Some Aspects of the Ambiguities of Bengali Non-finite Verb Forms

Niladri Sekhar Dash, Linguistic Research Unit, Indian Statistical Institute, Kolkata, India

An empirical analysis of orthographic forms and lexicosyntactic functions of non-finite verbs (NFVs) in Bengali reveals two interesting things: (a) orthographic forms create confusion in the analysis of their lexical identities and assignment of their part-of-speech (POS) values, and (b) lexicosyntactic functions generate ambiguities in the decipherment of their appropriate semantic roles in text. These two issues are primarily taken into analysis and investigation in this paper. The present work briefly reports on the following issues: compilation of a large lexical database of NFVs from modern Bengali text corpora; classification of NFVs based on their orthographic forms; division of NFVs into roots and suffixes to understand their formation; analysis of morpho-syntactic roles of suffixes in NFV formation; identification of morphophonemic concatenation patterns when suffixes join with verb roots to generate NFVs; understanding the patterns of grammatical mapping between roots and suffix in generation of valid NFVs; identification of semantic information embedded in suffixes; and annotation of their lexicosyntactic roles with information obtained from the analysis. The NFVs are first divided into different sub-groups according to their structures (e.g., roots ending in a consonant, and ending in -ā, -i, -e, -u, -o), while their suffixes are sub-grouped based on their nature of conjoining with roots. For annotation, verb roots are suggested to be stored in a root database, while suffixes are proposed to be stored in a suffix database, and a matching algorithm is proposed to concatenate appropriate root-suffix pairs. To increase robustness and accuracy in annotation, some grammatical mapping rules are proposed to generate possible root-suffix combinations. The findings of this study can make a valuable contribution to Bengali language description, language teaching, morphological processing, text annotation, machine learning, and dictionary compilation.

Keywords: *roots, non-finite verbs, suffixes, grammatical mapping, annotation*

1 Introduction

This paper is based on an analysis of a large number of Bengali non-finite verbs (NFVs) that we collect from a Bengali written text corpus. Our goals are to structurally analyze these verbs to collect data and information that may contribute towards understanding their form and role in the language. It can also provide the necessary information and insight for developing a method for their part-of-speech annotation, sense disambiguation, lexical analysis, language teaching, dictionary compilation, and language description.

The NFVs have a variety of functions in Bengali some of which are understood when the sense of a NFV is linked with that of a finite verb (FV). They add liveliness, tension, mood and aspectual features in the text where they are used (Thompson 2012: 78). The use of NFVs in Bengali speech and writing is very high. They are often used immediately before a finite verb in a sentence although there are instances of deviations. Generally, an NFV has no independent 'action role' in a sentence; because, in its independent identity, it always fails to indicate any complete sense of action. It has to depend on necessary information provided by a finite verb. The main functions of a NFV in Bengali are, therefore, not confined to denoting a sense of incompleteness of an action but also a sense of continuation and repetition of an action.

The whole idea of classifying Bengali verbs into NFV and FV is primarily based on the semantic function they play in texts. The term 'non-finite' is ambiguous in the sense that it refers to both the semantic and syntactic functions of a verb when it is used in a sentence. In practice, the use of an NFV in a sentence neither gives a complete sense of the meaning of an action denoted in the sentence nor does it provide a full grammatical structure to a sentence. For example, if we look at the sentence given below (1), we find that the sentence is complete in form and meaning and the presence of an NFV in the middle of the sentence, plays a supporting role in the completion of the sense of the verb.

(1) *āj bikele āmi kājṭā śeṣ kare ýābo*
Today_afternoon_I_ the-work_ finish_ doing_ shall-go
I shall go after finishing the work in the afternoon today.

The above example (1) shows that it carries an NFV (i.e., *kare*), which fails to denote a sense of completion of an action. It needs the support of a finite verb to complete its grammatical and semantic functions. This signifies that although NFVs and finite verbs are interdependent parts of the verbal system of Bengali, they work together to convey meaning collectively. What is unique here is that the orthographic form of the NFV (i.e., *kare*) is ambiguous. Due to this factor, in a context-free situation, it can be identified as a finite verb, NFV, noun, adjective, or a postposition and discourse element if we do not consider the information of context of its use. Thus the word *kare* can be identified in five different parts of speech, although we know that there are very few Bengali verbs which carry so many possible pars-of-speech values. As a noun, it is used in the sense of "tax" tagged with a locative case marker '-e' (e.g., *kare (< kar +-e*) "in tax". A typical use of this form is found in Bengali newspapers: *Sarkār ebāre kare beś kichhutā chhāṛ diyechhe* "the government has given some rebate in tax this time".

In this paper, we report that we compile a large number of NFVs from a large written Bengali text corpus. We classify NFVs based on their orthographic forms, split them into root and suffix, analyze their morphosyntactic roles, identify morphophonemic conjoining rules when a suffix is added to a root, define the grammatical mapping of root and suffix in the generation of final forms, gather semantic information from a suffix, and propose methods for annotation of NFVs based on information collected from analysis. We argue that Bengali learners need a clear understanding of the forms and functions of NFVs to enrich their knowledge of Bengali and this study can be quite important and useful for them. The information and data that we present here can also be used in machine learning, morphological processing, text annotation, and dictionary compilation.

In Section 2, we present a short description on grammatical analysis of Bengali NFVs; in Section 3, we briefly discuss problems and challenges involved in identification of NFVs in written Bengali texts; in Section 4, we refer to some examples to explain the nature of sense variations of NFVs and their use in different parts-of-speech; in Section 5, we present a brief account on contextual use of NFVs as found in Bengali texts; in Section 6, we present some simple statistical counts on percentage of use of NFVs in modern Bengali; in Section 7, we focus on the nature of ambiguity in NFVs to understand how ambiguity is generated by them; in Section 8, we present a brief analysis on deformation of verb roots while using NFV suffix; in Section 9, we divide NFVs into root and suffix parts to define the patterns of grammatical mapping between root and suffix; in Section 10, we propose a method to annotate NFVs with information gathered from their structural and functional analysis; and in

Section 11 (conclusion), we highlight the importance of this study in language description, teaching, dictionary development and language technology.

2 Grammatical Analysis of Bengali Non-Finite Verbs

The morphological analyses of modern Bengali NFVs that are available to date, are primarily based on grammatical-cum-semantic interpretation of NFVs in the language (Chatterji 1926 (1993), Chattopadhyay 1995, Sen 1993, Chaki 1996, Chakrabarti 1985, Majumdar 1993, Sarkar and Basu 1994, Bhattacharja 1998, Shahidullah 2003, Thompson 2010, Thompson 2012). Suniti Kumar Chattopadhyay (1995: 297-298) argues that Bengali verb roots are not tagged with *bibhaktis* (i.e., case markers) but with *pratayas* (i.e., suffixes). Because of this reason, Bengali NFVs should be considered as 'non-flexional verbs' in the sense that they cannot be inflected (i.e., these are structurally non-flexional) although they are semantically infinite. Muhammad Shahidullah (2003: 79-83), on the other hand, classifies NFVs based on their surface forms and semantic functions in texts. According to him, NFVs are formed when suffixes *-iyā, -ite,* and *-ile* are added to verb roots. He argues that the use of NFVs in Bengali texts is controlled by various semantic features (*e.g., sequence, purpose, goal, continuity, potentiality, propriety, contemporaneity (quality of belonging to the same period of time), necessity, desire, order*) than by other properties. To understand these semantic functions, we have to interpret these NFVs with proper reference to co-words occurring in different sentential contexts. On the other hand, after considering the etymology, meaning, and orthography of verb-ending suffixes, Sukumar Sen identifies four types of NFVs in Bengali (Sen 1993: 251-253).

(a) Conjunctive (*Lyabartha bā pūrbakālik asamāpikā*): produced by adding NFV suffixes *-i* and *iyā* with a verb root, e.g., *kari* "doing", *baliyā* "saying".
(b) Conditional (*Bhūtartha bā ýadyartha asamāpikā*): produced by adding NFV suffix *-ile* with a verb root, e.g., *karile* "doing", *balile* "saying".
(c) Gerundial (*Śatrartha bā bartamān asamāpikā*): produced by adding NFV suffix *-nte* with a verb root, e.g., *chalante* "going", *phalante* "resulting".
(d) Infinitive (*Tumartha bā uddeśak asamāpikā*): produced by adding NFV suffix *-ite* with a verb root, e.g., *dharite* "to catch", *marite* "to die".

