphological attribute is present there and they are suffixed to the pronoun. If the case markers are not suffixed to the pronoun (appears as post positions), one need not value the case marker as it is an optional attribute. Case marking is mandatory only when it is written together with the word. Genitive case shows gender marked morphological forms among the case markers. Genitive case in Hindi agrees with the head of the noun phrase, i.e., the possessed and not the possessor. E.g., *uskI* ***kitAba,***

uskA ***ghara***

merA, merI, mere etc.

Distributive:

Distributive is an optional attribute for pronouns. It is the feature for the pronoun which refers to distribution of the reference denoted by the pronoun. This is a function of reduplication; therefore, distribution would be easily identifiable in presence of reduplication. E.g.,*koi koi, kuCha kuCha* etc.

Emphatic and **Honorificity** are tagged as 'y' only if they are morphologically present. Honorificity is found only in second and third person pronouns.

Distance: distance is an optional feature throughout the framework. The values for distance are – proximal, and distal. This information is conveyed by different lexical items, e.g.,

Is(proximal) ko, Us(distal) ko; yah(proximal) – vah(distal) etc.

However, Distance has been kept as an optional attribute as it does not play any significant syntactic role, rather semantic role.

The main difference between a pronoun and a demonstrative is that a demonstrative should be followed by a noun, pronoun or adjective, but a pronoun has the distribution of a noun.

e.g., *यह\PPR.0.sg.3.dir.0.n.n.0.n एक आवद्गयक घार्त होगी*

3.2. REFLEXIVE (PRF)

A reflexive pronoun is a pronoun that is preceded by the noun or pronoun to which it refers (its antecedent), e.g., the reflexive pronoun is '*Apa',* the oblique form is '*apane*'.

A reflexive pronoun is inflected for **gender** and **number** in the following instances –

apanA, apanI(gender) and apane (number)

satIsha ko ***apanA*** *ghara sabase acChA lagtA hai.*

sImA ***apani*** *bahana ko/ se bahota pyAra karti hai*

sAre la.Dke ***Apane*** *ghar cale gaye*

Case: annotate them according to their morphological forms.

Case markers: only genitive is marked with a different morphological form. E.g.,

sImA apane ko bahot khubsurat mAnati hai

usne apane se hI bulAyA

Emphatic: usually the emphatic information is conveyed by a particle hI, bhI etc. but they can be either written separately or suffixed to the main pronoun. Annotate the value if the particle is suffixed to the pronoun.

3.3. RECIPROCAL (PRC)

The reciprocal is ‘*apne Ap*’. Cases are annotated if they are marked morphologically.

3.4. RELATIVE (PRL)

A relative pronoun is a pronoun that links two clauses into a single complex clause. It is called a relative pronoun because it relates to the word that it modifies. Relative and wh pronouns bear the same attribute list and they are similar to the guidelines referred in the pronominal list. E.g.,

Singular : **Jo-jisa-jisakA/ jisake/jisakI**;

Plural : **Jo-jina-jinakA/ jinake/jinakI** etc

Number is not morphologically marked in relative pronouns; hence annotate them as ‘0’ in number.

e.g., jo, jis, jin

This includes the correlative pronouns too.

3.5. WH PRONOUN (PWH)

Number is not morphologically marked in wh-pronouns; hence annotate them as ‘0’ in number.

e.g., ***kyA- kisa- kisakA/kisakI/kisake, kona- kina-kinakA/kinake/kinakI*** *etc.*

4. NOMINAL MODIFIER (J)

Nominal modifier is the category which usually modifies nominal constructions (nouns and pronouns) in the sentence. Adjectives and quantifiers have been put into the same group of nominal modifiers on the basis of their function. Intensifiers are also put into this category though intensifiers can modify verbal domains too. Intensifiers basically form the type of ‘non-numeral quantifier’

The types and attributes of nominal modifiers are –

TYPES	ATTRIBUTES
Adjectives (JJ)	**Gender, Number, Case**
Quantifier (JQ)	**Gender, Number, Case, Emphatic, Numeral**

4.1. ADJECTIVE (JJ)

An adjective modifies a noun; hence it is kept as a type of nominal modifier in the framework. Though adjectives are not always followed by nouns, it can be used as a predicate too. The first kind is called an attributive adjective and the second type is called a Predicative adjective. An adjective can function as a noun if not followed by a modified noun; in that case it is called an absolute adjective. However, these do not make any difference in the attribute set of the tagset. Nor do the comparative and superlative adjectives. An adjective in Hindi is inflected for gender, number and case.

