The Nelralec Tagset (NT-01) and Guidelines for Manual Tagging

by Andrew Hardie, Ram Lohani, Bhim Regmi and Yogendra Yadava

Table of Contents

General Notes and Guidelines

Open and closed categories

- Some of the categories in this tagset are open, and some are closed. In terms of the tagset, this simply means whether or not an exhaustive listing of the word-forms that can take that tag is possible.

- A *closed category* is one where it is possible to list the contents exhaustively; an *open category* is one where it is not. This division of the categories is specified is so that, in automatic tagging, if a tagger encounters a word not listed in its lexicon, the program can safely assume that this word-form cannot possibly be a member of a closed category.

- The tags RR and UU, and any tags beginning with N, J, F (except FZ), or V, indicate **open categories**. The tags FZ, RD, RJ and RK, and any tags beginning with C, I, P, D, M, or Y, indicate **closed categories**.

- Examples given in *italics* represent an exhaustive listing of a closed category. Not all closed categories are listed exhaustively in this document.

Tokenisation and clitics

- A text is composed of a string of words. Individual instances of words are referred to as *tokens*. So, the sentence before this contains 9 tokens.

- When a text is tagged (manually or automatically) it is divided into tokens. This is called *tokenisation*. Usually, each token will receive a single tag to indicate the grammatical category of that tag.

- Sometimes, a single graphical word will contain multiple elements which we would like to analyse separately. To do this, we must tokenise them separately, i.e. break the graphical word apart into two tokens, each of which can receive a tag.

- In part-of-speech tagging, the form which is removed from the start or end of another word and made into a separate token of its own is sometimes called a *clitic*. This is parallel to, but conceptually distinct from, the use of the term *clitic* in morphology and phonology. In morphology and phonology, a form can be a clitic even if it is written as a separate word; but for the purposes of tagging a clitic *must* be written as part of the same word. This is because tagging is performed on the written form of language and has no access to the spoken form; thus, when assessing the dependence of one form on another, we have to analyse that dependence in terms of whether or not they are written with a space or not – we cannot use pronunciation as a guide as is possible in non-automated linguistic analysis.

- For the purposes of manual tagging, clitics may have been split off already, or they may not have. In the latter case the manual tagger must split them off.

- The following convention is used in this document when exemplifying tokens that result from the splitting-up of another token: when an example begins or ends in the symbol #, this indicates that the form being exemplified is a form that has been separated – i.e. it is either a clitic that has been separated, or a form from which a clitic has been separated.

The issue of dialect

- This tagset has been compiled with reference to widely published grammars of Nepali. These grammars in the main describe standard Nepali. The tagset may not be sufficient for other dialects of Nepali.

- When an example in the tagset definition is enclosed in [square brackets], that example has been reported to be limited to particular dialects.

Compound words

- One of the primary bases on which words are assigned particular tags is the inflectional patterns they follow.

- Inflection in Nepali is based primarily on suffixes rather than prefixes.

- Therefore, compounds (whether written as a single word or with a hyphen) will be tagged according to the nature of the *last* element of the compound. For instance, छोरा-छोरी would be given the same tag as छोरी.

Words with derivational suffixes

- Adding a derivational suffix to a word sometimes changes its class. For instance, *raamro* is JJM and *raaamrarii* is RR.

- However, there are also derivational suffixes which do not change the class of the word they attach to. For example, the suffix *jii*[1] which may attach to a proper noun as an honorific marker does not change the class of the word: the word is still a proper noun and would receive an NP tag.

Reduplication in Nepali

- Reduplicated words are a special case of compound words.

- When words are reduplicated, there are three ways that this can be represented in the written form.

[1] There are other such affixes: *saaheb, jyuu*, etc.

- The reduplicated forms might be written as separate words (i.e. with a space between the first instance of the form and the second instance); or they might be given as a hyphenated form (i.e. with a hyphen or dash between the first instance and the second instance); or they might be written as a single word (i.e. with no space or hyphen between the first instance and the second instance).

- If the reduplicated form is written as two separate words (*space between*), then it constitutes *two tokens* and each part should be given its own tag. In most instances this will be the same tag on both the first form and the second form.

- If the reduplicated form is written as a single word or as a hyphenated word, then it constitutes *one token* and should thus receive one tag. What tag it receives will be determined, as for any other token, by an analysis of the token's word-class, morphology and syntactic status. This will often be the same tag as the corresponding non-reduplicated word would have received.

Gender, number and case inflection patterns in Nepali

This section explains the model of gender-number-case (GNC) inflection that this tagset is based on. It should be noted that the use of this model does not imply that the tagset authors are endorsing this as a full and accurate model of Nepali GNC inflection. Rather, this model is an abstraction designed to form a basis for manual and automatic classification of word-tokens.

If some important inflectional distinctions made in Nepali are omitted in the analysis this tagset embodies, this should not be taken to imply that we consider these distinctions not to exist or to be unimportant. POS tagging frequently represents no more than the first stage of analysis; the omission of a distinction from the tagset merely implies that we consider that that distinction would be better dealt with at a subsequent stage of analysis.

- While in many languages case, gender and number are best analysed together, in Nepali case and number are indicated by elements similar to (and tagged similarly to) postpositions, whereas gender is marked by an inflected affix. Thus, distinct considerations apply to geder and to case-number, as discussed below.

- In Nepali, there are a large number of forms that are considered to be *either* postpositions *or* case-suffixes. There is no 100% agreement on what is or is not a suffix.

- Some Nepali grammatical traditions consider the following forms to be suffixes: *haruu*, *ko/kii/kaa*, *le*, *laaii* (whose meanings are, respectively, plural/collective; genitive; ergative/instrumental; and accusative/dative), and consider all other postposition/suffix forms to be postpositions.

- However, in terms of part-of-speech tagging it is difficult to draw a consistent distinction between *haruu*, *ko/kii/kaa*, *le*, *laaii* and other postpositions such as *maa*, *baaTa, sanga, dekhi*, etc.

- For this reason, the tagging rule for *all* these forms is that they are tagged separately to the noun, adjective, pronoun or other form to which they are attached.

- This means that postpositions are necessarily tokenised separately to the forms that precede them.

- Some Nepali categories mark gender by means of a three-way *o/ii/aa* distinction, where *o* is masculine, *ii* is feminine, and *aa* is "other" (it can indicate masculine plural, or feminine plural, or oblique case marking motivated by a following postposition, or honorific marking). This ending is always bound much more closely to the root than the case postpositions described above. There is also one postposition, *ko/kii/kaa*, which is marked for gender in this way.

- However, this gender marking is not found on all members of the gendered categories. Many words are "unmarked" – that is, they have a single invariant form regardless of gender. For example, adjectives are a gendered category, speaking generally: they are

marked for gender agreement with the noun they modify. But many, many adjectives are actually *unmarked* – that is, there is nothing in their morphology to indicate their gender.

- Nepali nouns have natural gender rather than grammatical gender. Animate females are feminine, all other nouns are masculine.