This classification is, however, questioned by Shishir Bhattacharja on the ground that these suffixes are not available in modern Bengali texts (Bhattacharja 1998: 229-233). Although the argument of Bhattacharja is not wholly true, there is some truth in it. Our observation in this context is that some NFV suffixes, as stated in Sukumar Sen (1993), are indeed available in modern Bengali texts, although in a modified form. We agree with Bhattacharja's claim that the examples that are presented by Sen mostly belong to old and middle Bengali texts. It was necessary to show examples of modern Bengali NFVs mentioned in his list. Hanne-Ruth Thompson (2012: 77) argues that "Every verb has four non-finite forms" which seems to be true to a large extent. She classifies NFVs in Bengali into four types, namely the following (Thompson 2012: 77):

(a) Verbal Noun: Made by adding marker *-ā* to verb roots having CVC/VC structure, e.g., *lekh-* "write", *dekh-* "see", *kar-* "do", *bal-* "say", *oṭh-* "rise", *ān-* "bring", *ās-* "come";

adding *-oyā* to verb roots having CV, e.g., *ha-* "be", *de-* "give", *ne-* "take"; and adding *-no* to verb roots with CaCa/CVCA/VCa, e.g., *chālā-* "drive", *ghumā-* "sleep", *oṭhā-* "lift"; and adding *-no* to verb roots with VCo/CVCo, e.g., *ego-* "advance", *bero-* "go out".

(b) Imperfective participle: The suffix *-te* is added with the high stem of the verb, e.g., *kar-* "to do", *dhar-* "to catch", *phel-* "to drop".

(c) Perfective Participle: The suffix *-e* is added to the high stem of the verb, e.g., *par-* "to wear", *śun-* "to hear", *gun-* "to count".

(d) Conditional participle: The suffix *-le* is added to the high stem of the verb, e.g., *dhar-* "to hold", *par-* "to wear", *khel-* "to play".

We, however, deviate from this scheme of analysis as our primary goal is to analyze them to understand their morphological structure and formation as well as retrieve information on their functional role when used in texts (Dash 2002). From the perspective of orthography, there are two types of verbs, namely, non-affixed and affixed (Fig 1).

(a) Non-affixed verb: It includes those verbs that are used in their root forms without suffixes, e.g., *kar* "you do", *dekh* "see", *bal* "say", *thāk* "stay", *śon* "listen", *thām* "stop".

(b) Affixed verb: It includes those verbs that are used with suffix, e.g., *karche* "doing", *dekhla* "saw", *balbe* "will say", *thākbe* "will stay", *śunchi* "hearing", *thāmla* "stopped".

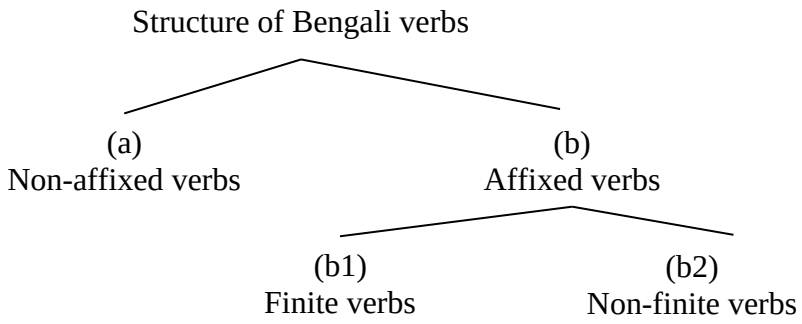

Figure 1: Orthographically two types of verbs in Bengali

The classification given above (Figure 1) is purely a structural one and the verbs in the first set (i.e., non-affixed verbs) are actually verb roots, which can be affixed based on the need of the contexts of their use. We, therefore, do not claim them to be a different type of verbs; rather they are the verbs, which are found in texts in their basic root forms. The second layer of the classification (Fig. 1) is more complex as it takes the structural and semantic aspects of verbs. Here verbs are divided into two types: (a) finite verbs (FVs) which are made with specific sets of suffixes and which denote a sense of completeness of an action, e.g., *balchhi* "I/we am/are saying", *balcchilām* "I/we was/were saying", *balba* "I/we shall say", *ballām* "I/we said", *baltām* "I/we used to say"; and (b) non-finite verbs (NFVs) which are made with another set of suffixes and which imply a sense of incompleteness of an action, e.g., *bale* "saying", *baliyā* "having said" *balle* "having said", *balte* "to say". In this paper, we are going

to focus on the NFVs only as our goal is to understand their ambiguous identity and to develop a strategy for their better interpretation description and application.

3 Identification of NFVs in Bengali Texts

The NFVs are identified in Bengali texts based on their two linguistic criteria: surface form and sense denotation. With regard to the first criterion, Bengali NFVs use specific suffixes based on which we can identify if a verb is an NFV. There are, however, some examples, where a NFV suffix is identical with that of a FV. This creates ambiguities for the NFVs at their structural and functional levels (discussed in Section 7). If we analyze the structure of NFV suffixes, we can divide them into two types based on their use in two different varieties of Bengali verbs: *sādhu* or chaste variety which is quite restricted in use in Bengali and *chalit* or colloquial variety which is quite frequent in use in the language.

(2) a. Suffixes used to produce *sādhu* or chaste NFVs:
- *-iyā* (*kariyā, dekhiyā, chiniyā*),
- *-ite* (*karite, dekhite, chinite*),
- *-ile* (*karile, dekhile, chinile*),
- *-āile* (*karāile, dekhāile, chināile*),
- *-āiyā* (*karāiyā, dekhāiyā, chināiyā*),
- *-āite* (*karāite, dekhāite, chināite*).

b. Suffixes used to produce *chalit* or colloquial NFVs:
- *-e* (*kare, dekhe, chine*),
- *-te* (*karte, dekhte, chinte*),
- *-iye* (*kariye, dekhiye, chiniye*),
- *-le* (*karle, dekhle, chinle*),
- *-ye* (*giye, peye, dhuye*),
- *-āte* (*karāte, dekhāte, chenāte*).

The percentage of use of *sādhu* (chaste) NFVs is much less (18%) than that of *chalit* (colloquial) NFVs (82%) in modern Bengali texts (Dash 2005:172). Among the *chalit* forms, the percentage of NFVs made with the suffix *-e* is the highest (34.3%) followed by that of *-te* (15.7%) and *-iye* (14.4%) (Table 1).

Table 1: *sādhu* and *chalit* NFV suffixes in Bengali

sādhu suffix	%	*chalit* suffix	%
-iyā	7.5	-e	34.3
-ite	5.2	-te	15.7
-ile	2.3	-iye	14.4
-āile	1.4	-le	9.6
-āiyā	1.0	-ye	5.8
-āite	0.6	-āte	2.2
Total	18	Total	82

We find that some NFV suffixes are identical in form with some suffixes used for finite verbs and nouns. This creates a problem in the identification of NFVs in texts as well as their frequency count, and lexicosemantic analysis. We had to manually check each form to determine if it was an NFV based on its form and grammatical role in a sentence. Another unique characteristic feature is that some of the NFV suffixes are quite robust and productive. They are not only used with NFVs and FVs but also with compound verb roots and nouns to generate 'denominalized verbs' (i.e., *verbs generated from nouns by way of adding verbal suffixes with noun stems*) of various types. For instance, NFV suffixes like *-iye, -āiyā, -āile, -āle, -āite, -āte* are tagged with nouns to generate new NFVs (Table 2). This phenomenon leads us to have a closer look into the structure of NFVs to understand the roles of suffixes in the formation of these words.

