Chapter 4

FOUNDATIONAL CONSIDERATIONS

In this chapter we present the different word classes of Nepali Language with their grammatical constituents and their annotation procedure. We also discuss several important issues related to Computational Morphology of Nepali Language. In this chapter we also presented and described the Nepali tagset which is used for our experiment and tagset is a very important issue which can affect the tagging accuracy.

4.1 Nepali Language overview

Nepali language is originally belongs to the Indo-Aryan branch of indo-European family [50]. This language takes its root from Sanskrit which is the classical language of India. Nepali language was known as Gurkha, Gurkhali or Khas Kura. In the 11th century AD Nepali Language developed from the Brahmi script. Nepali Language is written with the Devanagari alphabet. Nepali is spoken by more than 40 million people, mostly in Nepal, Bhutan, Myanmar, West Bengal and other parts of India. Linguistically, Nepali is most closely related to Sanskrit and Hindi. A large proportion of the technical vocabulary written in Nepali is influenced by Sanskrit. [65, 66].

Nepali is written in the Devanagari script and there are 12 vowels and 36 consonants in this language. The script is written from left to right [65]. There is no provision of capital

and small letters in the script [51, 56, 57]. The Nepali alphabets are written in two separate groups, namely the vowels and the consonants and it is shown below.

Vowels: अ (a), आ (a), इ (i), ई (i), उ(u), ऊ(u),

ए (e), ऐ(ai), ओ(o), औ(au), ऑ(am), ऒ(ah)

Consonants: क(ka), ख(kha), ग(ga), घ(gha), ङ (na),

च(ca), छ (cha), ज(ja), झ (jha), ञ(na),

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), रि (ra), ल (la), व(va), स(sa),

ष(sa), श(sa), ह (ha), क्ष(ksa), त्र(tra), ज्ञ(gya)

4.2 Computational morphology

Computational morphology is one of the important parts of computational linguistics [54] which deals with the processing of words in both their graphemic, i.e. written form and phonemic, i.e. spoken form.It includes the analysis of word formation methods, morpheme

segmentation, hyphenization and error correction etc. These tasks may appear very easy to a human but they create difficult problems to a computer program.

The Nepali Grammar consists of both the inflected and uninflected forms; it is called as open and closed classes as well. These constitute the parts of speech of the Nepali Grammar [54]. Noun, adjective, verb and adverb come under the open class where as pronoun, coordinating conjunction, subordinating conjunction, interjection, postposition, vocative and nuance particle come under the closed class.

4.2.1 Noun and Lexical Morphology

In Nepali language there is no gender marked on noun and generally the noun is not inflected and only has a nominative (unmarked) form. If we see the traditional grammar it shows nouns has an inflected form for gender, number and seven cases.

eg. chhoro 'son' vs. chhori 'daughter'

but modern grammar prefers to treat such form as separate lexical items, independent of each other. Nepali noun shows inflectional contrasts for singular vs. plural eg. मानिस(manis) मनिसहरू(manisharu) and for seven different cases. The case-number suffixes are shown in the table below.

Cases	**Singular**	**Plural**
Nominative(Nm)	---	हरु
Accusative(Ac)	लाई	हरुलाई
Instrumental(In)	ले	हरुले
Dative(Dt)	लाई	हरुलाई

Ablative(Ab)	बाट	हरुबाट
Genitive (Gn)	को	हरुको
Locative(Lc)	मा	हरुमा

Table 4.1: The number and case suffixes of nouns

The inflection table of noun kitab/किताब 'book' looks like as shown below

Cases	**Singular**	**Plural**
Nominative(Nm)	किताब	किताबहर
Accusative(Ac)	किताबलाई	किताबहरलाई
Instrumental(In)	किताबले	किताबहरले
Dative(Dt)	किताबलाई	किताबहरलाई
Ablative(Ab)	किताबबाटि	किताबहरबाटि
Genitive (Gn)	किताबको	किताबहरको
Locative(Lc)	किताबमा	किताबहरमा

Table 4.2: The number and case inflection of noun Book/किताब

The most frequent noun-forming derivational suffixes are याइ, आइ. and some of the characteristics of Nepali nouns are explained below

i) When the suffix -हरु added to nouns, then the noun becomes a plural form.

