THE DHAKA UNIVERSITY JOURNAL OF LINGUISTICS
@2025 Department of Linguistics

Vol. 18, No. 35, June 2025, P. 61-98
ISSN : 2789-3014, eISSN : 2075-3098
http://dx.doi.org/10.70438/dujl/1835/0003

Morphosyntactic Analysis of Bangla Matrix Verbs: A Computational Perspective

Naira Khan*

Abstract: This paper analyses verb morphology of Bangla matrix verbs or simplex predicates using a computational lens. It analyses the structure of matrix verbs in terms of Tense-Aspect-Modality (TAM) markers as well as pronominal and valence marking, and proposes a conceptual Finite State Transducer (FST) that captures the rich inflectional feature-values in a concatenative manner. Often linguistic theoretic models and computational analyses are at loggerheads due to the units assumed. Linguistic theory is often grounded in psycholinguistic reality, while computational perspectives prioritise the optimisation of the implementation algorithm resulting in units that do not port over to linguistic theoretical frameworks. This paper analyses verb morphology in terms of TAM markers from a linguistic theoretic perspective and then extends the analysis by modeling a FST of said verbform with TAM and pronominal markers. The analysis is limited to simplex predicates or matrix verbs of Standard Colloquial Bangla (SCB), and can be extended to complex predicates by means of the concatenation of TAM markers on the vector or light verb, since the pole in a complex predicate remains static (Butt 2010, Masica 1993). The Finite State Transducer proposed here is a theoretical model, which can be implemented by researchers working with Bangla verb morphology and Natural Language Processing (NLP) tools. The goal is to develop a more accurate and linguistically informed model of verb morphology for Bangla simplex and complex predicates that can serve both descriptive and applied purposes. This research work contributes to both the theoretical understanding of Bangla morphology and the practical advancement of NLP tools for low-resource languages.

Keywords: Bangla verb morphology, Morphosyntax, Simplex and complex predicates, TAM markers, Finite State Transducers, Computational linguistics, Natural Language Processing.

* Associate Professor, Department of Linguistics, University of Dhaka, Bangladesh
E-mail : nairakhan@du.ac.bd

1. Introduction

Ranked at 6th or 7th as one of the most widely spoken languages of the world (Ethnologue 2025), Bangla is an inflectional language possessing rich verbal morphology that comprises simplex and complex predicates. The matrix verb or simplex predicate morphosyntactically patterns with the vector of a complex predicate and is semantically similar to the pole of the complex predicate (See §2.1). The matrix verb hence consists of the verb root and is marked with multiple inflectional suffixes that encode for Tense-Aspect-Modality (TAM) as well as person and valence. As these matrix verbs inflect with multiple recurring partials comprising TAM and pronominal marking, they pose a challenge for a linguistic analysis that maintains a computational lens in order to highlight the concatenative factor in combinatorial well-formedness for the purpose of computational modeling. Despite being one of the largest languages in the world in terms of speakers, where digitisation and Natural Language Processing (NLP) is concerned, Bangla is one of the most under-resourced languages of the world.

Bangla NLP has gained momentum in recent years (Islam 2009) with NLP applications such as Parts of Speech (POS) tagging (Alam et.al. 2016), grammar checkers (Alam et.al.2007), Named Entity Recognition (NER) (Chowdhury et.al. 2018), morphological analysis (Dasgupta & Khan 2004, Dasgupta et.al. 2005, Abdullah et.al. 2007), syntactic parsing (Mosaddeque & Haque 2004), Machine Translation (MT) (Sennrich 2016), Text-to-Speech (TTS) (Alam et. al. 2007), Speech Recognition (Paul 2009), and Optical Character Recognition (OCR) (Hasnat et.al.2008, Chaudhuri & Pal 1998). Over the years NLP research in Bangla has even extended to sentiment analysis (Hasan et.al. 2020, Hasan et.al. 2013) and emotion detection (Das & Bandyopadhyay 2010). However, despite growing interest in South Asian languages within the NLP community, Bangla remains underrepresented in computational research compared to languages such as Hindi. Existing approaches to Bangla NLP range from rule-based and statistical systems (Sengupta 1993, Sengupta & Chaudhuri 1993) to more recent deep learning models (Alam et.al. 2020, Bhattacharjee et.al. 2021, Hasan et.al. 2019), yet where