Table 2: Denominalized verbs made from nouns by using NFV suffix

Noun	NFV Suffix		Final NFV form	Gloss
hāt	-iye	→	hātiye\NFV\	having stolen
hāt	-āiyā	→	hātāiyā\NFV\	stealing
hāt	-āile	→	hātāile\NFV\	having stolen
hāt	-āle	→	hātāle\NFV\	having stolen
hāt	-āite	→	hātāite\NFV\	to steal
hāt	-āte	→	hātāte\NFV\	to steal
mukh	-iye	→	mukhiye\NFV\	being eager
ghum	-iye	→	ghumiye\NFV\	having slept
ghām	-iye	→	ghāmiye\NFV\	having sweated

4 NFVs in Modern Bengali Text Corpus

To understand the nature of the use of NFVs in Bengali, we collected a large number of NFVs from a modern Bengali text corpus of around a hundred thousand sentences. We collected them through the application of a concordance programme on the Bengali corpus. We analyzed them with reference to the contexts of their use in sentences to interpret their structure and role in the language. In this section, we summarize the information that we obtained from the analysis of their surface forms, functional roles, and contextualized senses to show how they are used and understood in modern Bengali.

4.1 *Use of NFVs in Different Part-of-Speech*

This is a new finding that shows that NFVs are used in other parts-of-speech in Bengali texts. It happens when words, which are often used as NFVs, are also used in other parts-of-speech. It is possible due to ambiguities generated from the contexts of their use in different sentential environments.

- NFV as discourse element: *āmi tār nām dhare ḍāki* "I call him by his name".
- NFV as conjunctive: *tumi base theke kājṭā kariye nāo* "Sitting here, you get the work done".
- NFV as postposition: *śaharer theke grām bhālo* "Village is better than town".
- NFV as infinitive: *lokṭāke tumi dekhte pele nā* "You did not get to see the man".
- NFV as Adjective: *eman khaṭiye mānuṣ āge dekhini* "I have never seen such a workaholic man".

The role of context in the use of NFVs is discussed in Section 5, while the issue of ambiguity is addressed in Section 7.

4.2 *Sense Variation of Bengali NFVs*

While studying their sense denotation, we note that Bengali NFVs indicate 14 different types of senses such as the following.

- Incompleteness: *sekhāne giye khabarṭā deba* "After reaching, I shall give the news".
- Continuity: *dekhte dekhte anek din keṭe gela* "Many days passed by in the meantime".
- Variance: *dekhe śune bichār karte habe* "To be judged after considering all aspects".
- Recurrence: *bale bale āmi klānta* "I am tired of saying".
- Condition: *tumi ḍākle āmi āsba* "I shall come if you invite."
- Sequence: *Se hese balla se kathā* "He said that while laughing."
- Result: *Beśi kheye se base paṛechhe* "He sits down after eating more".
- Propriety: *mā bābār kathā mene chalte hay* "One has to obey words of parents".
- Contemporarity: *tumi ḍākte se sāṛā diyeche* "He responded after you called him".
- Necessity: *upāy nei, āmāke ýete habe āj*. "No way, I have to go today".
- Desire: *tomār hāter rānnā khete chāi ekbār* "Want to taste your cooking once".
- Instruction: *tumi ekhan āste pāro* "You may come now".
- Uncertainty: *khelāṭā āj nāo hate pare* "The play may not happen today".
- Doubt: *jānā thākle tomāy bale dite pārtām* "Could have told you if I knew it".

The senses that are classified and cited above include different senses of NFVs that we note in modern Bengali texts. Some of the senses are not found in earlier works (Majumdar 1993, Sarkar and Basu 1994, Sen 1993), while some other senses are differently interpreted (Thompson 2012). However, it should be stated here that this list is not exhaustive and the number of senses may increase if more examples are interpreted.

5 Contextualized Use of NFVs

The contexts of the use of NFVs, as noted in Bengali texts, supply various kinds of information about their occurrence in contextualized frames. Some of our findings contradict observations made by earlier scholars as their observations are mostly based on intuitions than on examples of actual use. From the corpus, we identify nine different types of contexts of use of NFVs, as the following examples show.

- NN-NFV-NN: āgun lege\NFV\ ghar\NN\ puṛechhe\FV\ “The house is burnt with fire”.
- NN-NFV-FV: tār hāte hāt\NN\ rekhe\NFV\ balla\FV\ “He said it keeping hands in hand”.
- NFV-NFV-FV: āmāy path dekhiye\NFV\ niye\NFV\ chala\FV\ “Guide me the path”.
- NFV-NFV-NN: base\NFV\ base\NFV\ dintā chale gela “The day passed by sitting”.
- PP-NFV-FV: dokān theke\PP\ kinte\NFV\ habe\FV\ “Have to be bought from a shop”.
- ADJ-NFV-FV: ākāśṭā kālo\ADJ\ haye\NFV\ āschhe\FV\ “The sky is becoming dark”.
- ADV-NFV-FV: tār galā spaṣṭa\ADV\ śunte\NFV\ pelām\FV\ “I clearly heard his voice”.
- FV-NFV-NFV-FV: se āschhe\FV\ jene\NFV\ theke\NFV\ gelām\FV\ “I stayed back knowing that he is coming”.
- FV-NFV-NFV-NFV-FV: se kāl āsbe\FV\ jānte\NFV\ pere\NFV\ theke\NFV\ gelām\FV\ “I stayed back knowing that he will come tomorrow.”

The examples given above exhibit different contextual frames of occurrence of NFVs in written Bengali texts, which may differ from spoken texts. Generally, it is stated in standard Bengali grammar that NFVs occur immediately before finite verbs, which is also supported by many early studies. We, however, find many new patterns which are not attested in early works. In fact, before this corpus-based study, we had no idea that a finite verb can occur immediately before an NFV or a noun can occur immediately after an NFV. It helps us to know different contexts of the use of NFVs in sentences, which can be used to describe the modern Bengali language as well as to teach it to the learners.

6. Frequency of Use of NFVs

Nearly twenty years ago, an attempt was made to know the frequency of use of NFVs in written Bengali texts. From the analysis of a written Bengali text corpus, it is claimed that one of every three sentences carries at least one NFV along with a finite verb (Dash and Chaudhuri 2000). With regard to frequency of use, it is reported if only the finite verbs and NFVs are taken into comparison, then finite verbs (59%) can record a higher percentage of use than NFVs (41%) in Bengali (Dash 2005: 225-226). The use of NFVs is further reduced when they are compared with words of other parts-of-speech (Table 3).

Table 3: Use of words of different part-of-speech in Bengali

Part-of-Speech	%-age
Nouns	33.23
Pronouns	5.23
Demonstratives	2.84
Finite Verbs	18.33
Non-Finite Verbs	5.05
Adjectives	7.09
Adverbs	5.27
Postpositions	2.75
Conjunctions	3.35
Particles	2.66
Quantifiers	3.07
Residuals	10.03
Others	1.10
Total	100

On the other hand, when we calculate the use of NFVs within different sub-types of verbs in Bengali (Table 4), we find that NFVs show the second highest percentage (28.92%) of use preceded by affixed finite verbs (47.97%). It is an interesting feature to note that NFV suffixes, including both *sādhu* (chaste) and *chalit* (colloquial) forms, are only a few in number, but their use in the language is quite frequent and productive, due to which it records a high percentage of use in the language.

Table 4: Use of different verb sub-types in Bengali

Verb sub-types	%
Verb_Non-affixed	0.35
Verb_Finite_Affixed	47.97
Verb_Finite_Negative	1.03
Verb_Non-finite	28.92
Verb_Infinitive	5.11
Verb_Gerund	15.66
Verb_Auxiliary	0.96
Total	100

Among NFV suffixes, *-e* records the highest use followed *-te, -iye, -le, -iyā, -ite, -ye, -āte*, and *-ile*. Among these, *-ite, -ile* and *-iyā* are used for *sādhu* (chaste) NFVs, while others are used for *chalit* (colloquial) NFVs (Table 1). It also shows that the use of *sādhu* (chaste) NFVs is much less (18%) than *chalit* (colloquial) NFVs (82%) in the language, which indirectly hints at the gradual loss of use of *sādhu* (chaste) forms in the language. The orthographic similarities between NFVs and finite verbs can create confusion in the proper identification of NFVs. To overcome this problem, during frequency count, each NFV is manually checked to confirm its lexical identity.