- In most cases gender is not marked on nouns in the way that it is marked on some adjectives. For a minority of nouns, there are pairs of masculine and feminine nouns related through the *–o/ii* distinction, for instance *keTo* "boy" / *keTii* "girl". But there are also numerous feminine nouns that end in *ii* without there being a masculine equivalent in *o* (e.g. *aaimaaii*, "woman").

- Thus the *o/ii/aa* distinction is ignored on nouns for the purpose of this system of POS tagging, but *not* on pronouns, adjectives, non-finite verbs, etc., where the distinction is motivated by agreement and is thus clearly inflectional rather than lexical-derivational, as is the case for nouns.

- In short: gender on nouns is a lexical-derivational feature, and is thus ignored. Gender on adjectives is an agreement feature and is thus not ignored.

Nouns

- There is one tag for proper nouns, another for common nouns. While number, case and gender are all relevant grammatical categories for Nepali nouns, they are not marked on the tags for reasons discussed in detail in the preceding section, and summarised below:
    - Case and number endings are tokenised separately to the noun and treated as types of postpositions
    - Gender is not included in the tags for nouns because it is not an inflectional category in the same way that it is for adjectives.
        - The grammatical genders of Nepali are masculine and feminine. The assignment of gender to nouns is *natural*: nouns denoting female humans are feminine; nouns denoting male humans are masculine; the default for all other nouns is masculine.
        - While some nouns carry gender marking, this is lexical/derivational rather than inflectional. For this reason, the tagset does not indicate the difference between masculine and feminine nouns. Words like adjectives and verbs which agree with nouns may show the gender.

- This means that *keTii* and other feminine common nouns would be tagged NN, and *siitaa* and other feminine proper nouns would be tagged NP.

- For nouns that have separate "nominative" and "oblique" forms (e.g. *keTo*, *keTaa*), these forms are tagged alike; for the purposes of this tagset they are treated as variant realisations of the noun base; contrast adjectives, for which there are different categories (JM/JF/JO).

- Some nouns have an additional form, sometimes referred to as the *illative* case (e.g. base form *ghar*, illative *ghara*). However, since the final *–a* is unwritten (or, to be more precise, the absence of the final *–a* in *ghar* is unwritten), this "illative" is never apparent in the written language; hence it will not be distinguished here.

- Where a proper noun consists of two words, those words will sometimes be compounded together. If this happens, the word receives only a single tag (proper noun tag for the case, number and gender of the word as a whole).

- Similarly, if a proper noun (or, for that matter, a common noun) is compounded together with an honorific marker such as *saaheb*, *sar*, *baabu*, *jyuu*, *jii*, *sarakara* etc., the whole thing should receive the appropriate noun tag.

Category definition	**Examples (Latin)**	**Examples (Devanagari)**	**Tag (Latin)**	**Tag (Devanagari)**
Common noun	keTo, keTaa, kalam	केटो, केटा कलम	NN	
Proper noun	raam	राम	NP	

Adjectives

- There are five adjective categories, one for unmarked adjectives, three for the three different inflected forms that marked adjectives may take, and one for Sanskrit-derived comparative and superlative adjectives.

- There are some adjectives (e.g. *bahulaahaa*) which invariably use the *–aa* ending to agree with all masculine nouns. That is, *bahulaahaa* is always used for the masculine singular, and the word **bahulaaho* does not exist (*bahulaahii* is feminine). In this case, the JO tag is used for *bahulaahaa* regardless of what type of noun it modifies.

- There are some adjectives which end in *–ii* or *–aa* but are invariable (i.e. the ending doesn't change). Examples are *saphaa* and *dhanii*. These are given the JX tag.

Category definition	**Examples (Latin)**	**Examples (Devanagari)**	**Tag (Latin)**	**Tag (Devanagari)**
Masculine adjective	moTo, raamro	मोटो, राम्रो	JM	
Feminine adjective	moTii, raamrii	मोटी, राम्री	JF	
Other-agreement adjective	moTaa, raamraa	मोटा, राम्रा	JO	
Unmarked adjective	saphaa, dhanii, asal	सफा, धनी, असल	JX	
Sanskrit-derived comparative or superlative adjective[2]	uccatar, uccatam	उच्चतर, उच्चतम	JT	

[2] Note these are, like JX adjectives, unmarked.

Pronouns and determiners

First and second person and honorific pronouns

- The tags for these words consist of three characters. The first is P (for personal pronoun); the second is either M for first person or T for second person[3] or X for unspecified person. The third indicates level of honorific: N for non-honorific, M for medial-honorific, H for high-honorific, R for royal-honorific, or X if level of honorific is not marked or not relevant (this is the case for the first person pronouns).

- There is an argument that the first person pronoun *haamii* is plural, and should therefore be tagged differently to *ma*. While acknowledging this argument, we note that this is a lexical plural rather than a inflectional plural, and the usual way of indicating number in Nepali is through the *haruu* sffix, which is handed separately; thus, *haamii* is in this system tagged as PMX.

- Second person pronouns have three levels of honorific: non-honorific, medial-honorific, and high-honorific.

- In this classification scheme, no pronouns are classed as third person. Instead, the forms often described as third person pronouns are classed either as *determiner-pronouns* (see below), or as *additional honorific pronouns*.

- The additional honorific pronouns ("additional" because they are additional to the honorific second person pronouns) are said to have *unspecified person*. This is because they can be used in both the third person and the second person (in the case of *mausuph*, its use as a second person pronoun is reported to be frowned on, but *is* known to occur). These tags begin PX because these pronouns have no specific person. These pronouns have two levels of honorific: high-honorific and royal-honorific.

- Some sources (e.g. Acharya 1991: 106-107, Schmidt et al. 1993: xiv-xv, 537, 586) report that *yahaa~* should be considered a second person pronoun and *wahaa~* a third person pronoun. However, authorities differ on this; for example Hutt and Subedi (2003: 31) suggest that both are to be considered third person pronouns (albeit one proximal and one distal in meaning).

- For this reason, the neutral course of action has been taken in this document of leaving the person of *yahaa~* and *wahaa~* and the associated inflected forms as *unspecified*.

- Moreover Schmidt et al. (1993: 537, 586) report that *yahaa~* and *wahaa~* can also be used as adverbs. Other reports have suggested that this is rare or does not occur at all. However, allowances must be made for this usage, in case it is encountered.

[3] In the context of the personal pronouns, M refers to "m-" (the first-person root form) and T to "t-" (the second-person root form).

- So the rule for tagging *yahaa~* and *wahaa~* is as follows:
    - If the word means "here" and "there", rather than referring, then it should receive the tag RD (see below), like *tyahaa~*.
    - If the word means "you" or "he" or "she" (i.e. if refers pronominally to an honoured person), then it should receive the tag PXH.

First person pronouns

- Note that there are alternate forms for *ma* when followed by certain postpositions. The form followed by the ergative postposition is *maile*, not **male*, for example.

- The first person pronouns are not followed by the genitive postposition. Instead, there are inflected forms in *ro/rii/raa*. Since *ro* does not display any of the mobility associated with *ko/kii/kaa* it has been treated as a single unit (a possessive pronoun).