e.g. .राजा(king) + हरु = राजहरू (kings)

ii) When demonstratives यो and यो change to यी andती the then the noun becomes plural form. e.g. यी मान्छे

iii) When the noun preceded by a numeral, then the noun usually remains singular. For e.g. पांच साल

4.2.2 Adjectives and Lexical Morphology: The main rule of Nepali adjectives is to qualify a noun or noun phrase. In Nepali language, adjective that ends in-o/ाो inflects for number, and gender giving different form of adjective root. Inflections of Nepali adjectives are illustrated in the table 4.3 below

Singular number Masculine	Feminine	Plural number Masculine/Feminine
राम्रो	राम्रो	राम्रा
बाठो	बाठी	बाठा
लाटो	लाठी	लाटा
कालो	काली	काला
मोटो	मीठो	मोटा

Table 4.3 Inflections of adjectives

Many words in Napali language are borrowed from Hindi and Sanskrit languages and these words don't show inflection so they form a category of uninfected adjectival forms. These words show the same distribution and functions as adjectives, e.g.

अरसलकेटो(Aasalketo)-- 'good boy'

अरसलकेटाहर(Aasalktaharu)--- 'good boys'

अरसलकेटी(Aasalketi)--- 'good girl'

अरसलकेटीहर(Aasalketiharu)---‘good girls'

Here अरसल/good takes same form for both singular and plural masculine and feminine forms.

4.2.3 Verbs and Lexical Morphology:

Verb is the most complex structure in Nepali grammar and verbs are quite inflected. It means that verbs have to agree with the subject’s condition, so in different situations, we use different forms of verbs. For example when we speak to a senior person, we use honorific speech and we use different verb form. Again we use different verb form, when we speak to a junior person.

Example, we say *तपाई भात खानुहुंचा[honour]*

त भात खान्चास[without respect]

There are three levels of honorific and the difference in gender is also marked in low grade honorific forms, all these measures make the verb's inflectional system fairly complicated.

i) The infinitive suffix – नु is used to mark simple or Compound verb e.g. खानु , लागनु.

ii) In Nepali language the verb usually appears at the end of the sentence.

e.g. किताब कहाँ छ ? Where is the book? (छ locates)

iii) Nepali verbs have special negative verb forms. The negative forms are

छ --→छैन
हो --→ होइन

4.2.4 Affixes

Affixes are bound morphemes which can only be attached to a root, stem and base word to form a new word. In an agglutinative language like Nepali, affixes play very important role in the formation of various words [65] and deriving several word classes. In Nepali language there are two sets of affixes, one for the prefix and another for the suffix [67].

For example, if a root word 'सुत्' is combined with the suffix 'एको' then the resulting form becomes 'सुतेको'. The suffix 'एको': 'eko' changes it's form to become 'ेेको' : 'eko'. Here 'ए' is a vowel and 'ेे' is a vowel symbol of 'ए'. So the set of suffixes will contain 'ेेको'. These affixes represent the category of bound morphemes and indicate certain syntactic category when combined with suitable free morphemes or roots. This is most common in Nepali language with verbs. However, not all bound morphemes essentially may be assigned the syntactic categories.

4.2.4.1 Prefix:

A prefix is an affix which is placed before the root of a word, mostly used to form word class. Prefixes are very limited in Nepali. There are few prefixes which have come into

Nepali from Sanskrit language and few are discussed below with words formation.

अति(ati) Prefix

अति(prefix)/ +क्रमण = अतिक्रमण(atikraman)

अति(prefix)/ +क्रम = अतिक्रम(atikram)

अति(prefix)/ +रक्ति = अतिरक्ति(atirakti)

अधि(adhi) Prefix

अधि(prefix)/ + करण = अधिकरण

अधि(prefix)/ +कम = अधिकम

अधि(prefix)/ +कारी = अधिकारी

अधि(prefix)/ +पति = अधिपति

उप(upa) prefix

उप(prefix)/ + देश = उपदेश

उप(prefix)/ + नाम = उपनाम

उप(prefix)/ + हार = उपहार

उप(prefix)/ + चार = उपचार

4.2.4.2Suffix:

A suffix is an affix which is placed after the root of a word. In Nepali, suffix plays an important role to form new word class. So various word forms can be constructed by suffixation of respective markers and various suffixes can be added one after another, by which meaning is also added. There are numerous suffixes used in Nepali [65].