7 Ambiguity in Bengali NFVs

Lexical ambiguity is an important aspect of a natural language. Words, in both context-free and context-bound situations, can convey multiple senses, items, and ideas to generate various possible information. This allows us to derive several readings, which differ based on lexical features, lexical sub-categorization, selection features, syntactic aspects, semantic properties, idiomatic readings, discourse function, and others (Sinclair 1991: 105). Many scholars have discussed this issue in detail to understand the nature of ambiguity and semantic flexibility in word meaning (Ullmann 1962, Leech 1974, Yule 1985, Cruse 1986, Todd 1987, Palmer 1995, Pustejvsky 1995, Boguraev and Pustejvsky 1996Kreidler 1998, Cruse 2000, Ravin and Leacock 2000, Bouilon and Busa 2001). In most cases, it is noted that it is related to *polysemy* where words carry multiple senses (Lascarides and Copestake 1998) and *homonymy* where unrelated meanings share the same surface representation of words (Fellbaum, 2000, Ravin and Leacock 2000). The Bengali NFVs are no exception. They also carry ambiguity and semantic flexibility. However, their ambiguity falls under 'structural ambiguity', where ambiguity is caused due to similarity in surface forms of suffixes. This kind of ambiguity is mostly noted in the case of suffixed verbs where multiplicity of sense is caused due to the use of similar suffixes. This phenomenon is not confined to single verbs only; it is also found in compound and reduplicated verbs, which carry similar suffixes.

In the case of structural ambiguity, the condition is that the affixed verbs, in spite of belonging to different parts-of-speech, look identical in surface form, both in root and suffix parts. Since this kind of use is common in Bengali words, it is a serious problem in the identification of NFVs and their senses. We find that most of the NFV suffixes (e.g., *-e, -ite, -te, -ile, -le, -iye*) are identical in form to that of finite verbs, nouns, and adjectives. Therefore, in a context-free situation, the addition of a suffix with a verb root, a noun, or an adjective creates confusion. We can consider the final form of the word as a finite verb, an NFV, an inflected noun, or an affixed adjective. For instance, the word *khāṭ* can be tagged with identical suffix to generate four sets of identical surface forms, which are different in parts-of-speech and sense, as the following four sets show.

Set 1: Outputs as Finite Verbs

No	Root	Suffix	Final	POS	Gloss	Type
1	khāṭ	-e	khāṭe	FV	Works	A1
2	khāṭ	-ite	khāṭite	FV	Used to work (chalit)	A2
3	khāṭ	-te	khāṭte	FV	Used to work (chalit)	A3
4	khāṭ	-ile	khāṭile	FV	Having worked (chalit)	A4
5	khāṭ	-le	khāṭle	FV	Having worked (chalit)	A5
6	khāṭ	-āite	khāṭāite	FV	Used others to work (chalit)	A6
7	khāṭ	-āte	khāṭāte	FV	Used others to work (chalit)	A7
8	khāṭ	-āile	khāṭāile	FV	You made others work (chalit)	A8
9	khāṭ	-āle	khāṭāle	FV	You made others work (chalit)	A9

Set 2: Outputs as NFVs

No	Root	Suffix	Final	POS	Gloss	Type
1	khāṭ	-ite	khāṭite	NFV	“Used to work” (sādhu)	B1
2	khāṭ	-te	khāṭte	NFV	“Used to work” (chalit)	B2
3	khāṭ	-ile	khāṭile	NFV	“You worked” (sādhu)	B3
4	khāṭ	-le	khāṭle	NFV	“you worked” (chalit)	B4
5	khāṭ	-iyā	khāṭiyā	NFV	“Having worked” (sādhu))	B5
6	kheṭ	-e	kheṭe	NFV	“Having worked” (chalit)	B6
7	khāṭ	-āite	khāṭāite	NFV	“Used others to work” (sādhu)	B7
8	khāṭ	-āte	khāṭāte	NFV	“Used others to work” (chalit)	B8
9	khāṭ	-āile	khāṭāile	NFV	“Made others work” (sādhu)	B9
10	khāṭ	-āle	khāṭāle	NFV	“Made others work” (chalit)	B10
11	khāṭ	-āiyā	khāṭāiyā	NFV	“making others work” (sādhu)	B11
12	khāṭ	-iye	khāṭiye	NFV	“making others worked” (chalit)	B12

Set 3: Outputs as Nouns

No	Root	Suffix	Final	POS	Gloss	Type
1	khāṭ	-e	khāṭe	NN	on the cot (chalit)	C1
2	khāṭ	-iyā	khāṭiyā	NN	cot made of thread (chalit)	C2

Set 4: Outputs as Adjective

No	Root	Suffix	Final	POS	Gloss	Type
1	khāṭ	-iye	khāṭiye	ADJ	“hard working” (chalit)	D1

There are 24 forms in four sets. The first set (A1-A9) is the finite verb; the second set (B1-B12) is NFV, the third set (C1-C2) is noun, and the fourth set (D1) is adjective. In each set, we use only one form (i.e., *khāṭ*) either as a verb root (in A, B, and D) or as a noun stem (in C). When the word *khāṭ* is used as a verb root, it means “to work” (A, B, and D), and when used as a noun stem, it means “cot” (C). At the non-inflected level, we get two different meanings: verb root and noun stem; at the inflected level, we find many new meanings of the form. When we compare the final forms, we make some interesting observations about their form and function in different parts-of-speech.

(a) Comparing between A1 (i.e., *khāṭe*), B6 (i.e., *kheṭe*), and C1 (i.e., *khāṭe*), we find that in all three cases, the suffix *-e* is used either with a verb root (A1 and B6) or with a noun stem (C1) to generate three (semi-)identical forms. In A1, it denotes a sense of completeness (*-e* is a finite verb suffix); in B6, it denotes a sense of incompleteness (*-e* is an NFV suffix); and in C1, it denotes a sense of location (*-e* is a locative case marker for noun). Moreover, in B6 (i.e., *kheṭe*) the original root vowel (/a/) undergoes a change (i.e., vowel height assimilation) due to the presence of the case marker *-e* in the following syllable (*khāṭe >kheṭe : ā > e/-e*).

(b) Comparing between A2 (i.e., *khāṭite*), A3 (i.e., *khāṭte*) on one hand and B1 (i.e., *khāṭite*) and B2 (i.e., *khāṭte*) in other, we observe that the suffixes *-ite* and *-te* are used with the same verb root (i.e., *khāṭ-* “to work”) but in different senses. In A2 and A3,

suffixes are used in a sense of completeness of an action, while in B2 and B3, they are used in a sense of incompleteness of an action.

(c) Comparison between A4 (i.e., *khāṭile*) and A5 (i.e., *khāṭle*) on one hand and B3 (i.e., *khāṭile*) and B4 (i.e., *khāṭle*) in other, we find that the suffixes *-ile* and *-le* in A4 and A5 generate a sense of condition. On the other hand, they denote a sense of completeness in B4 and B5 with relevant information for the person (Second), number (singular and plural), and tense (simple past).

(d) Comparing between B12 (i.e., *khāṭiye*) and D1 (i.e., *khāṭiye*), we find that the suffix *-iye* in B12 denotes non-finiteness of action with a sense of causation, while in D1, it denotes an adjectival sense indicating a high level of efficiency of an individual.

Thus, we find that *khāṭe* records three different senses (NFV, FV, and NN), *khāṭite* and *khāṭte* record two different senses each (FV and NFV), *khāṭile* and *khāṭle* record two different senses (FV and NFV), and *khāṭiye* record two different senses (NFV and ADJ). This implies that identical suffixed words can vary in parts-of-speech and meanings. We also find some relevant morphosemantic information about these forms when we analyze them in context-free situations (i.e. when these forms are not used in sentences). Since they refer to multiple senses as context-free independent lexical items, they are free to belong to different parts-of-speech without referring to their actual grammatical identities. Therefore, when we analyze their structural ambiguity to understand their grammatical role and meaning, we integrate their morphophonemic information with their surface forms. Also, we refer to their contexts of use to determine in which part-of-speech they are used in texts.