Category definition	**Examples (Latin)**	**Examples (Devanagari)**	**Tag (Latin)**	**Tag (Devanagari)**
First person pronoun	*ma, haamii, mai#*	म, हामी, मै#	PMX	
First person possessive pronoun with masculine agreement	*mero, haamro*	मेरो, हाम्रो	PMXKM	
First person possessive pronoun with feminine agreement	*merii, haamrii*	मेरी, हाम्री	PMXKF	
First person possessive pronoun with other agreement	*meraa, haamraa*	मेरा, हाम्रा	PMXKO	

Second person pronouns

- As with *ma*, *ta~* has a variant form that occurs before *le*. There are also possessive pronouns for *ta~* and *timii*.

Category definition	**Examples (Latin)**	**Examples (Devanagari)**	**Tag (Latin)**	**Tag (Devanagari)**
Non-honorific second person pronoun	*ta~, tai#*	तँ, तैँ#	PTN	
Non-honorific second person possessive pronoun with masculine agreement	*tero*	तेरो	PTNKM	
Non-honorific second person possessive pronoun with feminine agreement	*terii*	तेरी	PTNKF	
Non-honorific second person possessive pronoun with other	*teraa*	तेरा	PTNKO	

agreement				
Medial-honorific second person pronoun	*timii*	तिमी	PTM	
Medial-honorific second person possessive pronoun with masculine agreement	*timro*	तिम्रो	PTMKM	
Medial-honorific second person possessive pronoun with feminine agreement	*timrii*	तिम्री	PTMKF	
Medial-honorific second person possessive pronoun with other agreement	*timraa*	तिम्रा	PTMKO	
High-honorific second person pronoun	*tapaai~, hajur*	तपाईँ, हजुर	PTH	

Additional honorific pronouns

Category definition	**Examples (Latin)**	**Examples (Devanagari)**	**Tag (Latin)**	**Tag (Devanagari)**
High-honorific unspecified-person pronoun	*yahaa~, wahaa~*[4]	यहाँ, वहाँ	PXH	
Royal-honorific unspecified-person pronoun	*sarkaar*[5]*, mausuph*	सरकार, मौसुफ	PXR	

Reflexive pronoun

- The reflexive pronoun *aaphuu* occurs in its own category. The tags for this category are like the tags above, except that the tag contains "RF" ("reflexive") instead of letters indicating person and honorific level.

- There are no special tags for the phrasally-constructed reciprocal pronoun form *ek arkaa*; this is a syntactic category rather than a morphosyntactic one.

- Like ma and ta~ it has a corresponding possessive pronoun: *aaphno*_PRFKM, *aaphnii*_PRFKF, *aaphnaa*_PRFKO.

[4] There is an alternative form for *wahaa~*: *uhaa~*, उहाँ. This form would also take PXH.

[5] When not being used to refer to a person in this pronominal fashion, *sarkaar* is a noun (meaning "government").

Category definition	Examples (Latin)	Examples (Devanagari)	Tag (Latin)	Tag (Devanagari)
Reflexive pronoun	*aaphuu*	आफू	PRF	
Possessive reflexive pronoun with masculine agreement	*aaphno*	आफ्नो	PRFKM	
Possessive reflexive pronoun with feminine agreement	*aaphnii*	आफ्नी	PRFKF	
Possessive reflexive pronoun with other agreement	*aaphnaa*	आफ्ना	PRFKO	

Pronoun-determiners: demonstrative, relative, and interrogative

- The pronoun-determiners have two uses: as part of a noun phrase (the determiner function[6]), or as a noun phrase in their own right (the pronoun function).

- Morphologically, they fall into five groups: a series beginning in t-, one in y-, one in u- or w-, one in k- and one in j-. However, the first three series are functionally rather similar: all are deictic demonstratives. The K-series are interrogative pronoun-determiners and the J-series are relative pronoun-determiners. There are also adverbs that fall into these 5 groups (see below).

- These three groups are indicated by single letters as follows:

  D Demonstrative (*yo, tyo, u; yas, tyas, us; yini, tini, uni*; etc.)
  K Interrogative (*ko, kas*, etc.)
  J Relative (*jo*, *jas*, etc.)

- Most of these pronoun-determiners are not marked for gender. However, some are: *yasto, tyasto, usto, kasto, jasto*; *yatro, tyatro, utro, katro, jatro*. This means that there are four basic groups of tags for these pronoun-determiners: D?X (unmarked), D?M (masculine), D?F (feminine),D?X (other) where ? = D, K or J. These endings are similar to those on adjectives:

yasto keTo *this-kind-of boy ("such a boy")* DDM NN	raamro keTo *good boy* JM NN
yastaa keTaaharuu *this-kind-of boys ("such boys")* DDO NN IH	raamraa keTaaharuu *good boys* JO NN IH

[6] This might also be described as an adjectival function.

- It is expected that the difference between *yasto* (etc.) as an anaphoric pronoun, taking case markers, and *yasto* (etc.) as a determiner, taking no case markers, would be something that would be desirous to indicate at a subsequent level of analysis, namely in a syntactic parsing of the text.

Demonstrative pronoun-determiners

- These tags begin DD for demonstrative determiner, and then X, M, F, or O depending on their gender marking.

- In the table below, only y-forms are listed as examples, although the corresponding t-forms and w/u-forms would take the same tags e.g. *u* and *tyo*, etc., would take the same tags as *yo*, etc. The list of forms given here is not exhaustive, although this is a closed category.

- While *yo* is clearly unmarked in the sense we are using the term here (as it has no form for the feminine) it does have a variant form, *yas*, which occurs before postpositions. As with nouns, this variant form receives the same tag as the usual form. So DDX would be used to tag *yas* in *yasle* or *yasmaa*: e.g. *yas*_DDX *maa*_II. A similar consideration applies to *yi* and *yin*.

Category definition	**Examples (Latin)**	**Examples (Devanagari)**	**Tag (Latin)**	**Tag (Devanagari)**
Masculine demonstrative determiner	yasto, yatro	यस्तो, यत्रो,	DDM	
Feminine demonstrative determiner	yastii, yatrii	यस्ती, यत्री	DDF	
Other-agreement demonstrative determiner	yastaa, yatraa	यस्ता, यत्रा	DDO	
Unmarked demonstrative determiner	yo[7], yas#, yi, yin#, yinii, yati, yatti	यो, यस#,यी, यिन#, यिनी, यति, यत्ति	DDX	

Interrogative pronoun-determiners

- The tags for these words (the k-forms) follow the same pattern as those for the y-, ty- and w-forms. The same comments apply.

- Note that when *ke* is used to indicate a yes-no question, it does not receive a DKX1 tag. It receives a special tag instead, the QQ tag for "question marker". See also below.

[7] As the emphatic form of *yo*, *yahii* would be tagged in the same way: DDX.

Category definition	Examples (Latin)	Examples (Devanagari)	Tag (Latin)	Tag (Devanagari)
Masculine interrogative determiner	kasto, katro	कस्तो, कत्रो	DKM	
Feminine interrogative determiner	kastii, katrii	कस्ती, कत्री	DKF	
Other-agreement interrogative determiner	kastaa, katraa	कस्ता, कत्रा	DKO	
Unmarked interrogative determiner	ko, kas#, ke, kun, kati	को, कस#, के, कुन, कति	DKX	

Relative pronoun-determiners

- The tags for these words (the j-forms) follow the same pattern as those for the ty-, w- and y-forms. The same comments apply.