करण(karan) Suffix

अनु +करन(suffix) = अनुकरन(anukaran)

उप +करन(suffix) = उपकरन(upakaran)

अधि +करण(suffix) = अधिकरण(adhikaran)

गति(gati) Suffix

प्र +गति(suffix) = प्रगति(pragati)

स +गति(suffix) = सगति(sagati)

दु +र्गति (suffix) = दुर्गति(durgati)

Special Characteristics of Nepali Prefixes and Suffixes: Here, we list down the special characteristics of Nepali Prefixes and Suffixes:

a) The suffix चाहिं may be added to nouns, pronouns and adjectives. When चाहिं is added to adjectives, it turns them into nouns, and may usually be rendered into English as 'the ... one' ठुलोचाहिं 'the big one'

मेरोचाहिं 'mine, my one'

In the same manner चाहिं may be added to the demonstrative and pronominal adjectives, and to a possessive formed with the postposition को

योचाहिं	this one
कुनचाहिं	which one?
राजकोचाहिं	Raj's one.

.

c) Masculine words (**Caste, Profession, Title** etc) are changed into feminine words by the addition suffix नी' (*nii*)

मालिक (*maalik*) = = मालिक् (*maalik*-) + नी (*nii*) --→ मालिक्नी

Some of the examples of prefixes and suffixes being used in Nepali words are given below:

(अ)	सहमित	-अ	prefix सहमित head word
(अ)	समान	-अ	prefix समान head word
खा(नु)		-नु	suffix खा head word
गर(ंनु)		-नु	suffix गर् head word
प(नु)		-नु	suffix प head word

धु(नु) -नु suffix धु head word

While it is evident from the above examples to some extent, the regular pattern of word formation suggest that generally the nouns take the prefix अ to form compound words and that verbs take the suffix नु thus resulting into new words. Such word formation pattern may be exploited for creating effective affix rules. The creation of affix rules to be later associated with head words substantially will free oneself from exhaustively including all the possible words of the language in the lexicon or the dictionary for spell checking purpose.

4.3 Nepali Tagset Review

Tagset development forms a foundation of any computational processing endeavour. The first pre-requisite for automated POS tagging is a tagset that is a set of exhaustive categories into which any token of the language can be placed. While the nature of the language is that there will always be words that are hard to classify, or are ambiguous between two categories [66], the tagset categories should be designed in such a way so as to minimize such problems. The fundamental problems in POS tagging task stem from the fact that a word can take different lexical categories depending on its context. The tagger has to resolve this ambiguity and determine the best sequence for a sentence. Tags are also applied to punctuation markers, thus tagging for natural language is the same process as tokenization for computer languages, although tags for natural languages are much more ambiguous.

NELRALAC tagset is the first tagset developed in Nepali language which consists of 112 tags. The error rates of annotation could be much higher when the size of the tagset is large and due to large size of tagset, it was not successful. But later, a smaller sized POS Tagset was developed which consists of just 43 tags and the design of this Nepali POS Tagset was inspired by the PENN Treebank POS Tagset [41].

4.3 Description of Nepali Tagset

For designing a Nepali tagset, apart from following the Eagles Guidelines and the Penn tree bank tagset, many other Indian tagging guidelines like IL-POST, ILMT and Sanskrit tagset were taken into consideration. After careful consideration a hierarchical tagset was favoured. The whole design of the tagset developed so far revolves around three distinct features into which the grammatical schema is distributed. The features are Category, Type and Attribute.