Gender:

Gender is marked on the adjectives only if it is an –A ending adjective. Otherwise, masculine gender is taken to be the default value of gender in adjectives. In all the other cases morphological inflection is null. E.g., *achhA laRkA, acche laRke, acchi laRki, acchi laRkiyaN etc.*

Other examples -

नई\JJ.fem.pl.dir ***लाइनों***\NC.fem.pl.obl.0;

सम्पन्न\JJ.mas.sg.dir वर्ग\NC.mas.sg.dir.gen के\PP.neu.sg.gen लोग\NC.mas.sg.dir.0 ही उठाया करते थे

Adjectives ending in consonants or vowels other than –A, remain the same throughout the paradigm. E.g.,

<u>*chatur*</u>   *laRkA/ laRke/ laRkI/ larkiyAN*

Number:

The above criteria are same for number and case too. –A ending masculine singular direct adjective have corresponding plural direct and oblique form ending in –e. these attributes are equally applicable to the predicative adjectives too.

laRkA/ laRkI/ larke/ larkiyaN acchA/acchI/accheN hE

- When adjectives are not followed by a noun, annotate them as nouns, e.g., acchoN\NC.mas.pl.obl.0 acchoN\NC.mas.pl.obl.0 se pAlA paRA

However, followings are some tagged examples of adjective:

एक\JQ.0.sg.dir.n.crd महत्त्वपूर्ण\JJ.0.sg.dir भूमिका\NC.0.sg.dir.0

कार्य\NC.mas.sg.dir.0 को\PP.0.sg.acc पूरा\JJ.mas.sg.dir करने\NV.obl.0 का\PP.mas.sg.gen प्रस्ताव

संयोजकता\NC.0.sg.dir.0 प्राप्त\JJ.0.sg.dir होती\VM.fem.0.0.0.0.0.0.ifn.n है

स्वर्णिम\JJ.0.sg.dir चतुर्भुज\NC.0.sg.dir.0 में

चयनित\JJ.0.sg.dir सेक्द्रानों\NC.0.pl.obl.0 की\PP.fem.sg.gen मल्टी-प्लेक्सिंग

संबंधित\JJ.0.sg.dir परियोजनाओं

अनेक\JQ.0.pl.dir.n.nnm महानगरीय\JJ.0.sg.dir परियोजनाएं

4.2. QUANTIFIERS (JQ)

A quantifier is a word which quantifies the noun, i.e., it expresses the noun's definite or indefinite number or amount e.g.,

दसवीं\JQ.0.sg.dir.n.ord योजना\NC.mas.sg.dir.0;

Gender is not marked in cardinals. It is marked in non-numerals and ordinals. Gender as an attribute is present in only –A ending quantifiers, e.g.,

thorA sA pAni, thoRi si sharArat, wo pahlI bAr, pahlA nashA etc.

Number is also marked in some of the numerals. Usually it is not marked. Annotate it according to its inherent numeral qualities e.g., '*ek*' would be singular, but '*anek*' would be plural.

Kuch (pl) acche larke aye the,

Kuch (pl) acchi larkiyaN ayii thii

Kuch(pl) samay (uncountable)

Annotate the **case** if it is morphological in nature. Oblique case is marked in quantifiers (as well as adjectives) only in –A ending ones, eg. Acche larke, acchi larkiyA, sundar larke,

Annotate **emphatic** as an attribute if it is suffixed to the quantifier, e.g., thoRA hi vs. thoRAhi chawal

Numerals: the values for numerals are: cardinal, ordinal and non-numeral. Any word other than cardinal and ordinals are annotated as non-numerals.

  - Ordinal: quantifiers those denote the orders
  - Cardinal: number words
  - Non-numerals: quantifiers other than numbers and ordinals, which includes existential or universal quantifiers, modifiers, intensifiers etc.

- Modifiers can modify a noun, as well as a verb. In that case, we presume that there is an ellipsis of a noun and hence the construction looks like a verbal modifier.
- When a quantifier is not followed by a noun annotate it as noun.

e.g.,

एक\JQ.0.sg.dir.n.crd महत्त्वपूर्ण\JJ.0.sg.dir भूमिका\NC.0.sg.dir.0 निभाती\VM.fem.sg.3.prs.sim.dcl.fin.n हैं\VAUX.0.sg.3.prs.sim.dcl.fin.y

विभिन्न\JQ.0.pl.dir.n.nnm राज्यों में अनेक\JQ.0.pl.dir.n.nnm महानगरीय\JJ.0.sg.dir परियोजनाएं\NC.0.pl.dir.0

कुछ\JQ.0.pl.dir.n.nnm राज्य लागत विभाजन व्यवस्था के संबंध में

सभी\JQ.0.pl.dir.n.nnm नई\JJ.fem.sg.dir परियोजनाओं के संबंध

5. DEMONSTRATIVES (D)

Demonstratives form a class of words which deictically refers to the entity or the object. It refers to a spatial deixis for the location of the referent (i.e., object or the entity). In Hindi, the forms are same in demonstrative and pronouns, but the only difference is that the demonstrative always followed by a noun or a pronoun. Types and attributes of demonstratives are the following. Notice that all types of demonstratives contain the same set of attributes.