Category definition	Examples (Latin)	Examples (Devanagari)	Tag (Latin)	Tag (Devanagari)
Masculine relative determiner	jasto, jatro	जस्तो, जत्रो,	DJM	
Feminine relative determiner	jastii, jatrii	जस्ती, जत्री	DJF	
Other-agreement relative determiner	jastaa, jatraa	जस्ता, जत्रा	DJO	
Unmarked relative determiner	jo, jas#, je, jati, josukai[8]	जो, जस#, जे, जति, जोसुकै	DJX	

Pronoun-determiners: general

- The "general" pronoun determiners do not follow the five-way morphological pattern outlined above. Syntactically and conceptually, however, words like *arko* and *aruu* are similar to words like *yatro* (and dissimilar to adjectives such as *raamro*, which is the next-most-similar category). For this reason they are grouped in this separate category.

Category definition	Examples (Latin)	Examples (Devanagari)	Tag (Latin)	Tag (Devanagari)
Masculine general determiner-	arko	अर्को	DGM	

[8] *junsukai*, *jesukai*, and *jatisukai* would also belong in this category.

pronoun				
Feminine general determiner-pronoun	arkii	अर्की	DGF	
Other-agreement general determiner-pronoun	arkaa	अर्का	DGO	
Unmarked general determiner-pronoun	aruu	अरू	DGX	

Verbs

Introduction

- Nepali has a vast number of verb inflections. This is because compounding is an extremely productive process in the verb system of Nepali. Different combinations of various non-finite forms of the main verb and various inflected forms of secondary verbs creates a very large number of tense-aspect-mood combinations[9].

- If every distinction made in the verb system were to be indicated by a separate tag, then the tags for verbs would become entirely unmanageable. Thousands of tags would be required for the verbs alone.

- Furthermore, there is an issue of consistency. For many compound verb forms, some variation exists as to whether they are written as one word or two. If specific tags were defined for inflectional forms that were actually compound verbs, then it would become necessary for those tags to be applied across two tokens if the form were written as two words.

- To avoid this, the following tagging system for verbs uses this rule:

Every compound verb is tagged according to the last verb in the compound.

- That means that only the last identifiable verb in a compound verb is taken into account when deciding the tag. If there is only one identifiable verb, then the whole thing is taken into account. For example:

Verb form	Take into account…
garis	garis
garnubhaena	bhaena
garnecha	cha
garirahibaksanthyo	baksanthyo
garidiyo	diyo
garirahanuhuncha	cha
garirahekaa	rahekaa
gardina~	gardina~

[9] The term "compound verb" is used in a wide sense here, to refer to *any verb that consists of two or more elements that are identifiable as belonging to separate independent verb lexemes*. So, for instance, *garidiyo* would be a straightforward example of a compound verb; however, for the purposes of this description, words such as *garthyo*, *garcha*, *garirahyo*, and *garnecha* will be considered to be compound verbs as well. Compounds consisting of noun+verb or adjective+verb can be tagged using the same rule: only consider the second element, i.e. the last verb.

garibaksanuhu~dorahecha	cha
garnuhune	hune

- Once it is determined how much of the verb should be taken into account, its ending can be looked up on the charts below and its tag worked out.

- If a form which can be written as a single word is written as two words, then each is considered individually.

- For example, *garidiyo* (and similar constructions) would usually be written as one word, but if it were written as two words (*gari diyo*), each part would receive a separate tag according to the rules laid out below.

- These means that there are no specific tags for high or royal honorific forms of the verb, since these are always compounds: *garnubhayo*, etc. being a compound with *hunu*, and *garibaksyo*, etc. being a compound with *baksanu* / *baksinu*.

- As a further simplifying measure, certain aspects of the verb morphology system are ignored. These are passive, causative, and negative. This means that:
    - Any passive verb is tagged the same as the corresponding active verb
        - E.g. tag *garinu* the same as *garnu*
    - Any causative verb is tagged the same as the corresponding non-causative verb
        - E.g. tag *garaau~cha* the same as *garcha*
    - Any negative verb is tagged the same as the corresponding positive verb
        - E.g. tag *garnechaina* the same as *garnecha*

- No distinction is made in the tagging between auxiliary verbs and main verbs.

Non-finite inflections: infinitives, participles, command forms and others

- For the purposes of this tagset description, a verb-form is considered *finite* if it is marked for person, and *non-finite* if it lacks marking for person. (This division is for *convenience of reference only* and does not constitute a claim about finiteness in Nepali.)

- Many of the non-finite forms occur embedded at the start of a longer verb (e.g. **gardai**thyo, **garnu**huncha). However, in accordance with the general rule, they are **not** tagged separately in this case.

- This means that non-finite tags are **only** used *if the non-finite form is written as a separate word*, or *if the non-finite form is at the end of the longer verb*.

Infinitives

- The form referred to here as the *ne*-participle, but described elsewhere variously as the *imperfect participle* or the *infinitival participle*, is not included in this section, but is grouped with the participles below.

Category definition	Examples (Latin)	Examples (Devanagari)	Tag (Latin)	Tag (Devanagari)
Infinitive verb	garnu, garna, garnaa[10], nagarnu, nagarna, nagarnaa	गर्नु, गर्न, गर्ना, नगर्नु, नगर्नु, नगर्ना	VI	

Participles

- The following groups of non-finite forms are grouped together for convenience as *participles*, although they should probably not all be considered "participles" in the most precise grammatical sense[11].

- They fall into the following broad groups:
  - the *e(ko)*-participle, sometimes called the *perfect participle*
  - the *d*-participle, which is used for three functions, all covered by four cartegories:
    - the converb/participle function (also known as the *conjunctive participle*, the *progressive participle*, and the *simultaneous converb*): ending *–dai*, which has a variant form in *–daa*
    - the modifier function, where the d-participle follows the adjectival pattern of inflection: so *–dii* and *–daa* are the "feminine" and "other" forms
    - the *d*-participle is also used as an element of compound verbs (where it does not receive a separate tag, as per the usual rule)

[10] The infinitive in *–naa* is a spelling variant of the infinitive in *–na* when followed by a postposition.
[11] Some are "converbs" and some may be more precisely analysed as infinitival.

- the *ne*-participle (described by Acharya 1991 as the *imperfect participle* and by Schmidt et al. 1993 as the *infinitival participle*)
- the sequential converbs, also called *absolutive participles*, of which there are three, which all receive the same tag:
  - the *era*-participle
  - the *ii*-participle
  - the *iikana*-participle

- The *e(ko)*-participle is so-called because it is based on a participial form in *–e*. However, that form rarely emerges. It may be found in the phrase *gae saal* "last year", or before the postposition *jati*, "about (of an amount)", e.g. *garejati* "as much as (something e.g. work) done". The participle is usually found followed by *ko/kii/kaa*. The form *gareko* would be tagged as *gare*_VE *ko*_IKM (with the *ko* tagged as a separate unit as per usual).