8 Deformation of Bengali Verb Roots

We observe that the majority of Bengali verb roots can have any one of the following three structures: VC, CV, or CVC [V = Vowel, C = Consonant]. A vowel-ending verb root ends with *-ā, -i, -ī, -e, -o, -u,* and *-ū*. This information is useful when we find that the addition of a verb suffix with a verb root does not usually change the primary form of a root. Verb suffixes are generally attached to roots without causing any morphophonemic change in the root part. However, there are exceptions to this rule. Some verb roots may undergo a morphophonemic change when a suffix is tagged to roots. During suffix addition, the roots *ān-, ās-, kāch-, kāṭ-, khāṭ-, māt-, pāt-, ýā-, khā-, gā-, chā-, dhā-, pā-, uṭh-, gun-, khul-, chhuṭ-, kin-, gil-, chin-, chir-, jit-, mil-, di-, and ni-* undergo change. There are several phonological and grammatical factors (e.g., *vowel harmony, segment assimilation, change in tense class, change of sādhu (chaste) to chalit (colloquial) form of verbs*) which are responsible for causing different kinds of morphophonemic changes. We record those factors that operate behind a change in a vowel in a verb root; without this, it is difficult to understand the processes of morphological change and concatenation between roots and suffixes.

In some early works, some attempts are made to understand the kinds of changes that take place in roots while roots are tagged with suffixes to generate final forms (Sengupta and Chaudhuri 1993, Dash, Chaudhuri and Kundu 1997, Sengupta 1997). These studies are, however, neither elaborate nor exhaustive. Keeping this information in view, we investigate the patterns of change in verb roots while they are tagged with suffixes to produce NFVs. In our view, by applying the following rules, it is possible to shed light on possible changes in verb roots while they are tagged with suffixes for producing NFVs. In most cases, these are

instances of vowel height assimilation due to the presence of a vowel of a different class in a suffix that is tagged to a verb root.

(a) Root *āg-* changes into *eg-* when it is tagged with NFV suffix *-iye, -ule, -ute* (e.g., *āg- + -iye > egiye* "moving forward", *āg- + -ule > egule* "going ahead", *āg- + -ute > egute* "to go ahead"). The vowel /a/ in the root is changed into /e/ due to the presence of /i/ in the suffix.

(b) Root *ās-* changes into *es-* when it is tagged with NFV suffix *-e*, e.g., *ās- + -e > ese* "coming". The vowel /a/ in the root changes into /e/ because of the presence of /i/ in the suffix.

(c) Root *uṭhā-* changes into *oṭhā* when it uses NFV suffix *-te* and *-le* (e.g., *uṭhā- + -te > oṭhāte* "to uplift", *uṭhā- + -le > oṭhāle* "having uplifted"). The vowel /u/ in the root is changed into /o/ because of the presence of /a/ in the suffix. Strikingly, this kind of change does not happen in the case of *sādhu* (chaste) NFVs. In such cases, the vowel in the root usually remains unchanged in spite of the presence of the appropriate condition (e.g., *uṭhā- + -ite > uṭhāite, uṭhā- + -ile > uṭhāile*). It rarely happens for *chalit* (colloquial) forms where the vowel in the root remains unchanged even if it uses the NFV suffix *-te* and *-le* (e.g., *uṭhāte* and *uṭhāle*). The use of such forms in modern Bengali texts is very rare.

(d) Roots *kāṭ-, nāch-, pāt-, bhāb-, mār-, hār-* and *chhāṛ-* change into *keṭ-, nech-, pet-, bheb-, mer-, her-* and *chheṛ-*, respectively, when these roots use NFV suffix *-e* (e.g., *kāṭ- + -e > keṭe* "cutting", *nāch- + -e > neche* "dancing", *pāt- + -e > pete* "laying", *bhāb- + -e > bhebe* "thinking", *mār- + -e > mere* "killing", *hār- + -e > here* "losing", *chhāṛ- + -e > chheṛe* "leaving"). In these cases, the vowel /a/ in the root changes into the vowel /e/ due to the presence of the vowel /e/ in the suffix.

(e) Roots *khā-* and *pā-* change into *khe-* and *pe-*, respectively, when they use the suffix *-te, -le,* and *-ye* (e.g., *khā- + -te > khete* "to eat", *pā- + -te > pete* "to get", *khā- + -le > khele* "eating", *pā- + -le > pele* "getting", *khā- + -ye > kheye* "having eaten", *pā- + -ye > peye* "getting"). For these verbs, the vowel /a/ in the root changes into the vowel /e/ due to the presence of the vowel /e/ in the suffix.

(f) Roots *gā-, dhā-, chā-, chhā-* change into *ge-, dhe-, che-* and *chhe-*, respectively, when they use the suffix *-ye* (e.g., *gā- + -ye > geye* "having sung", *dhā- + -ye > dheye* "having run", *chā- + -ye > cheye* "having wanted", *chhā- + -ye > chheye* "having shaded"). For these verb roots, the vowel /a/ in the root changes into /e/ due to the presence of the vowel /e/ in the suffix.

(g) Root *ýā* changes into *ge-* when it uses the suffix *-le* (e.g., *yā- + -le > gele* "going"). Here both root vowel and root consonant undergo change. In the case of vowel, the vowel /a/ in the root changes into vowel /e/ due to the presence of the vowel /e/ in the suffix. On the other hand, the consonant ý /ʤ/ in root changes into *g* /g/ due to other phonological factors not being clearly understood. Surprisingly, in the case of its *sādhu* (chaste) form, this kind of change does not take place. The vowel in the root remains unchanged in spite of the use of the *sādhu* suffix with root (e.g., *ýā- + -ile > ýāile,* not **gāile*).

(h) Root *ýā-* changes into *ýe-* when it uses the suffix *-te* (e.g., *ýā- + -te > ýete* "to go"). In this case, the vowel /a/ in the root changes into /e/ due to the presence of the vowel /e/ in the suffix. In its *sādhu* (chaste) form, this does not occur. Here, the vowel in root

remains unchanged in spite of the use of the *sādhu* suffix with root (e.g., *ýā- + -ite > ýāite* not **gāite*).

(i) Root *ýā-* changes into *gi-* when it uses the suffix *-ye* (e.g., *ýā- + -ye > giye* "having gone"). In this case, /a/ in the verb root changes into /e/ due to the presence of /e/ in the suffix. Also, the consonant *y* /ʤ/ in root is changed into *g* /g/ in the final form.

(j) Root *ýā-* changes into *gi-* when it uses the suffix -y*ā* (e.g., *ýā- + -yā > giyā* "having gone"). This form (i.e., *giyā*) goes through various morphophonemic changes before it gets its final form (i.e., *ýā- + -iyā > ýāiyā > giyā*). Initially, the vowel /a/ in the root is retained in the intermediate form (*ýāiyā*). However, when the consonant *ý* /ʤ/ in the root is changed into *g* /g/, the vowel in the root is deleted and the vowel /i/ of the suffix occupies the place vacated by the root vowel.

9 Grouping Verb Roots and NFV Suffixes

We observe that NFVs occupy a major share in the total occurrence of conjugated verbs in modern Bengali. The idea of 'non-finiteness' is defined semantically but it is determined structurally (based on its usage) following some specific suffixes attached to roots as well as on the amount of information derived from contexts of use in texts. The morpho-syntactic processes that are operated to produce NFVs in Bengali involve two major processes in the following manners:

(3)	Root + NFV suffix =	NFV
	$kar_{[FV\_RT]}$ + $-te_{[NFV\_Suffix]}$ >	$karte_{[NFV]}$ "to do"

(4)	Root + causative suffix + NFV suffix =	NFV
	$kar_{[FV\_RT]}$ + $-ā_{[Causative\text{-}Suffix]}$ + $-te_{[NFV\_Suffix]}$ >	$karāte_{[NFV]}$ "forcing others do"

The suffixes of NFV are two types: simple and causative. There are 10 suffixes (e.g., *-e, -ile, -ite, -iyā, -le, -te, -ule, -ute, -yā, -ye*) for simple NFVs and 12 suffixes (e.g., *-āile, -āite, -āiyā, -āle, -āte, -iye, -oāile, -oāite, -oāiyā, -oāle, -oāte, -oyāiyā*) for causative NFVs including both *sādhu* (chaste) and *chalit* (colloquial) forms. The suffixes are used at the root-final position. We note that there are some restrictions in the valid grammatical mapping of suffixes with verb roots. The restrictions are summarized in the following manners.

(i) The suffix *-iyā* is used with roots ending in *-u* and a consonant, e.g., *dhu- + -iyā > dhuiyā* "having washed", *kar- + -iyā > kariyā* "having done".