.The tagset for Nepali currently includes 43 tags and covers almost all the grammatical categories in the Nepali language. By the reference of Penn Treebank[61] tagset, the tagset of the Nepali is designed and it also based on BIS(Bureau of Indian Standards) framework. The short description of tag set used here is given follow:

Category	POS Tag ID No	POS Name	POS Tag	Examples
Noun	1	Common Noun	NN	तिया,

				कीटो,केटा,कलम
	2	Proper Noun	NNP	राजू
Pronoun	3	Personal Pronoun	PP	म, हामी
	4	Possessive Pronoun	PP$	मेरो, हाम्रो
	5	Reflexive Pronoun	PPR	आफू
	6	Marked Demonstrative	DM	अर्को
	7	Unmarked Demonstrative	DUM	त्यो
Verb	8	Finite Verb	VBF	खायो
	9	Auxiliary Verb	VBX	थियो
	10	Verb Infinitive	VBI	खान, गर्नु, गर्न,नगर्नु
	11	Prospective Participle	VBNE	हिदने मान्छे
	12	Aspectual Participle	VBKO	थियो
	13	Other Participle Verb	VBO	दिएको
Adjective	14	Normal/Unmarked	JJ	असल

	15	Marked Adjective	JJM	राम्रो
	16	Degree Adjective	JJD	अधिकार
Adverb	17	Manner Adverb	RBM	दिलो हिद्छा
	18	Other Adverb	RBO	यहाँ बसा
Intensifier	19	Intensifier	INTF	धेरै चलाख
Postpositions	20	Le-Postposition	PLE	हरिले
	21	Lai-Postposition	PLAI	भिलाई
	22	Ko-Postposition	PKO	रामको
	23	Other Postpositions	POP	ताबुल्माथि
Conjunction	24	Coordinating	CC	रा
	25	Subordinating Conjunction	CS	किनभने
Interjection	26	Interjection	UH	ओहो
Number	27	Cardinal Number	CD	एक
	28	Ordinal Number	OD	पहिलो
Plural Marker	29	Plural Marker हरु	HRU	हरु
Question Word	30	Question Word	QW	को

Classifier	31	Classifier	CL	दस्जना
Particle	32	Particle	RP	खाई
Determiner	33	Determiner	DT	त्यो कीटो
Unknown Word	34	Unknown Word	UNW	नेकोम्प्रेनस
Foreign Word	35	Foreign Word	FW	गुड
Punctuation	36	sentence Final	YF	? !
	37	sentence Medieval	YM	, : ;
	38	Quotation	YQ	‘ ‘ “ “
	39	Brackets	YB	( ) { |
Header List	40	Header List	ALPH	का
Symbol	41	Symbol	SYM	%
Abbreviation	42	Abbreviation	FB	म.पु.पु.

Table 4.4: List of Part-of-Speech tags for Nepali Language

4.3.1 Noun:

Nouns in Nepali are broadly classified as Proper and Common Nouns

Proper Noun (NNP): A proper noun is a noun that in its primary application refers to a unique entity, such as London, Jupiter, Sarah, or Microsoft, as

distinguished from a common noun, which usually refers to a class of entities . NNP tag is given for proper noun.

. **Common Noun** (NN): A common noun is a word that names people, places, things, or ideas. They are not the names of a single person, place or thing. NN tag is assigned to denote common nouns.

4.3.2 Pronoun

All Pronouns are tagged separately. A pronoun is a word takes the place of a noun. In Nepali Pronouns are inflected for number, gender and case like Nouns but are distinguished from Nouns by having a category of person also. Thus all Pronouns will be tagged with their gender, number and person features

Personal Pronoun (PP): All first and second person pronouns come under personal pronoun and PP tag is assigned to this category.

.

Possessive Pronoun (PP$): A pronoun is a word that replaces a noun in a sentence, making the subject a person or a thing. Possessive pronouns are pronouns that demonstrate ownerships. For example, my pen, my father.

Reflexive Pronoun (**PPR):** A reflexive pronoun is a pronoun that is preceded or followed by the noun, adjective, adverb or pronoun to which it refers within the same clause.

Category	POS Name	POS tag	Example
	Personal Pronoun	PP	म, हामी

Pronoun	Possessive Pronoun	PP$	मेरो, हाम्रो
	Reflexive Pronoun	PPR	आफू
	Marked Demonstrative	DM	अर्को
	Unmarked Demonstrative	DUM	त्यो

Table 4.5: Classification of Pronoun with example

4.3.3 Demonstrative:

Demonstratives are words like 'this' and 'that', used to indicate which entities are being referred to and to distinguish those entities from others. They are typically deictic, their meaning depending on a particular frame of reference.

Marked Demonstrative (DM**):** Marked Demonstratives include those which inflect for gender and number.