TYPES	ATTRIBUTES
Absolute Demonstrative (DAB)	**Number, Case, Emphatic**
Relative Demonstrative (DRL)	**Number, Case, Emphatic**
Wh Demonstrative (DWH)	**Number, Case, Emphatic**

5.1. ABOSOLUTE DEMONSTRSTIVE (DAB)

Demonstratives have the same form of the pronouns, but they are different in their distribution than the pronouns. Demonstratives are always followed by a noun, adjective or another pronoun, where a pronoun is not followed by another noun, pronoun or adjective. Reduplication of pronouns, such as, ‘*jo jo*’, ‘*jisa jisa ko*’ etc are treated as pronouns, and not demonstratives.

e.g.,

is/ us *kitAb ko idhara do*

yah/ vah *ghar acchA nahi hai*

For all the attributes in demonstrative, go by the rules stated in the pronouns.

e.g., *उन\DAB.pl.dir.n कार्यों को,* ***यह*** *आयोजन होगा,* **इस** *अवसर पर,* ***यह*** *सपना देखा था,* ***इसी*** *वजह से,* ***इन*** *नहरों का प्रबंधन,* ***ये*** *आँकड़े कतई आश्चर्यजनक नहीं हैं*

Wrongly annotated:

इसका\DAB आकार, इसका\DAB प्रमुख\JJ\ कारण यह\DAB था\VAUX, अब\DAB शिरशोव समुद्र विज्ञान संस्थान, यही\DAB कारण है, यहाँ\DAB उल्लेखनीय है

5.2. RELATIVE DEMONSTRATIVE (DRL)

Relative demonstratives are non-distinguishable from relative pronouns, except for that a demonstrative is ALWAYS followed by a noun, pronoun or adjective.

Jis Adami ko ye bAta patA hai…

Wrongly annotated

जिसका\DRL.sg.dir.n स्वतंत्र अस्तित्व

5.3. WH DEMONSTRATIVES (DWH)

Wh-demonstratives are non-distinguishable from wh-pronouns, except for that a demonstrative is followed by a noun, pronoun or adjective. The change in the morphological form is found.

6. ADVERB (A)

An adverb belongs to a group of words that modifies the verb, adjective or the sentence.

Types and attributes of adverbs are -

TYPES	ATTRIBUTES
Adverbs of Manner (AMN)	**Case, (Case Marker), Emphatic**
Adverbs of Location (ALC)	**Case, (Case Marker), Emphatic**

6.1. ADVERBS of MANNER (AMN)

These adverbs are the adverbs which modifies the way the action is described in the verb, e.g.,

तकरीबन, सर्वथा, अत्यंत, हाल\AMN ही में, लगभग, अचानक, फिर etc.

*Some other adverbs like - **Aise** hI kAma karnA cahiye, **kayse, jaise, vaise** itnA* etc.

- Manner adverbials are sometimes coined with '*se*', '*pUrvak*' etc. too, e.g., jaldi se A jAo
- In the above cases, if the postposition i.e., '*se*' is attached to the preceding word, it is tagged as an adverb; otherwise, they are tagged as post positions. e.g., *jaladi\AMN.obl.0.nse*

6.2. ADVERBS of LOCATION (ALC)

All the words denoting time and place come under ALC, hence it includes words denoting words or phrases that spatially and temporally modifies the verb. E.g.,

aba, yahA.N, wahA.N, taba, udhara, kidhara, kahA.n, Aja, kal etc.

- ALCs are not similar to the NSTs. NSTs consist of a closed class of words which can be used as a post position too. However all the attributes present in adverbs will be tagged according to their morpho-syntactic feature.

e.g.,

aba\ALC.dir.0.n **yAhA.N\ALC.dir.0.n** bArISha hone wAli hai

kAhA.N\ALC.dir.0.n the itane dina?

Taba/ALC.obl.0.n **ki/PP.fem.sg.gen** bAta hai

e.g.,

जब\DRL.sg.dir.n.dst **इस**\DAB.sg.dir.n.prx **सागर**\NC.mas.sg.obl.gen **का**\PP.mas.sg.gen **आकार**

जहाँ\ALC.dir.0.n **इतिहास**\NC.mas.sg.obl.gen **का**\PP.mas.sg.gen **सहारा**, **वहाँ**\ALC.obl.abl.n **से**,

यहीं\ALC.obl.loc.n **पर**,

7. POSTPOSITION (PP)

A post position is the functional word that occurs after the word to indicate the inflectional markers. Post positions indicate grammatical relations between two parts of speech.