(ii) The suffix *-yā* is used with roots ending in *-i* and a consonant, e.g., *di- + -yā > diyā* "having given", *ān- + -iyā > āniyā* "having brought".

(iii) The suffix *-ye* is used with roots ending in *-i* and *-u*, e.g., *ni- + -ye > niye* "having taken", *dhu- + -ye > dhuye* "having washed".

(iv) Suffixes *-te* and *-le* are used with roots ending in *-i, -u*, and a consonant, e.g., *di- + -te > dite* "to give", *di- + -le > dile* "giving", *śu- + -te > śute* "to lie", *śu- + -le > śule* "lying", *bal- + -te > balte* "to say", *bal- + -le > balle* "having said", *ān- + -te > ānte* "to bring", *ān- + -le > ānle* "bringing".

(v) Suffix *-e* is used with roots ending in a consonant only, e.g., *kar- + -e > kare* "having done", *ān- + -e > ene* "having brought", *gun- + -e > gune* "having counted".

(vi) Suffixes *-ule* and *-ute* are used with root *eg*, e.g., *eg-* + *-ule* > *egule* "going ahead", *eg-* + *-ute* > *egute* "to go ahead".
(vii) Suffixes *-ole* and *-ote* are used with root *pichh*, e.g., *pichh-* + *-ole* > *pichhole* "going backward", *pichh-* + *-ote* > *picchote* "to go backward".

Based on the information given above, we divide NFVs into different sub-types in accordance with their possible concatenations for valid root and suffix pairing along with related grammatical and semantic information. While the basic semantic information is primarily stored in the root part, grammatical information is normally found in the suffix part. The patterns of combinations between root and suffix (for simple and causative forms) are divided into 12 sub-groups for generating a valid NFV (Table 5). The number will increase if emphatic particles (i.e., *-i* and *-o*) are added at the end of each suffix.

Table 5: Different groups for valid NFV root-suffix pairing

Group	Root	Simple	Causative
1	kar-, dekh-, khul-, mil-, hār-, kin-, śun-	-iyā, -ile, -le, -ite, -te, -e	-āiyā, -āile, -āle, -āite, -āte, -iye
2	khā-, gā-, pā-, ýā-	-iyā, -ite, -ile	-oāiyā, -oāile, -oāle, -oāite, -oāte, -iye
3	ān-, kāṭ-, mākh-, ās-	-ite, -te, -ile, -le,	-āiyā, -āite, -āte, -āile, -āle, -iye
4	en-, keṭ-, mekh-, es-	-e	
5	khe-, ge-, pe-	-ye, -te, -le	
6	gi-	-ye, yā	
7	ýe-	-te	
8	di-, ni-	-ye, -te, -le, -yā	
9	de-, ne-		-oyāiyā, -oāite,-oāte, -oāile, -oāle
10	śu-, dhu-	-iyā, -ye, -ite, -te, -ile, -le	-āiyā, -āite, -āte, -āile, -āle, -iye
11	śo-, dho-		-āiyā, -āite, -āte, -āile, -āle
12	eg-	-ule, -ute	
13	pichh-	-ole, -ote	-āiyā, -āile, -āle, -āite, -āte, -iye

10 POS Annotation of Bengali NFVs

To annotate Bengali NFVs, we consider several linguistic features: (a) phonological features, (b) morphological features, (c) morphophonemic changes of verbs, and (d) information on the contextual use of verbs. An annotation process should also include all the basic requirements that are essential for a system: (i) morphological information, (ii) part-of-speech information, and (iii) semantic information. To address these needs we may need to use models and theories that are applied in English and other Indian languages (Antony *et al.* 2010, Atwell *et al.* 2000, Avinesh and Karthik 2007, Baskaran *et al.* 2008, Dandapat 2009, Dash 2015, Dash

2021, Dhanalakshmi *et al.* 2009, Ekbal *et al.* 2007, Garrette and Baldridge 2013, Ide and Pustejovsky 2017, Kumar and Josan 2010, Manning 2011, Mishra and Mishra 2011, Nagata *et al.* 2018; Naseem *et al.* 2009, Nguyen and Verspoor 2018, Pammi and Prahallad 2007, Rao and Yarowsky 2007, Rao *et al.* 2007, Ray *et al.* 2010, Saharia 2009, Sastry *et al.* 2007, Schulz and Kuhn 2016, Shambhavi and Ramakanth 2010, Shambhavi *et al.* 2012, Shrivastava and Bhattacharyya 2008, Singh and Jha 2015, Toutanova and Manning 2000, Wallis 2007, Wallis 2014, Wallis 2020, Yang and Eisenstein 2016). Also, we should refer to those models and techniques that are proposed for processing Bengali words (Sengupta and Chaudhuri 1993, Dash *et al.* 1997, Sengupta 1997, Saha and Debnath 2004, Dandapat 2007, Chakrabarti 2011, Dash 2013, Dash 2021).

During part-of-speech annotation of NFVs, we suggest giving importance to orthographic forms of Bengali NFVs. This is mooted on the argument that all Bengali NFVs are available in texts in conventional orthographic forms for analysis and annotation. We also suggest referring to all phonological rules to explain morphophonemic changes that concatenate roots and suffixes. The approach is useful for implementing an annotation process in a straightforward manner with a maximum amount of linguistic information. Moreover, with necessary modifications, it may be applied to annotate words of other parts-of-speech in Bengali and other languages which have words with similar morphological forms.

An annotation process, at an early stage, involves several important tasks: extraction of orthographic, grammatical, and semantic information from the analysis of NFVs. At the linguistic level, it involves analysis of morphemes used to form NFVs and extraction of grammatical, syntactic, and semantic information from these properties. It can be done at context-free and context-bound levels to understand the form and function of NFVs as well as gather information for designing strategies for the annotation of these words.

(a) Context-free level: Analyze NFVs in isolated situations to understand and extract their grammatical information, and
(b) Context-bound level: Analyze NFVs in sentential contexts to gather information about their meanings and their syntactic roles.

The process of part-of-speech annotation should start with the development of a semi-supervised and manually controlled annotation tool. Once it identifies an NFV in the sentence, we analyze its form and context to confirm if it is a NFV. The logic is—if a Bengali verb is to be marked as a NFV it has to be suffixed because there is no NFV in Bengali which does not carry a suffix. Therefore, it must have a root and a suffix. Both parts must grammatically agree (as in Table 6) to be annotated as a valid NFV. The root and suffix should satisfy all conditions to be rightly annotated. To find an NFV, the tool should have information on the following 4 types:

(a) Information about morphophonemic change that might have occurred during the joining of suffixes with root,
(b) Semantic information from the root about the nature of action,
(c) Information from suffixes about the grammatical and semantic nature of a verb, and
(d) Information about the contextual environment of its use in texts.

The information summarized above is to be stored for the sense disambiguation of NFVs in Bengali. Following this method, it is possible to annotate most of the NFVs accurately. However, in some situations, it may produce more than one output due to 'structural ambiguity' caused by identical affixes. In such a situation, we should separately store these forms in a list for further investigation, analysis, and disambiguation. By following this, we can annotate NFVs in the language. The following example (Figure 2) shows how we can annotate NFVs in Bengali texts, which can be applied to Bengali finite verbs as well.

BNGAS 00102	tārā sei bāṛite chaṛiye chiṭiye thāke
	tārā\PR_PRP\ sei\DM_DMR\ bāṛite\NN_NN\ chaṛiye\VR_VM_VNF\ chiṭiye\VR_VM_VNF\ thāke\VR_VM_VF\ .\RD_PUNC\
BNGTV 00578	puliś khũje pete sei ṭākā phiriye dey bhadralokke.
	puliś\NN_NN\ khũje\VR_VM_VNF\ pete\VR_VM_VNF\ sei\DM_DMR\ ṭākā\NN_NN\ phiriye\VR_VM_VNF\ dey\VR_VM_VF\ bhadralokke\NN_NN\ .\RD_PUNC\

Figure 2: A sample Bengali text where NFVs are annotated

The method that we suggest here is useful in the identification and annotation of almost all NFVs in Bengali. Due to their polysemous/homonymous identity, some NFVs may be annotated as finite verbs, nouns, and adjectives. We can overcome such ambiguities if we can annotate nouns, verbs, and adjectives along with NFVs and utilize information on their contextual use in the annotation. The uniqueness of this approach is that part-of-speech annotation can be jointly applied with morphological processing and the process of sense disambiguation of words.