Category definition	**Examples (Latin)**	**Examples (Devanagari)**	**Tag (Latin)**	**Tag (Devanagari)**
Masculine *d*-participle verb	gardo, nagardo	गर्दो, नगर्दो	VDM	
Feminine *d*-participle verb	gardii, nagardii	गर्दी, नगर्दी	VDF	
Other-agreement *d*-participle verb	gardaa, nagardaa	गर्दा, नगर्दा	VDO	
Unmarked *d*-participle verb	gardai, nagardai	गर्दै, नगर्दै	VDX	
e(ko)-participle verb	gae (as in *gae saal*), gare (as in *garejati* or *gareko*)	गरे	VE	
ne-participle verb::	garne, nagarne	गर्ने, नगर्ने	VN	
Sequential participle-converb	garera, gariikana, garii, nagarera, nagariikana, nagarii	गरेर, गरीकन, गरी, नगरेर, नगरीकन, नगरी	VQ	

Command forms

- Command forms are variously referred to as *imperatives*, *request forms*, etc.

- The optative forms of the verb (see below) can also be used to issue polite commands or requests. However, the optative has a separate set of personal endings.

- There are three different command forms: each is given a separate tag.

- Examples based on the verb *jaanu* have been added to the usual examples from *garnu* in the table below.

Category definition	Examples (Latin)	Examples (Devanagari)	Tag (Latin)	Tag (Devanagari)
Command-form verb, non-honorific	gar, jaa	गर्, जा	VCN	
Command-form verb, mid-honorific	gara, jaau, jaao	गर, जाऊ, जाओ	VCM	
Command-form verb, high-honorific	garnos, jaanos	गर्नोस्, जानोस्	VCH	

Other non-finite forms

- The remaining non-finite forms fall into the following broad groups:
    - The *e*-form (often referred to as *subjunctive* or *conditional*), **not** to be confused with the *e*-participle.
    - the *i*-form, which is the form in which a non-final verb in a compound appears. This form is sometimes referred to as the *passive root* – this is **not** to be confused with the *ii*-participle. It would most frequently be part of a compound, but if it is written alone, it would receive the tag given below.

Category definition	Examples (Latin)	Examples (Devanagari)	Tag (Latin)	Tag (Devanagari)
Subjunctive / conditional *e*-form verb	gare, nagare	गरे, नगरे	VS	
i-form verb	gari	गरि	VR	

"Finite" inflections: forms with inflection for person and tense

- There are six finite forms, which are indicated in the third person singular by the following typical endings: *-yo, -thyo, -echa, -cha, -necha, -laa*. The descriptive labels given to them by Acharya (1991) and other sources, together with the labels used in this tagset description ("here"), are listed below:

Form	Acharya's description	Other description(s)
-yo	Past – known – simple	Past, simple past Here: *yo*-past
-thyo	Past – known – habitual	Past-habitual Here: *thyo*-past
-echa	Past – unknown	Evidential past Here: *echa*- past
-cha	Present	Non-past, present Here: *cha*-present
-necha	Future – definite	Certaintive; definitive Here: *necha*-future
-laa	Future – nondefinite	Possibilitive; probabilitive Here: *laa*-future

- There is an additional finite form, the *optative*. (This is given separate tags: see the following section.)

- Many of these tenses follow similar patterns of inflection. There are in fact only nine sets of endings[12]: four positive, three negative, and two unique patterns (*ho* and *cha*).

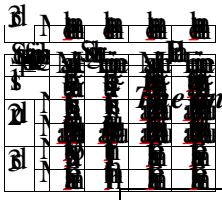
- The sets of endings in question are as follows:

  Unique verb tenses: ho (its negative follows [o]ina)
  cha (its negative follows ina)
  (bhayo, thiyo both follow the usual pattern of the yo set)

  Positive: The yo set
  The thyo set [NOT identical to thiyo etc.]
  The laa set [also used in positive with na- as prefix]
  The optative set [in –os etc.] (not included here, see below.)

  Negative: The oina set
  The ena set
  The daina set

[12] Nine is a high estimate: many of the listed sets of endings are very similar to one another, and further analysis could reduce the number different patterns of verb endings considerably.



The readings are tabulated below.

Unique tenses: positive		**Singular**		**Plural**	
		Masculine	***Feminine***	***Masculine***	***Feminine***
1st		hu~ chu	hu~ chu	hau~ chau~	hau~ chau~
2nd	***N***	hos chas	hos ches	hau chau	hau chau
	M	hau chau	hau chyau	hau chau	hau chau
	N	ho cha	ho che	hun chan	hun chan

Sets of endings: negative		**Singular**		**Plural**	
		Masculine	***Feminine***	***Masculine***	***Feminine***
1st		oina~ ina~ dina~	oina~ ina~ dina~	oinau~ enau~ dainau~	Oinau~ enau~ dainau~
2nd	***N***	oinas inas dainas	oinas inas dinas	oinau enau dainau	oinau enau dainau
	M	oinau enau dainau	oinau inau dinau	oinau enau dainau	oinau enau dainau
3rd	***N***	oina ena daina	oina ina dina	oinan enan dainan	oinan enan dainan
	M	oinan enan dainan	oinan inan dinan	oinan enan dainan	oinan enan dainan

- The following distinctions operate on finite verbs:
  - Person: first, second, third
  - Gender: masculine (default), feminine
  - Number: singular, plural
  - Honorific level: non-honorific, medial-honorific[13]

- Therefore, in theory there should be 3 x 2 x 2 x 2 = 24 tags. However, not all possible combinations of features have separate forms in Nepali. There is, for instance, no specific form in any tense-mood for the second-person non-honorific feminine plural. There is just a single second-person plural form.

- In general, only second and third person singular verbs are marked for gender and honorific level. Plural verb forms are not marked for gender or for honorific level. First person verbs are not marked for gender or for honorific level.

- Similarly, at the medial-honorific level, there is no distinction between singular and plural, except for feminine verbs (which are listed separately below).

- This means that if we take only gender, number and honorific level into account, and if we treat alike those categories which are merged together, only ten tags are necessary.

[13] As noted, the higher honorific levels (high, royal) are conveyed through compound verbs.

- In many cases gender is not marked on verbs. Even where it is marked, it is reported that the "masculine" verbs may be used in some varieties of Nepali with feminine subjects (see Hutt and Subedi 2003: 37-38).

- For this reason, the tagset considers "masculine" to be the default option. So the six "masculine" tags listed in the table below, are in fact not defined as masculine: they are simply the usual, default forms. The feminine forms, discussed subsequently, and indicated by adding F to the tag, are defined in contrast to the default.

- The different positions in the tags for finite verbs contain the following letters:

	Person	Honorific	Number	Gender
VV…	M	N	1	( )
	T	M	2	F
	Y	X		

- The tags start with "VV": the second "V" identifies them as general finite verbs, in contrast to the optative verbs in the next section (VO) and the various non-finite forms discussed above.

- In the column for *person* above, the M stands for *ma* (=first person), the T for *ta* (= second person), and the Y for *yo* (= third person).