The attributes for postpositions are-

TYPES	ATTRIBUTES
Postposition (PP)	**Gender, Number, Case Marker**

Gender and **number** attributes are present in postposition only for the genitive markers, e.g., kA, ke, kI, and not usually for others. However, tag these attributes with their value according to if they are physically marked in the postposition itself. If it is only for the genitive markers, then annotate the rest as ‘0’, hence the default value is ‘0’ in number and gender in postpositions.

Since case markers are mostly marked by the postposition in Hindi, it is mandatory to tag the case markers in the post positions, unlike nouns.

e.g.,

उन कार्यों को\PP.0.sg.acc पूरा करने पर\PP.0.sg.loc बल दिया...

परियोजनाओं से\PP.0.sg.abl परिपक्वता वाले खण्डों\NC.0.pl.obl.0 की\PP.fem.sg.gen क्षमता\NC.0.sg.dir.0 में\PP.0.sg.loc वृद्धि...

8. PARTICLE (C)

A particle is a word that does not belong to one of the main parts of speeches, is invariable in form, indeclinable and typically has grammatical or pragmatic meaning. It does not have any attributes.

TYPES	ATTRIBUTES
Coordinating (CCD)	
Subordinating (CSB)	
Classifier(CCL)	
Interjection (CIN)	
Others (CX)	

kuCha bhI ***nA\CX*** *kahate huye vah calA gayA* *mujhe   Ama*   ***nahi\CX***   *khAnA*   *hai*

8.1. COORDINATING PARTICLES (CCD)

Coordinating particles are those particles which act as conjunctions that link constituents without syntactically subordinating one to the other. These are similar to English- *and, or* and *but*; e.g.,

aur\CCD.n, etc.

*क्षमता   में   वृद्धि   होगी   **और\CCD.n** ट्रैफिक   के   आवागमन   की   बाधाएं*

8.2. SUBORDINATING PARTICLE (CSB)

A subordinating particle is a particle that acts as conjunction that links constructions by making one of them a complement of another. E.g., par\CSB.n, parantu\CSB.n, ki\CSB.n etc.

e.g., ***पर\CSB.n*** *कार्य   को   पूरा   करने   का   प्रस्ताव...*

*जिनसे   पत्तनों/ उद्योगों   की   संयोजकता   प्राप्त   होती   है   **तथा\CSB.n** जिन*
परियोजनाओं   से   परिपक्वता

8.3. CLASSIFIER PARTICLE (CCL)

A classifier particle acts as unit nouns, e.g., *500 crore\CCL.n, ityAdi\CCL.n* etc.

8.4. INTERJECTION (CIN)

Words that express emotion are interjections, e.g., *wAh\CIN.n, shAbAsh\CIN.n, are\CIN.n* etc.

8.5. OTHERS (CX)

This tag is used for negative particles and other particles which cannot be grouped under the above mentioned types. E.g., *merA to\CX.n yahi etc.*

परिपक्वता वाले\CX.n खण्डों की

पहले ***ही\CX.n*** *सहमत हो गए हैं ।*

9. PUNCTUATION (PU)

The punctuation marks are ‘ I’ ‘,’ “” ‘;’ ‘:’ ‘?’ ‘!’ ‘’ etc. They are tagged as I \PU. They do not have any attribute

10. RESIDUAL (R)

Residuals are the words those cannot be categorized under any category-type described so far. Residuals do not have any attributes.

The types of Residual are-

TYPES	ATTRIBUTES
Foreign word (RDF)	
Symbol (RDS)	
Others (RDX)	

10.1. FOREIGN WORD (RDF)

Foreign words are those words which are written in any **foreign script** other than Hindi. Any borrowed words from another written in the same (Hindi)script should not be confused with RDF. E.g.,

16\RDF, buildings\RDF, Alexander\RDF the\RDF great\RDF, ˈaɪzək\RDF ˈæzɪˌmɑv\RDF, Исаак\RDF Озимов\RDF etc.

10.2. SYMBOL (RDS)

Symbols are characters which are not used as punctuation marks. They are not used as alphabets of the language also, e.g., *(\RDS,)\RDS, $\RDS, &\RDS, +\RDS, %\RDS, @\RDS, #\RDS* etc.

10.3. OTHERS (RDX)

This tag is given to words that are written in Hindi numerals. E.g., *१३१०\RDX, २\RDX* etc.