Unmarked Demonstrative (DUM): Unmarked Demonstratives cover those demonstratives that do not inflect for gender and number. However, they vary in their forms if they are followed by postposition.

4.3.4 Verb:

Verbs in Nepali are quite highly inflected, agreeing with the subject in number, gender, status and person. They also inflect for tense, mood, and aspect. Apart from these inflected finite forms, there are also a large number of participial forms.

Possibly the most important verb in Nepali, as well as the most irregular, is the verb हुनु hunu 'to be, to become'.

Finite verb (VBF): Finite Verbs include verbs with tense inflections. By implication, it assumes imperatives and optatives as well. VBF is assigned for this group.

.’

Reflexive pronoun (VBX): A reflexive pronoun is a pronoun that is preceded or followed by the noun, adjective, adverb or pronoun to which it refers (its antecedent) within the same clause. There are two auxiliary verbs in Nepali –छand हो . These and their corresponding past forms receive separate tag, VBX.

Verb infinitive (VBI): The infinitive form of a verb is the verb in its basic form. Verbs that end with न,नु and ना that basically refer to the infinitiveness are grouped in this class. VBI is assigned for this class.

Participle verb: Participles share behavioural properties with other lexical categories such as adjectives and adverbs. Besides, they also come as a member in a verbal chain.

Prospective participle verb (**VBNE**): This category of verbs that ends in ने modifies nouns.

Aspectual participle verb (**VBKO):** The non-finite forms of the verbs that end with को/एको are considered as aspectual participles. In Nepali, they may appear as a member in a verbal chain and may play the adjectival role.

Other participle verb(VBO): The non-finite form of verbs that end with एर/ई/ईकन and दो/दा/द /दै are grouped in this class. Basically, these verbs characterize manner.

Category	POS Name	POS tag	Example
Verb	Finite Verb	VBF	खायो
	Auxiliary Verb	VBX	थियो
	Verb Infinitive	VBI	खान, गर्नु, गर्न,नगर्नु
	Prospective Participle	VBNE	हिदने मान्छे
	Aspectual Participle	VBKO	थियो

Table 4.6: Classification of Verb with example

4.3.5 Adjective:

Adjectives may be divided into declinable and indeclinable categories. Declinable are marked, through termination, for the gender and number of the nouns they qualify. The declinable endings are -o for the "masculine" singular, -ī for the feminine singular, and -ā for the plural. e.g. sānokitāb "small book", sānī eṭī "small girl", sānākalamharū "small pens"

Unmarked Adjective (JJ): The modified word does not inflect for gender and number in unmarked adjectives and it is assigned by the tag JJ

Marked Adjective (JJM**)**: Any Adjective which are ends with ओ,ई and आ come under this group and JJM tag is assigned for it.

Degree Adjective (**JJD**): Generally adjectives that end with तर and तम are come under this group and JJD tag is assigned for it.

4.3.6 Adverb:

Nepali adverbs are parts of speech. Generally they're words that modify any part of language other than a noun. Adverbs can modify verbs, adjectives (including numbers), clauses, sentences and other adverbs.

today -→aaj - **आज**

now ahile - **अहिले**

tonight aajaraati - **आज**

Manner (RBM): Manner of the verbs is described by this type of adverb and RBM tag is used to assign to them.

Other Adverb (RBO**)**: Location and time come under this category and RBO tag is assigned to them.

4.3.7 Intensifier (INTF)

Intensifier is a linguistic term for a modifier that makes no contribution to the meaning of propositional a clause but serves to enhance and give additional emotional context to the word it modifies. INTF tag is assigned for Intensifier.

.

4.3.8 Postposition:

In English language, prepositions are often placed before the words. But in Nepali, they may come after adjective and verbs. There are lots of postpositions in Nepali; some of them are as follow:

a) -baaTâ /dekhi from

b) -sângâ with

c) -tirâ towards/to

Le-Postposition (**PLE):** The marker ले is often used for this type of category and the tag PLE is used to assign this type of category.

. **Lai-Postposition** (**PLAI):** The accusative/dative marker that refers to the patient/recipient function is identified as Lai-Postposition. PLAI tag is assigned forth is class.