- Note that number here is **very different** to number as marked on nouns, adjectives, etc. On those words, number related solely to the presence or absence of *haruu* (which is tagged as a separate word). By contrast, for finite verbs, number is an inflectional category indicated by the same endings that indicate person and (sometimes) gender/honorific.

Category definition	**Examples (Latin)[14]**	**Examples (Devanagari)**	**Tag (Latin)**	**Tag (Devanagari)**
First person singular verb	gare~, garthe~, garina~, chu, hu~, garnechu	गरेँ, गर्थेँ ,गरिनँ, छु, हुँ, गर्नेछु	VVMX1	
First person plural verb	garyau~, garthyau~, garenau~, chau~, hau~, garnechau~	गर्यौँ, गर्थ्यौँ, गरेनौँ, छौँ, हौँ, गर्नेछौँ	VVMX2	

[14] It should be remembered that only the last element of a verb is taken into account. So any verb that ends in *cha* – such as *garibaksecha* or *garnubhaecha* – will also be tagged VYN1. At first consideration, it seems highly counter-intuitive that a high or royal honorific verb should be tagged as a "third person non-honorific singular verb". However, what is actually happening is that the *–cha* at the end is being tagged (quite correctly) as "third person non-honorific singular verb" and the rest of the compound verb (*garibakse–*, *garnubhae–*, etc.) is being ignored. The analysis of the high/royal inflections (along with other compound verb features, such as the progressive form with *rahanu*) is in effect being removed from the domain of morphosyntactic categories and added to the domain of derivational morphology, in accordance with this section's general aim, which is to simplify the analysis of verbs. Of course if the earlier part of the compound verbs (including high and royal honorific forms) were written as separate words for any reason, they could be given the appropriate non-finite tag.

Second person non-honorific singular verb	garis, garthis, garinas, chas, hos, garnechas	गरिस्, गर्थिस्, गरिनस्, छस्, होस्, गर्नेछस्	VVTN1	
Second person plural (or medial-honorific singular) verb	garyau, garthyau, garenau, chau, hau, garnechau	गर्यौं, गर्थ्यौं, गरेनौ, छौ, हौ, गर्नेछौ	VVTX2	
Third person non-honorific singular verb	garyo, garthyo, garena, cha, ho, garnecha	गर्यो, गर्थ्यो, गरेन, छ, हो, गर्नेछ	VVYN1	
Third person plural (or medial-honorific singular) verb	gare[15], garthe, garenan, chan, hun, garnechan	गरे, गर्थे, गरेनन्, छन्, हुन्, गर्नेछन्	VVYX2	

- Turning to the feminine forms of the verb, there are four of them, because there are four places in the verb paradigm where a feminine verb can differ from a maximum verb – the second person and third person singular, non-honorific and medial-honorific.

- The primary rule is that a feminine tag should **only** be used if the feminine form is different to the masculine-default form. This is not always the case. For example, *cha* is VYN1 and *che* is VYN1F. But the negative form (*chaina*) is always tagged VYN1 regardless of whether it has a masculine or feminine subject, because there is no specifically feminine form here.

- In short, V_____F tags should only be given to words that cannot ever be used with a masculine subject.

Category definition	**Examples (Latin)**	**Examples (Devanagari)**	**Tag (Latin)**	**Tag (Devanagari)**
Feminine second person non-honorific singular verb	garlis, ches, garthis	गर्लिस्, छेस्, गर्थिस्	VVTN1F	
Feminine second person non-honorific singular verb	garthyau, chyau	गर्थ्यौं, छ्यौ	VVTM1F	
Feminine third person medial-honorific singular verb	garina, garii[16], che, garthii	गरिन, गरी, छे,	VVYN1F	

[15] This is identical in form to the non-finite *e*-form. Context should distinguish the two.
[16] Not to be confused with the non-finite *ii*-form, discussed above, which need not be feminine. Context should distinguish the two.

		गर्थी		
Feminine third person medial-honorific singular verb	garin, garthin, garinan, chin	गरिन्, गर्थिन्, गरिनन्, छिन्	VVYM1F	

Optative verbs

- There are separate tags for optative verbs, as these verbs behave differently in many ways to the other finite verbs (e.g. by taking a prefix to indicate the negative, rather than a suffix).
- The tags (beginning in VO-) are directly parallel to the VV- tags for general finite verbs.
- There are no feminine forms in the optative, and thus no tags for them.
- The endings of optative verbs are as follows:

Sets of endings: optative		**Singular**		**Plural**	
		Masculine	***Feminine***	***Masculine***	***Feminine***
1st		u~	u~	au~	au~
2nd	***N***	es	es	e	e
	M	e	e	e	e
3rd	***N***	os	os	un	un
	M	un	un	un	un

Category definition	**Examples (Latin)**	**Examples (Devanagari)**	**Tag (Latin)**	**Tag (Devanagari)**
First person singular optative verb	jaau~, garu~	जाऊँ, गरूँ	VOMX1	
First person plural optative verb	jaaau~, garau~	जाऔँ, गरौँ	VOMX2	
Second person non-honorific singular optative verb	gaes, gares	गएस्, गरेस्	VOTN1	
Second person plural (or medial-honorific singular) optative verb	gae, gare	गए, गरे	VOTX2	
Third person non-honorific singular optative verb	jaaos, garos	जाओस्, गरोस्	VOYN1	
Third person plural (or medial-honorific singular) optative verb	jaauun, garuun	जाऊन्, गरून्	VOYX2	

Adverbs

- There are three closed categories and one open category of adverbs. The three closed categories correspond to the categories of pronoun-determiners (i.e. D~K~J), to which they are morphologically related.

Category definition	Examples (Latin)	Examples (Devanagari)	Tag (Latin)	Tag (Devanagari)
Adverb	raamrarii, ekdam, chiTo	राम्ररी, एकदम, छिटो	RR	
Demonstrative adverb[17]	yataa, utaa, tyataa; ahile, ahilyai, yahii~, yasarii, aba	यता, उता, त्यता, अहिले, अहिल्यै, यहीं, यसरी, अब	RD	
Interrogative adverb	kataa, kahaa~, kahile, kasarii	कता, कहाँ, कहिले	RK	
Relative adverb[18]	jataa, jahaa~, jahile, jasarii, jaba	जता, जहाँ, जहिले	RJ	

[17] *tyahaa~* also belongs in this category, as do *yahaa~* and *wahaa~* if they are used as adverbs (although yahaa~ and *wahaa~* are also pronouns: see above). Adverbs in *t-* or *w-* that are parallel to adverbs in *y-* (e.g. *taba* parallel to *aba*; usarii and tyasarii parallel to yasarii) would be included here as well.

[18] *jahaa~sukai* and *jahilesukai* would also be members of this category.

Postpositions

- Postpositions are tagged separately to the nouns they follow (including the ones often described as affixes: *haruu*, *le*, *laai*, *ko/kii/kaa*). This is true regardless of whether they are written as part of the same word as the noun or as a different word.

- (See the comments in the section on **Nouns** for further discussion of this issue.)