Conclusion

This guideline focuses on the morpho-syntactic description of Hindi for facilitating the annotator with the tool. Basic description of the tool consists of the definition and descriptions of the tags which is instanced with examples from Hindi. Morphology in HIndi is syncretic which results into lack of one-to-one correspondence of tags and morpho-syntactic categories. This eventually leads to tag ambiguity while tagging. The framework adopted here tries to focus on the morpho-syntactic features of words to derive the appropriate attribute sets for the tags. Although natural language properties are quite systematic, exceptions are also found quite frequently. Any effort to categorize and classify natural language is thus a challenge for language technology research. This tagset tries to capture nuances of the language in the tagset, though this is a mammoth task to achieve; we expect documentation of finer nuances from the annotators for perfection of the tagset description. However, it is almost an impossible job to capture all the subtleties of a natural language. This guideline is specific to Hindi and aims to give clues in annotation helping in disambiguation in tagging. It would be appreciable if any

APPENDIX A: Hindi Tag set

CATEGORY	TYPE	ABBREVIATION	ATTRIBUTE VALUE SET
Noun		N	
	Common		*NC {1, 2, 5, (6)}*
	Proper		*NP {(1), 2, 5, (6)}*
	Verbal		*NV {5, (6)}*
	Spatio-temporal		*NST {5,(6)}*
Verbs		V	
	Main		*VM {1, 2, 3, 4, 7, 8, 9, 16}*
	Auxilliary		*VA {1, 2, 3, 4, 7, 8, 9, 16}*
Pronoun		P	
	Pronominal		*PPR {1,2, 3, 5, (6),(10),12,16}*
	Reflexive		*PRF {1,2, 5,(6), 12}*
	Reciprocal		*PRC {5}*
	Relative		*PRL {2, 5, (6), (10), 12, 16}*
	Wh-		*PWH {2, 5, (6), (10), 12, 16}*
Nominal Modifier		J	
	Adjectives		*JJ {1,2,5}*
	Quantifiers		*JQ {1,2,5,12, 17}*
Demonstratives		D	
	Absolutive		*DAB {2, 5, 12}*
	Relative		*DRL {2, 5, 12}*
	Wh-		*DWH {2, 5, 12}*
Adverb		A	
	Manner		*AMN {5, (6),12}*
	Location		*ALC {5, (6),12}*
Postposition		PP	PP*{1,2,6 }*
Particles		C	
	Coordinating		*CCD*
	Subordinating		*CSB*
	Classifier		*CCL*
	Interjection		*CIN*
	Others		*CX*
Punctuation		PU	
Residual		RD	
	Foreign word		*RDF*
	Symbol		*RDS*
	Other		*RDX*

ATTRIBUTE VALUES

Attributes-*Values* and	ABBREVIATIONS
2. Number	NUM
Singular	*(sg)*
Plural	*(pl)*
3. Person	PER
First	*(1)*
Second	*(2)*
Third	*(3)*
4. Tense	TNS
Present	*(prs)*
Past	*(pst)*
Future	*(fut)*
6. Case-marker	CSM
Ergative	*(erg)*
Accusative	*(acc)*
Instrumental	*(ins)*
Dative	*(dat)*
Genetive	*(gen)*
Sociative	*(soc)*
Locative	*(loc)*
Ablative	*(abl)*
Benefective	*(bnf)*
Vocative	*(voc)*
Purposive	*(pur)*
7. Aspect	ASP
Simple	*(smp)*
Progressive	*(prg)*
Perfect	*(prf)*
8. Mood	MOOD
Declarative	*(dcl)*
Imperative	*(imp)*
Habitual	*(hab)*
9. Finiteness	FIN
Finite	*(fin)*
Non-finite	*(nfn)*
Infinite	*(ifn)*
10. Distributive	DSTB
Yes	*y*
No	*n*

11. Emphatic	(EMPH)
Yes	*y*
No	*n*
12. Negative	(NEG)
Yes	*y*
No	*n*
13. Honorificity	HON
Yes	*y*
No	*n*
14. Numeral	NML
Ordinal	*(ord)*
Cardinal	*(crd)*
Non-numeral	*(nnm)*

APPENDIX B: SPECIAL CASES

wAlA CONSTRUCTION

wAlA meaning ‘one who’ is very productive in Hindi and can be combined with four major word types: viz. nouns, verbs, adjectives and adverbs. It should be noted that with each of these types they can have different functions depending on the context.

Interestingly, it can either be suffixed to the word or written separately immediately after the word without any change in the usage. They can also be inflected for gender and number depending on the word type it is attached to. We have made comprehensive analysis of different occurrences of wAlA and propose that it be handled as given below.