Ko-Postposition (PKO): These types of postpositions are used for gender marker and display the gender, number inflections.PKO tag is assigned for this class.

Other Postpositions (POP): All other postpositions including dependent postpositions like देख ,बाट , मा,सँग,सम are grouped into this class.

Category	POS Name	Tag Name	Example
Postpositions	Le-Postposition	PLE	हरिले
	Lai-Postposition	PLAI	भिलाई
	Ko-Postposition	PKO	रामको
	Other Postpositions	POP	ताबुल्माथि

Table 4.7: Classification of Postpositions with example

4.3.9 Conjunction

Conjunctions are used to join two sentences. Conjunctions are called 'संयोजक'(*Sanyojak*) in Nepali.Conjunction are classified into two categories: Coordinating Conjunction, Subordinate Conjunction.

अनि (ani)

Coordinating Conjunction (**CC):** The coordinating conjunctions are used to make connection. It is used for coordination between two clauses of equal

scope .This type of class is assigned by CC tag.

Subordinate Conjunction (CS): Subordinate Conjunctions are the conjunctions which are used to establishing a relationship between the similar clauses.

4.3.10 Interjection (UH):

An interjection is used to express an emotion or sentiment on the part of the speaker and assigned by the tag UH. Interjections are generally placed at the beginning of a sentence.

4.3.11 Number: Numbers are two types**:** cardinals and ordinals and they are tagged as CD and OD respectively.

Cardinal Number (CD)A Cardinal Number is assigned by the tag CD It is used for number, such as one, two, three, four, and five.

Ordinal Number (OD): Ordinal number is used to show the position of someone or something which is in a series. It is assigned by the tag OD.

4.3.12 Plural Marker (HRU): Plural Marker generally is a single word tag and it acts like a plural and collective marker.

4.3.13 Question word (QW):

All question words like 'को' are tagged as 'QW' in the present tagset.

4.3.14 Classifier (CL):

In Nepali language Classifiers are used for classify nouns. Classifiers are used for indicating the amount of the noun in question.CL tag represents the classifier.

4.3.15 Particle (RP):

Particles in Nepali are like prepositions in English but they come after the word. Therefore, they modify the word they succeed. RP tag will be used for them.

4.3.16 Determiner (DT):

A determiner is a word, phrase or affix that occurs together with a noun or noun phrase and serves to express the reference of that noun or noun phrase in the context. DT tag is assigned for Determiner.

4.3.17 Unknown (UNW):

Unknown words are those which are not known or identified in the process of tagging. UNW tag is assigned for unknown words.

4.3.18 Foreign Word (FW)

Foreign Word is used for tagging foreign words which are written in the Devanagari script or any other foreign scripts. FW is used for assigning foreign words.

4.3.19 Punctuation

Punctuation marks are symbols and common punctuation marks are the question mark, comma, exclamation point, apostrophe, quotation mark and hyphen. Four types of punctuations are used for our tagset and they are sentence final, sentence medieval, quotation and brackets. They are tagged as YF, YM, YQ and YB respectively.

Sentence final (YF): The YF tag is assigned for sentence final like ? । . ! ।॥

Sentence medial (YM): The YM tag is assigned for sentence medial like **, ; : :- / - :**

Quotation (YQ): The YB tag is assigned for quotation **marks** like ‘ ’ “ ” ' ' " ".

Brackets (YB): The YB tag is assigned for Square Brackets, Curly Brackets, angle Brackets.

e.g () {} []

Category	POS Name	POS tag	Example
Punctuation	sentence Final	**YF**	**? !**
	sentence Medieval	**YM**	**, : ;**
	Quotation	**YQ**	**‘ ‘ “ “**
	Brackets	**YB**	**( ) { |**

Table 4.8: Classification of Punctuation with example

4.3.20 Abbreviation (FB): In this category, FB is used for assigning the tag of abbreviation.

4.3.21 Header List (ALPH):

A character like alphanumeric symbols, brackets, comma etc., bullets, and stars comes under this category. ALPH is used to tag this group.

4.3.22 Symbol (SYM):

All special characters like #, @, &, $, %, etc. are tagged as ‘SYM’. This tag is similar to the Penn tag set tag ‘SYM’.

4.3.23 Null<Null>:

In this category an element of the text does not required any tag.