- If a postposition is written as part of the same word as the noun, it should be detached from the noun prior to being tagged, either manually or automatically.

- It is possible for multiple postpositions to occur in sequence (written separately or joined together). For example:

  raamsangabaaTa: raam_NP sanga_II baaTa_II
  ghara agaaDibaaTa: ghara_NN agaaDi_II baaTa_II

- Some postpositions (e.g. agaaDi and pachaaDi) are also used as adverbs. If used as adverbs they are tagged as RR.

- There are separate subcategories for the postpositions most popularly perceived as grammatical affixes.

- One of these, *ko*, has three tags, because it has inflected forms (*kii* and *kaa*).

Category definition	**Examples (Latin)**	**Examples (Devanagari)**	**Tag (Latin)**	**Tag (Devanagari)**
Postposition	agaaDi, pachaaDi, baaTa, dwaaraa, maa, maathi, saath, puurvak, tira, tarpha, vasa, sanga, binaa	अगाडि, पछाडि, बाट, द्वारा, मा, माथि	II	
Plural-collective postposition	haruu	हरू	IH	
Ergative-instrumental postposition	le	ले	IE	
Accusative-dative postposition	laaii	लाई	IA	
Masculine genitive postposition	ko	को	IKM	
Feminine genitive postposition	kii	की	IKF	
Other-agreement genitive	kaa	का	IKO	

postposition				

- Although it is referred to, for purposes of classification, as "genitive", the meaning *in context* of *ko/kii/kaa* is not always genitive. For example, *yatiko* has a different use and meaning to *yasko* (even though both are determiners followed by *ko*).

- Similarly, yatikii ad yattikii would have different meanings.

- But these considerations of meaning do not affect the tagging. *ko* is IKM whether it comes after *yas* or after *yati*; *kii* is IKF in both *yatikii* and *yattikii*.

- It is anticipated that the distinction between pronouns like *yasko* and pronouns like *yatiko* would be an important one to make at a higher level of analysis such as parsing.

- There are also some oddities of gender agreement when *ko/kii/kaa* comes after the royal pronoun *sarkaar*.
    - In positions where *sarkaarkii* or *mausuphkii* (or the equivalent forms with *haruu*) would be expected, *sarkaarkaa* or *mausuphkaa* occurs instead (e.g. *sarkaarkaa suputrii, mausuphkaa suputrii* "his/her-(royal) daughter", or *sarkaarkaa mahaaraanii*, "his queen". But *kaa* would be tagged as IKO here regardless.
    - However there are contexts where *sarkaarkii* can occur: for instance सरकारकी जय, सरकारहरूकी जय (*sarkaarkii jaya*, *sarkaarharuukii jaya* "long live his/her majesty, long live their majesty"). The feminine tag IKF would be used here, to reflect the morphosyntactic pattern, even though the meaning is not feminine in the usual sense.

Numerals and numeral classifiers

- There is a tag for cardinal numerals and several tags for ordinal numerals.
- The tag for cardinal numerals is used for numbers written as digits as well (whether in Devanagari or Latin numbers).
- The tag for cardinal numerals (MM) is also used for certain quantifying elements which, though they may be classed in conceptual terms as adjectives or nouns, behave in terms of their distribution in ways similar to the cardinal numerals. These words are: *kaiyau~* "several", *kehi*, "some/few", and *dherai*, "many".
- Ordinal numerals have parallel tags to adjectives: some are marked, some are not.
- Other words derived from numerals are tagged using the normal tags. For instance, *dohoro* would be tagged RR or JJX.
- Ordinal numerals derived from Sanskrit are tagged as MOX. For instance, *prathama*_MOX, *dvitiya*_MOX[19].
- Ordinal numerals pattern as marked adjectives; cardinal numbers pattern as unmarked adjectives.

Category definition	Examples (Latin)	Examples (Devanagari)	Tag (Latin)	Tag (Devanagari)
Cardinal number	ek, eu#, yau#, dui, tin, caar, paa~c	एक, दुई, तीन, चार, पाँच	MM	
Masculine ordinal number	pahilo, dosro, tesro, cautho	पहिलो, दोस्रो, तेस्री, चौथो	MOM	
Feminine ordinal number	pahilii, dosrii, tesrii, cauthii	पहिली, दोस्री, तेस्री, चौथी	MOF	
Other-agreement ordinal number	pahilaa, dosraa, tesraa, cauthaa	पहिला, दोस्रा, तेस्रा, चौथा	MOO	
Unmarked ordinal number	paa~cau~	पाँचौं	MOX	

[19] Note however that the words for the days of the lunar half-months that derive from these Sanskrit numerals would be tagged NN: so, *prathamaa*_NN, *dvitiyaa*_NN, and so on.

- Cardinal numerals frequently (but not always) occur with a *classifier* after them. The classifier may be attached as a clitic, or it may be written as a separate word.

- There are further tags for these classifiers. They are *wataa* and *janaa*.

- These classifiers can also attach to words other than cardinal numerals, e.g. *kati*.

- *waTaa* is marked for gender, but it does not have an *–o* form (similar to adjectives like *bahulaahaa* – see the section on adjectives above; therefore there are only two forms, namely although **not** according to the adjectival pattern as there are only two forms: *waTaa* ("other", MLO, although in this case "other" includes masculine singular) and *waTi* ("feminine": MLF).

- *waTaa* can occur in the –o form in one situation: when it is *euTo*. This would be tagged *eu*_MM *To*_MLM.

- However, *janaa* does not occur in different forms for different genders in this way.

- *waTaa* has three variant forms. The first variant is *#Taa*, which attaches to *dui* and to *eu#*, the contextual variant of *ek*, to form *euTaa* (or: *yauTaa*) and *duiTaa*. The other variants are and *#oTaa* and #*auTaa*[20], which can attach to other numerals (e.g. *caaroTaa* as a variant of *caarwaTaa*).

- If *waTaa* – in whatever form – is attached to a numeral or another word, it is separated from it in tokenisation. The same applies to *janaa*.

- When *#Taa* is separated from *eu#* or *yau#*, the remaining part of the word is tagged as MM.

Category definition	**Examples (Latin)**	**Examples (Devanagari)**	**Tag (Latin)**	**Tag (Devanagari)**
Masculine numeral classifier	*#To* [as in euTo]	#टो	MLM	
Feminine numeral classifier	*(#)waTii, #Tii, #oTii, #auTii*	(#)वटी, #टी, #ओटी	MLF	
Other-agreement numeral classifier	*(#)waTaa, #Taa, #oTaa, #auTaa*	#वटा, #टा, #ओटा, #औटा	MLO	
Unmarked numeral classifier	*(#)janaa*	(#)जना	MLX	

[20] The form with #*auTaa* is less preferred.

Conjunctions

- Relativisers such as *jo, je, jahaa~* are not counted as conjunctions – see pronouns and adverbs.

- This list of subordinating conjunctions is not comprehensive: others include *kinabhane*.