- With noun/ adjective/ adverb
  - When it is suffixed to the word, it should be treated as a single unit and assigned respective category
    1. With nouns: dudhawAlA, cAyawAlA,jhaNDe.NwAlAetc.
    2. With adjectives: acChAwAlA, harAwAlA etc
       - Note that if the wAlA construction is followed by a noun, it is considered as a nun; if not followed by a noun, it is considered to be a noun
    3. With verbal forms (gerunds): karanewAla, dekhnewAla

Earlier we decided:

- *when suffixed, the word should be annotated with a single tag as Participle-general*
- *when written separately after verb, they should be annotated as Verb-infinitive and Particle respectively*

Since we don’t have participles in Hindi and only Verb main and aux, the Participle-general should be changed to Verb-main-non-finite and the second can stay as it is (verb-main-infi + particle).

But, again we changed the assumption, due to-

- ✓ Firstly we are tagging karnA/ karne as verbal nouns (direct and oblique forms respectively). If that is the case then annotating them as verb-infinite would lead to tag ambiguity.
- ✓ wAlA constructions are similar to nominal categories (and not verbal)in their distribution.
  So-
- If we annotate them as noun-verbal; the problem with verb+wAlA is solved.
- karane wAlA would be tagged as VN+wAlA; wAlA deriving its tag depending on whether it is suffixed or not. But is it is written together, then the tag of the whole word would be ***JJ***, e.g.,
  AnewAlA\ JJ.mas.sg.dir *pal* *jAnewAla\ JJ.mas.sg.dir* *hai*

Ane\NV.obl.0 wAlA\CX.n pal jAne\NV.obl.0 wAla\CX.n hai

- When wAlA is written separately after the word they are considered as two separate words. The first word will be tagged with appropriate category, while wAlA will be tagged as Particle.

PROPER NOUN Vs. COMMON NOUNS

This is an important and problematic aspect of annotation as well as the guideline. We consider the nouns as PROPER NOUN if it denotes some name. The words which precede the name of a person, consider *shrI, jI, sAhAb* etc. are part of proper nouns and would be tagged as proper nouns. Consider,

beTI ***bacAMo*****\NP.0.0.0.0** *Andolan*

but in

beTI *bacAMo*\VM.0.sg.2.prs.sim.imp.n *kA* *nARA.*

APPENDIX B: EXAMPLES FROM HINDI

1. अगर\CSB स्थिति\NC.fem.sg.obl.loc में\PP.0.0.loc सुधार\NC.mas.sg.dir.0 न\CX आया\VM.mas.sg.3.pst.pft.dcl.fin.n तो\CX पाकिस्तान\NP.mas.sg.dir.0 दिवालिया\JJ.mas.sg.0 घोषित\JJ.0.0.dir किया\VM.mas.sg.3.0.pft.dcl.fin.n जा\VAUX.0.0.0.0.0.0.0.nfn.n.0 सकता\VAUX.mas.sg.3.0.sim.hab.nfn.n है\VAUX.0.sg.3.prs.sim.dcl.fin.n ।\PU

2. पर\CSB कार्य\NC.mas.sg.dir.0 को\PP.0.sg.acc पूरा\JJ.mas.sg.dir करने\NV.obl.0 का\PP.mas.sg.gen प्रस्ताव\NC.mas.sg.dir.0 है\VM.0.sg.3.prs.sim.dcl.fin.n ।\PU

3. २\RDX उन\DAB.pl.dir.n उद्योगों\NC.0.pl.obl.0 को\PP.0.sg.acc प्राप्त\JJ.0.sg.dir करने\NV.obl.0 पर\PP.0.sg.loc बल\NC.mas.sg.dir.0 दिया\VM.mas.0.0.0.0.0.0.nfn.n जाएगा\VM.mas.sg.3.fut.sim.0.fin.n

4. जिनसे\PRL.pl.obl.abl.n.n.n पत्तनों\NC.0.du.obl.0 की\PP.mas.sg.gen संयोजकता\NC.0.sg.dir.0 प्राप्त\JJ.0.sg.dir होती\VM.fem.0.0.0.0.0.0.ifn.n है\VAUX.0.sg.3.prs.sim.dcl.fin.n तथा\CSB जिन\DRL.pl.obl.n.0 परियोजनाओं\NC.0.pl.obl.0 से\PP.0.sg.abl परिपक्वता\NC.0.sg.dir.0 वाले\CX खण्डों\NC.0.pl.obl.0 की\PP.fem.sg.gen क्षमता\NC.0.sg.dir.0 में\PP.0.sg.loc वृद्धि\NC.0.sg.dir.0 होगी\VM.fem.sg.3.fut.sim.dcl.fin.n और\CCD ट्रैफिक\NC.0.sg.dir.0 के\PP.fem.sg.gen आवागमन\NC.0.sg.dir.0 की\PP.fem.sg.gen बाधाएं\NC.0.pl.dir.0 दूर\NC.fem.sg.dir.0 होंगी\VM.fem.sg.3.fut.sim.dcl.fin.n ।\PU