11 Conclusion

In this paper, we present an empirical, data-based, and manually done morphological analysis of Bengali NFVs which we collected from a modern Bengali written text corpus. We propose to utilize data and information derived from this analysis to develop a tool for part-of-speech annotation of Bengali NFVs. We refer to multiple senses in which NFVs are used in Bengali as well as analyze them to understand their orthographic forms, contextual roles, grammatical functions, and semantic information. Moreover, we look at the morphophonological processes that play crucial roles in changing the structure of NFVs in Bengali and analyze them to see how roots and suffixes concatenate, in various possible ways, to generate valid NFVs in Bengali.

The analysis of Bengali NFVs shows that orthographic and morphological information plays a crucial role in understanding their form and function in the language. This analysis also defines a guideline for compiling a list of NFVs from the Bengali corpus, classifying them based on their orthographic forms, dividing their roots and suffixes, analyzing their morphosyntactic roles, identifying their processes of concatenation, defining their grammatical mapping rules, identifying their semantic information embedded in contexts, and developing methods for their part-of-speech annotation. The results of this study may be applied to Bengali language description, language teaching, grammar writing, dictionary compilation, machine learning, morphological processing and lexical database

generation. The sets of rules that we propose here and the data that we collect can be used to develop a part-of-speech annotation tool for Bengali as well as a morphological analyzer for explaining the word formation processes used to form NFVs in Bengali. The major limitation of this study is that it does not try to develop an indigenous part-of-speech annotation tool that can annotate NFVs and address the challenges faced during the phases of development of the system.

Acknowledgement

This paper is first presented at the *48th All India Conference of Dravidian Linguistics (AICDL-48)*, Department of Linguistics, Bharathiar University, Coimbatore, India, during 25-26 February 2022. The author expresses his sincere thanks to the blind peer reviewers for their critical comments for improving the quality of the paper.

References

Antony, P.J.; Santhanu, P.M. and Soman, K.P. 2010. SVM-based parts-of-speech tagger for Malayalam. *Proceedings of the International Conference on Recent Trends in Information, Telecommunication & Computing* (ITC 2010), Kochi, Kerala, 2010, pp. 339-341.

Atwell, E.; Demetriou, G.; Hughes, J.; Schiffrin, A.; Souter, C. and Wilcock, S. 2000. A comparative evaluation of modern English corpus grammatical annotation schemes. *International Computer Archive of Modern English Journal.* 24: 7-23.

Avinesh, P.V.S. and Karthik, G. 2007. POS tagging & chunking using Conditional Random Field and Transformation-based learning. *Proceedings of the Workshop on Shallow Parsing for South Asian Languages (IJCAI-07),* IIIT-Hyderabad, India, pp. 21-24.

Baskaran, S.; Bali, K.; Bhattacharya, T.; Bhattacharya, P.; Chaudhury, M.; Jha, G.N.; Rajendran, S.; Sarvanan, K.; Sobha, K., and Subbarao, K.V. 2008. Designing a common POS tagset framework for Indian Languages. *Proceedings of the 6th Workshop on Asian Language Resources, Asian Language Resources in International Joint Conference on Natural Language Processing (IJCNLP-2008),* 11-12 January 2008, IIIT-Hyderabad, pp. 89-92.

Bhattacharja, S. 1998. *Sanjanani Vyakaran* (Generative Linguistics). Dhaka: Tcarou Prakashani.

Boguraev, B. and Pustejvsky, J. (eds.) 1996. *Corpus Processing for Lexical Acquisition.* Cambridge, MA: MIT Press.

Bouillon P. and Busa, F. (eds.) 2001. *The Language of Word Meaning.* Cambridge: Cambridge University Press.

Chaki, J. 1996. *Bangla Bhasar Byakaran (Grammar of the Bengali language)* Kolkata: Ananda Publishers.

Chakrabarti, B. 1985. *Ucchatara Bangla Byakaran (Higher Grammar of Bengali).* Kolkata: Akshay Malancha.

Chakrabarti, D. 2011. Layered parts of speech tagging for Bangla. *Language in India. www.languageinindia.com,* May 2011, Special Volume: Problems of Parsing in Indian Languages. Pp. 1-6.

Chatterji, S.K. 1926/1993. *The Origin and Development of the Bengali Language.* Calcutta: Rupa Publications.

Chattopadhyay, S.K. 1995. *Bhasa Prakash Bangala Vyakaran (The Grammar of the Bengali Language).* Kolkata: Rupa Publications.

Cruse, A. 1986. *Lexical Semantics.* Oxford: Oxford University Press.

Cruse, A. 2000. *Meaning in Language: An Introduction to Semantics and Pragmatics.* Oxford: Oxford University Press.

Dandapat, S. 2007. POS tagging and chunking with the Maximum Entropy model. *Proceedings of Workshop on Shallow Parsing for South Asian Languages (IJCAI-07),* IIIT-Hyderabad, India, pp. 29-32.

Dandapat, S. 2009. *Part-of-Speech tagging for Bengali.* Unpublished MS Thesis, Dept. of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, India.

Dash, N.S. 2002. Corpus generation and text processing. *International Journal of Dravidian Linguistics*. Vol. 31. No. 1. Pp. 25-44.

Dash, N.S. 2005. *Corpus Linguistics and Language Technology: With Reference to Indian Languages.* New Delhi: Mittal Publications.

Dash, N.S. 2013. Part-of-speech (POS) tagging in Bangla written text corpus. *Bhasa Bijnan o Prayukti: An International Journal on Linguistics and Language Technology*. 1(1): 53-96.

Dash, N.S. 2015. Marking words with part-of-speech (POS) tags within the text boundary of a corpus: problems, process, and outcomes. *Translation Today*. Vol. 9. No. 1. Pp. 5-24.

Dash, N.S. 2021. *Language Corpora Annotation and Processing.* Singapore: Springer Nature.

Dash, N.S. and Chaudhuri, B.B. 2000. The process of designing a multidisciplinary monolingual sample corpus, *International Journal of Corpus Linguistics*. 5(2): 179-197, 2000.

Dash, N.S., Chaudhuri, B.B. and Kundu, P.K. 1997. Computer parsing of Bangla verbs. *Linguistics Today.* 1(1): 64-86.

Dhanalakshmi, V.; Kumar, A.; Shivapratap, G.; Soman, K.P. and Rajendran, S. 2009. Tamil POS Tagging using Linear Programming, *International Journal of Recent Trends in Engineering*, Vol. 1. No. 2. Pp. 166-169.

Ekbal, A.; Mandal, S. and Bandyopadhyay, S. 2007. POS tagging using HMM and rule-based chunking, *Proceedings of the Workshop on shallow parsing in South Asian languages (SPSAL),* IJCAI 2007, IIIT-Hyderabad, India, pp. 31-34.

Fellbaum, C. 2000. Autotroponymy. In: Ravin, Y. and Leacock, C. (eds.) *Polysemy: Theoretical and Computational Approaches.* New York: Oxford University Press Inc. Pp. 52-67.

Garrette, D. and Baldridge, J. 2013. Learning a part-of-speech tagger from two hours of annotation. *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-13)* June 2013, Atlanta, GA, pp. 138-147.

Ide, N. and Pustejovsky, J. (eds.) 2017. *Handbook of Linguistic Annotation*. (Text, Speech, and Language Technology series), Springer.

Kreidler, C.W. 1998. *Introducing English Semantics.* London: Routledge.

Kumar, D. and Josan, G.S. 2010. Part-of-speech taggers for morphologically rich Indian languages: a survey, *International Journal of Computer Applications.* Vol. 6. No. 5. Pp. 1-9.

Lascarides, A. and Copestake, A. 1998. Pragmatics and word meaning. *Journal of Linguistics.* 34 (1998): 387-414.

Leech, G. 1974. *Semantics.* Middlesex, Middlesex, England: Penguin Books Ltd.

Majumdar, P.C. 1993. *Bangla Bhasa Parikrama (Survey of the Bengali Language).* Vol. II. Kolkata: Dey's Publishing.