Category definition	**Examples (Latin)**	**Examples (Devanagari)**	**Tag (Latin)**	**Tag (Devanagari)**
Coordinating conjunction	ra, tathaa	र, तथा	CC	
Subordinating conjunction appearing **after** the clause it subordinates	bhanne, bhanii, bhanera[21]	भने, भनी, भनेर	CSA	
Subordinating conjunction appearing **before** the clause it subordinates	ki, yadi, yaddyapi, kinaki	कि, यदि, यद्यपि, किनकि	CSB	

[21] The forms in *bhan-* are also non-finite forms of the verb *bhannu*, and if they are not used as conjunctions would have the normal verbal tags.

Particles

- This category of closed-class words are also referred to as *discourse-pragmatic particles* and *nuance particles*.
- Note that some of the particles have a wide variety of meanings (e.g. *na*).
- The full list of particles identified so far is given below:

nai	maatra	kewal	caahi~	pani	hai
na	ni	ra	re	kyaare	hagi
hau	byaare	ta	po	kyaa	lau
la	khai	ho			

Category definition	**Examples (Latin)**	**Examples (Devanagari)**	**Tag (Latin)**	**Tag (Devanagari)**
Particle	nai, caahi~, pani, hai, ra, re, kyaare, ho, khai	नै, चाहिँ, पनि, है, र, रे, क्यारे, हो, खै	TT	

Words derived using –ai

- The morpheme *–ai*, which indicates emphasis or focus, and may sometimes be translated as "only" or "very", is often considered to be a particle, like the words listed above.

- However, unlike those words, it occurs attached to another word (a noun, adverb, etc.), and never occurs as a word on its own.

- Furthermore, it can cause changes in the word it is attached to, if that word ends in a vowel.

- Furthermore, when attached to words, it can occur closer to the root than postpositions, though it does not always (e.g. *kalamaile*, *kalamleai*).

- For these reasons, the addition of *–ai* to a word will be considered an aspect of derivational morphology rather than inflection. Words which end in *–ai* will be treated as single words and tagged as such.

- Adding *–ai* to a word may change its category if it replaces an inflectional ending For instance, *raamro* is JM, whereas *raamrai* is JX (because the *–ai* affix does not inflect for gender/number).

- In other cases, adding *–ai* may not change the category. For instance, *sab* is JJX and *sabai* is still JJX. Similarly *kalamle* is NN—IE, *kalamaile*, *kalamleai* are both also NN—IE (*kalamai*_NN *le*_IE, *kalam*_NN *leai*_IE).

- Adding *–ai* to a possessive pronoun, or the genitive postposition will change its category. For instance, *aaphnai* (=*aaphno*+*ai*) is not PRFKM any more because it doesn't show masculine agreement: *aaphnai* can be used before nouns of any gender and number.

- In the main tagset, there are however no tags for unmarked (X) genitive case words – only for masculine (M), feminine (F) and other (O). That is, there is no "PRFKX" tag for *aaphnai*. So this tag is listed separately here (to avoid complicating the main listings of tags).

- A similar consideration applies to the other possessive pronouns, and to *kai* (*ko*+*ai*).

- **If other words are discovered where a new tag is needed for forms ending in *–ai*, they will be added, with appropriate comments, here.**

Category definition	Examples (Latin)	Examples (Devanagari)	Tag (Latin)	Tag (Devanagari)
Possessive reflexive pronoun without agreement	*merai*	मेरै	PMXKX	
Non-honorific second person possessive pronoun without	*terai*	तेरै	PTNKX	

agreement				
Medial-honorific second person possessive pronoun without agreement	*timrai*	तिम्रै	PTMKX	
Possessive reflexive pronoun without agreement	*aaphnai*	आफ्नै	PRFKX	
Unmarked genitive postposition	*kai*	कै	IKX	

Question marker

- The word *ke* is an interrogative pronoun (see above). But it is also used, with no referent, to indicate a yes-no question.

- When *ke* is used in this way, it receives the QQ tag. When it is used as an interrogative pronoun, to refer to an unknown entity, it receives a DK[…] tag, as described above.

Category definition	**Examples (Latin)**	**Examples (Devanagari)**	**Tag (Latin)**	**Tag (Devanagari)**
Question marker	*ke*	के	QQ	

Interjections

- The interjections given in the table below are a very small sample of the possible interjections.

Category definition	Examples (Latin)	Examples (Devanagari)	Tag (Latin)	Tag (Devanagari)
Interjection	oho, aahaa, hare	ओहो, आहा, हरे	UU	वव

Punctuation

- An ellipsis (…) will be tagged as a single instance of YM, rather than as three instances of YF.

Category definition	**Examples**	**Tag (Latin)**	**Tag (Devanagari)**
Sentence-final punctuation	? ! . । ॥	YF	
Sentence-medial punctuation	, ; : :- / –	YM	
Quotation marks	' "	YQ	
Brackets	( ) { } [ ]	YB	

Residual

- These tags are for various bits of text that are not really linguistic words.
- They must be used **sparingly**.
- They should only be used if something cannot possibly be put in one of the categories above.
- In particular, the FU tag for unclassifiable should **NOT** be used for Nepali words which are difficult to tag – the F-tags are **NOT** to be used as dustbin categories!
- Similarly, the FF tag for foreign words should **NOT** be used for loan-words, which are words from (in particular) Hindi, Sanskrit or English that have become part of the Nepali language.
- ***If it is at all possible***, words should be tagged using the Nepali tags[22]. The FF tag is for use only where that is not possible (e.g. if the word has an inflectional ending from another language on it, or is used as part of a sentence in another language, and so on).
- Words written in an alphabet other than Devanagari – such as Latin, Gujarati, Perso-Arabic, Japanese, or whatever – should be tagged FS.

Category definition	**Examples (Latin)**	**Examples (Devanagari)**	**Tag (Latin)**	**Tag (Devanagari)**
Foreign word in Devanagari			FF	
Foreign word, not in Devanagari			FS	
Abbreviation	M.P.P.	म. पु. पु.	FB	
Mathematical formula (and similar)	$e=mc^2$		FO	
Letter of the alphabet[23]			FZ	
Unclassifiable			FU	

[22] For instance, the Nepali word *koT* (कोट) "coat" is derived from English, but it has been fully incorporated into Nepali and as such it would be given the tag NN1 and ***NOT*** FF.

[23] This tag is for letters of the alphabet used on their own, not as words. Words which are made up of a single letter (for instance, म *ma*) should ***not*** be given that tag. For instance, in English, the letter "A" is a word on its own (the indefinite article): it would be tagged as an article when it is used as an article, but as FZ if it is used e.g. in the sequence "A is for Apple, B is for Boat …".

The NULL tag

- The NULL tag is given to XML elements – which obviously should not receive tags in the normal way of things – in any processing situation where they *have* to be given a tag.
- This tag indicates that the XML element is not a word and does not need a tag.
- It is, in effect, shorthand for "no tag needed".
- The NULL tag should always be added, where necessary, by computer. *A human analyst should never need to use the NULL tag*.

Category definition	**Examples (Latin)**	**Examples (Devanagari)**	**Tag (Latin)**	**Tag (Devanagari)**
Null tag: an element of the text which does not need a tag	<p>		NULL	NULL[24]

[24] There cannot be a Devanagari version of the NULL tag. NULL is NULL.