5. ३\RDX क्षमता\NC.0.sg.dir.0 को\PP.mas.sg.acc ,\PU विद्राेष\JJ.0.sg.dir रूप\NC.0.sg.dir.0 से\PP.0.sg.abl स्वर्णिम\JJ.0.sg.dir चतुर्भुज\NC.0.sg.dir.0 में\PP.0.sg.loc क्षमता\NC.0.sg.dir.0 को\PP.0.sg.acc बढ़ाने\NV.obl.0 के\PP.0.sg.0 उद्देद्रय\NC.0.sg.dir.0 से\PP.0.sg.0 चयनित\JJ.0.sg.dir सेक्द्रानों\NC.0.pl.obl.0 की\PP.fem.sg.gen मल्टी-प्लेक्सिंग\NC.0.sg.dir.0 से\PP.0.sg.ins

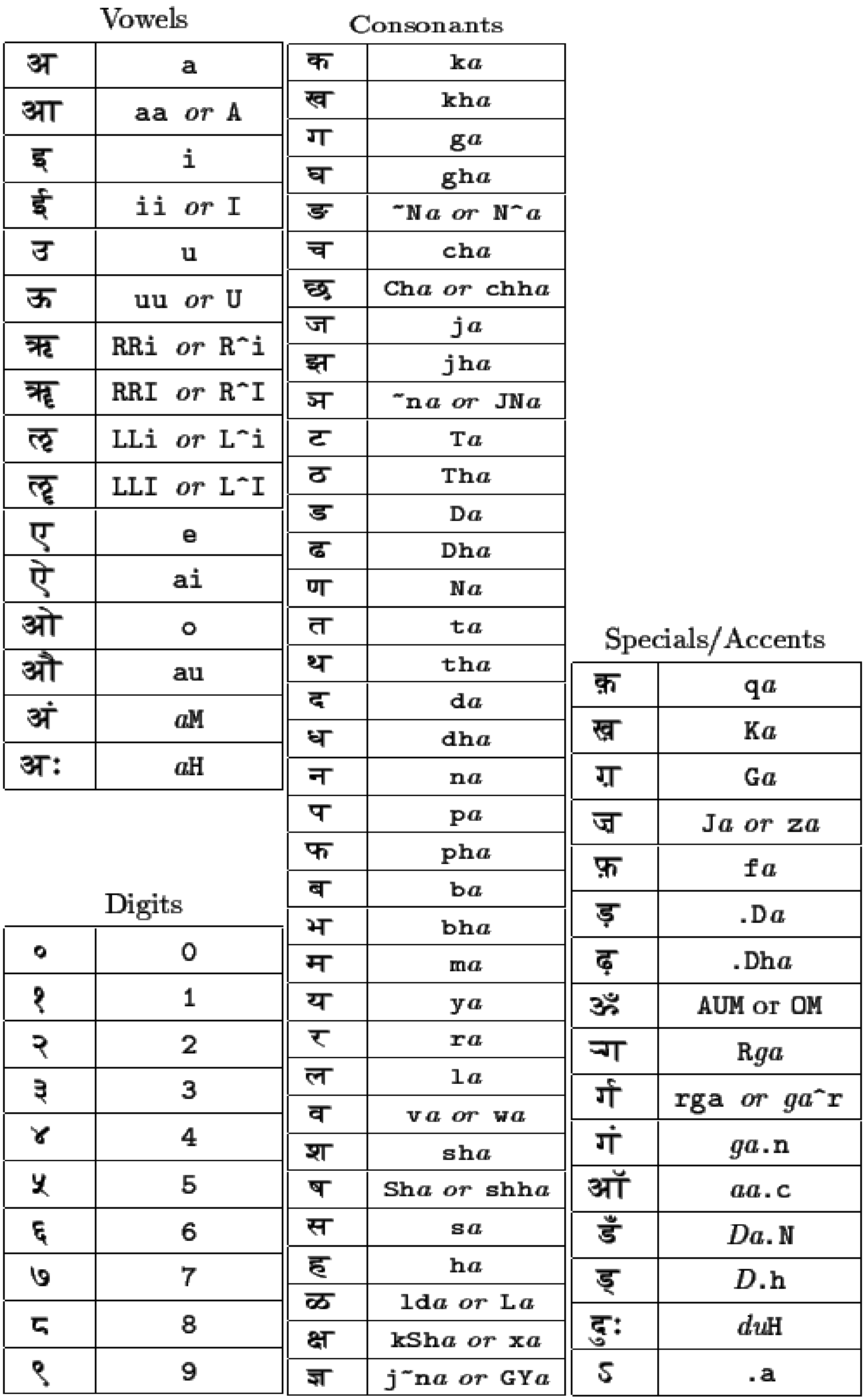
संबंधित\JJ.0.sg.dir परियोजनाओं\NC.0.pl.obl.0 को\PP.0.sg.acc
घा◌ुरू\JJ.0.sg.dir करना\NV.dir.0 आवद्गयक\JJ.0.sg.dir
होगा\VM.mas.sg.3.fut.sim.dcl.fin.n ।\PU