Manning, C.D. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing.* Vol. Part I, Tokyo, Japan, Springer-Verlag Berlin, February 20-26. Pp. 171-189.

Mishra, N. and Mishra, A. 2011. Part of speech tagging for Hindi corpus, *Proceedings of the International Conference on Communication Systems and Network Technologies*, Katra, Jammu, 2011, pp. 554-558.

Nagata, R.; Mizumoto, T.; Kikuchi, Y.; Kawasaki, Y. and Funakoshi, K. 2018. A POS tagging model designed for learners of English. Proceedings of the 2018 EMNLP Workshop W-NUT: *The 4th Workshop on Noisy User-generated Text, Association for Computational Linguistics*, pp. 39-48, Brussels, Belgium, November 01.

Naseem, T.; Snyder, B.; Eisenstein, J. and Barzilay, R. 2009. Multilingual part-of-speech tagging: two unsupervised approaches. *Journal of Artificial Intelligence Research,* Vol. 36. No. 1. Pp. 1-45.

Nguyen, D.Q. and Verspoor, K. 2018. An improved neural network model for joint POS tagging and dependency parsing. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Brussels, Belgium, Association for Computational Linguistics, October 31- November 1, pp. 81-91.

Palmer, F.R. 1995. *Semantics.* 2nd Edition. Cambridge: Cambridge University Press.

Pammi, S.C. and Prahallad, K. 2007. POS tagging and chunking using Decision Forests. *Proceedings of the Workshop on shallow parsing in South Asian languages (SPSAL), IJCAI 2007,* IIIT-Hyderabad, India, Pp. 33-36.

Pustejovsky, J. 1995. *The Generative Lexicon*, Cambridge, MA: MIT Press.

Rao, D. and Yarowsky, D. 2007. Part of speech tagging and shallow parsing of Indian languages. Proceedings of the Workshop on Shallow Parsing for South Asian Languages (IJCAI-07), IIIT-Hyd, India. pp. 17-20.

Rao, P.T.; Ram, S.; Vijaykrishna, R. and Sobha, L. 2007. A text chunker and hybrid POS tagger for Indian languages. *Proceedings of the Workshop on Shallow Parsing for South Asian Languages (IJCAI-07)*, IIIT-Hyd, India, pp., 9-12.

Ravin, Y. and Leacock, C. 2000. Polysemy: An Overview. In: Ravin, Y. and Leacock, C. Eds. 2000. *Polysemy: Theoretical and Computational Approaches.* New York: Oxford University Press Inc. Pp. 1-29.

Ray, P.R.; Harish, V.; Sarkar, S. and Basu, A. 2010. Part of Speech Tagging and Local Word Grouping Techniques for Natural Language Parsing in Hindi. *Proceedings of the International Conference on Natural Language Processing (ICON2003)*, Dept. of Computer Science and Engineering, IIT-Kharagpur, India, pp., 118-125.

Saha, G.K.; Saha, A.B., and Debnath, S. 2004. Computer-assisted Bangla words POS tagging. *Proceedings of (iSTRANS-2004)*, New Delhi, India, pp., 111-115.

Saharia, N.; Das, D.; Sharma, U. and Kalita, J. 2009. Part of Speech Tagger for Assamese Text. *Proceedings of the ACL-IJCNLP-2009 Conference*, Suntec, Singapore, pp., 33-36.

Sarkar, P. and Basu, G. 1994. *Bhasa Jignasa (Queries of Language).* Calcutta: Vidyasagar Pustak Mandir.

Sastry, G.M.; Chaudhuri, S. and Reddy, P.N. 2007. An HMM-based part-of-speech & statistical chunker for 3 Indian languages. *Proceedings of the Workshop on Shallow Parsing for South Asian Languages (IJCAI-07)*, IIIT-Hyderabad, India, pp., 13-16.

Schulz, S. and Kuhn, J. 2016. Learning from Within? Comparing POS tagging approaches for historical text. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, European Language Resources Association, pp. 4316-4322.

Sen, S. 1993. *Bhasar Itibritta (History of Language).* Calcutta: Ananda Publishers Ltd.

Sengupta, G. 1997. Three Models of Morphological Processing, *South Asian Language Review*. 7(1): 1-26.

Sengupta, P., and Chaudhuri, B.B. 1993. A Morpho-Syntactic Analysis Based Lexical Sub-system'. *International Journal of Pattern Recognition and Artificial Intelligence*. 7(3): 595-619.

Shahidullah, M. 2003. *Bangala Bayakaron (Bengali Grammar).* Dhaka: Mowla Brothers.

Shambhavi, B.R. and Ramakanth, P.K. 2010. Current State of the art POS tagging for Indian Languages: a study. *International Journal of Computer Engineering and Technology.* Vol. 1. No. 1. Pp. 250-260.

Shambhavi, B.R.; Ramakanth, K.P. and Revanth, G. 2012. A Maximum Entropy Approach to Kannada Part of Speech Tagging. *International Journal of Computer Applications.* Vol. 41. No.13., Pp. 9-12.

Shrivastava, M. and Bhattacharyya, P. 2008. Hindi POS tagger using Naive Stemming: harnessing morphological information without extensive linguistic knowledge, *Proceedings of the 6th International Conference on Natural Language Processing (ICON-2008),* CDAC, Pune India, 20-22 December 2008, pp., 1-8.

Sinclair, J. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Singh, S. and Jha, G.N. 2015, Statistical tagger for Bhojpuri employing Support Vector Machine. *Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI),* 2015, pp., 1524-1529.

Thompson, H-R 2010. *Bengali: A Comprehensive Grammar*. London and New York: Routledge (Taylor and Francis).

Thompson, H-R. 2012. *Bengali.* London Oriental and African Language Library 18, Amsterdam: John Benjamins.

Todd, L. 1987. *An Introduction to Linguistics.* Essex, UK: Longman York Press.

Toutanova, K. and Manning, C.D. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing & Very Large Corpora (EMNLP/VLC-2000),* pp. 63-70.

Ullmann, S. 1962. *Semantics: An Introduction to the Science of Meaning.* Oxford: Blackwell.

Wallis, S.A. 2007. Annotation, Retrieval, and Experimentation. In: Meurman-Solin, A. and Nurmi, A.A. (eds.) *Annotating Variation and Change*. Helsinki: Varieng, UoH (ePublished).

Wallis, S.A. 2014. What might a corpus of parsed spoken data tell us about language? In: Veselovská, L. and Janebová, M. (eds.) *Complex Visibles Out There. Proceedings of the Olomouc Linguistics Colloquium 2014: Language Use and Linguistic Structure*. Olomouc: Palacký University, Czech Republic, 2014, pp., 641-662.

Wallis, S.A. 2020. Grammar and Corpus Methodology. In: Aarts, B.; Popova, G. and Bowie, J. (eds.) *Oxford Handbook of English Grammar*. Part I: Chapter 4. Pp. 58-83. Oxford: Oxford University Press.

Yang, Y. and Eisenstein, J. 2016. Part-of-speech tagging for historical English. *Proceedings of NAACL-HLT 2016,* San Diego, California, Association for Computational Linguistics, June 12-17, pp., 1318-1328.

Yule, G. 1985. *The Study of Language: An Introduction.* Cambridge: Cambridge University Press.

Web Links

https://www.cs.umd.edu/~nau/cmsc421/part-of-speech-tagging.pdf
https://www.geeksforgeeks.org/nlp-part-of-speech-default-tagging/
https://www.guru99.com/pos-tagging-chunking-nltk.html
https://www.languageinindia.com
https://www.ldc.upenn.edu/Catalog/
https://www.nltk.org/book/ch05.html
https://www.shiva.iiit.ac.in/SPSAL2007/proceedings.php
https://www.sketchengine.eu/pos-tags/
https://www.tutorialspoint.com/natural_language_processing/

Niladri Sekhar Dash
Linguistic Research Unit
Indian Statistical Institute
203, B.T. Road, Kolkata -700108
West Bengal, India
e-mail: niladri@isical.ac.in
e-mail: ns_dash@yahoo.com

In SKASE Journal of Theoretical Linguistics [online]. 2023, vol. 20, no. 3 [cit. 2023-12-06]. Available on web page http://www.skase.sk/Volumes/JTL54/02.pdf. ISSN 1336-782X