6. विभिन्न\JQ.0.pl.dir.n.nnm राज्यों\NC.0.pl.obl.0 में\PP.0.sg.loc
अनेक\JQ.0.pl.dir.n.nnm महानगरीय\JJ.0.sg.dir परियोजनाएं\NC.0.pl.dir.0
चल\VM.0.0.0.0.0.0.0.ifn.n रही\VAUX.fem.sg.3.0.prg.dcl.nfn.n
हैं\VAUX.0.sg.3.prs.sim.dcl.fin.n ।\PU

7. कुछ\JQ.0.pl.dir.n.nnm राज्य\NC.0.pl.dir.0 लागत\JJ.0.sg.dir
विभाजन\NC.0.sg.dir.0 व्यवस्था\NC.0.sg.dir.0 के\PP.0.sg.gen
संबंध\NC.0.sg.dir.0 में\PP.0.sg.loc पहले\NST.dir.0 ही\CX
सहमत\NC.0.sg.dir.0 हो\VM.0.0.0.0.0.0.0.ifn.n
गए\VAUX.mas.sg.3.0.prf.dcl.fin.n हैं\VAUX.0.sg.3.prs.sim.dcl.fin.y
।\PU

8. सड़कें\NC.0.pl.obl.0 ,\PU हवाई\NC.0.0.dir.0 अड्डों\NC.0.pl.obl.0 ,\PU
रेलवे\NC.0.0.dir.0 स्टेद्रानों\NC.0.pl.obl.0 और\CCD.n
बन्दरगाहों\NC.0.pl.obl.0 के\PP.0.sg.gen साथ\PP.0.sg.soc
संयोजन\NC.0.sg.dir.0 कायम\NC.0.0.dir.0 करते\VM.0.0.0.0.0.0.0.ifn.n
हुए\VAUX.0.0.0.0.0.0.0.nfn.n ,\PU अन्तर-माडल\NC.0.sg.dir.0
परिवहन\NC.0.sg.dir.0 विकास\NC.0.sg.obl.0 में\PP.0.sg.loc भी\CX
एक\JQ.0.sg.dir.n.crd महत्त्वपूर्ण\JJ.0.sg.dir भूमिका\NC.0.sg.dir.0
निभाती\VM.fem.sg.3.0.sim.dcl.nfn.n हैं\VAUX.0.sg.3.prs.sim.dcl.fin.y
।\PU

APPENDIX C: ITrans List



Vowels

अ	a
आ	aa or A
इ	i
ई	ii or I
उ	u
ऊ	uu or U
ऋ	RRi or R^i
ॠ	RRI or R^I
ऌ	LLi or L^i
ॡ	LLI or L^I
ए	e
ऐ	ai
ओ	o
औ	au
अं	aM
अः	aH

Consonants

क	ka
ख	kha
ग	ga
घ	gha
ङ	~Na or N^a
च	cha
छ	Cha or chha
ज	ja
झ	jha
ञ	~na or JNa
ट	Ta
ठ	Tha
ड	Da
ढ	Dha
ण	Na
त	ta
थ	tha
द	da
ध	dha
न	na
प	pa
फ	pha
ब	ba
भ	bha
म	ma
य	ya
र	ra
ल	la
व	va or wa
श	sha
ष	Sha or shha
स	sa
ह	ha
ळ	lda or La
क्ष	kSha or xa
ज्ञ	j~na or GYa

Specials/Accents

क़	qa
ख़	Ka
ग़	Ga
ज़	Ja or za
फ़	fa
ड़	.Da
ढ़	.Dha
ॐ	AUM or OM
ऱ्ग	Rga
र्ग	rga or ga^r
गं	ga.n
ऑ	aa.c
डँ	Da.N
ड्	D.h
दुः	duH
ऽ	.a

Digits

०	0
१	1
२	2
३	3
४	4
५	5
६	6
७	7
८	8
९	9

References

1. Dandapat, S. 2008. *MSRI Part-of-Speech Annotation Interface*. Draft. Microsoft Research India Private Limited: Bangalore.
2. *Framework for a Common Parts-of-Speech Tagset for Indian Languages*. Draft. Microsoft Research India Private Limited: Bangalore.
3. Kachru, Y. 2006. *Hindi.* Amsterdam/Philadelphia: John Benjamins.

Contact:

Priyanka Biswas

t-pribis@microsoft.com,

biswas.priyanka@gmail.com