morphological analysis is concerned -more comprehensive, scalable solutions that balance linguistic accuracy with computational efficiency are required. Computational morphological analysis that is based on robust linguistic theoretic rigor will serve to inform the analysis as well as implementation and bridge the gap between theoretical and applied endeavours.

This paper aims to provide a detailed examination of Bangla verb morphology through a computational lens. It presents a computational analysis of Bangla simplex predicates, aiming to bridge the gap between theoretical linguistics and practical language technology. With a brief introduction (§1), we first present an overview of the structural characteristics of Bangla verbforms, highlighting key inflectional patterns in light of TAM, pronominal and valence marking (§2) with a view to previous analyses. We explore the differing perspectives when working in a linguistic framework versus a computational one in §3, and propose a computational model that integrates linguistic insights with algorithmic strategies by presenting a finite state transducer for Bangla matrix verbs. In §4 we acknowledge the limitations and explore future research possibilities in terms of implementation and evaluation metrics, and conclude the paper in §5.

2. The Morphological Structure of Bangla Matrix Verbs

The class of verbs in any language is often explored in terms of time-stability (Payne 1997) in that these are lexemes that express the least time stable concepts such as actions or events. In terms of morphosyntactic properties, prototypical verbs tend to be heads of verb phrases, predicates of clauses and encode events in a text. In terms of structural properties verbs encode for subject agreement and specify when the event took place, who the participants were and even whether the event really took place.

2.1 Basic Verb Structure

Bangla is a polysynthetic inflectional language with simplex and complex predicates exhibiting rich verb morphology (1-5) (Chatterji 2004, Chatterji 1926, Masica 1993, Jain and Cardona 2008).

Complex predicates in Bangla comprise a V1 V2 structure with the V1 remaining static in participial form and the V2 completing syntactic function by combining with inflectional markers (5). Vectors of a complex predicate, however, are light verbs, which act as an auxiliary in terms of the semantic content of the entire verbal complex (Paul 2003, Dasgupta 1977, Kachru 1980, Butt 2010).

The matrix verb or simplex predicate is a full form predicate in terms of semantic content similar to the V1 or pole of a complex predicate, but morphosyntactically patterns with the V2 or vector and combines with TAM and pronominal markers for syntactic function. Each matrix verb or simplex predicate can inflect for Tense-Aspect-Modality (TAM) as well as person and honorifics, creating a paradigm of 49 inflected forms for each root notwithstanding periphrastic forms, and another 48 for valence adjusting operations (§2.3).

Following Payne's (1997) framework of a verb, I delineate the basic structure of a Bangla matrix verb in (1):

1) ROOT- {valence-aspect-tense-person-honorific}/mode

2) kin- {-i -ech -il -en } / uk
buy- {caus perf past 3P-Hon } / jus
"kiniechilen/kinuk"

3) tini boi-ti kiniechilen
3P.Sg.Hon book-Det buy.Cause.Perf.Past.3P.Hon
"s/he had made them buy the book."

4) she kinuk
3P.Sg.Ord buy.Jus
"let him/her buy."

Building on Paul (2003) and incorporating (1), I propose the structure of a Bangla complex predicate as:

5) V1+participial V2+{val-asp-tens-pers-hon}/mod

Bangla verbs have bound roots that combine with inflections that are required by the syntactic environment and ground the concept of the root according to time, internal temporal structure, participant reference etc. Verb roots that exist in the syntax without any inflections are in the second person familiar imperative form in the present tense (§2.2.4.1).

2.1.1 Order of Morphemes

As we can see there is a very specific ordering of the morphemes where, the valence marker when present follows the root, the aspect marker precedes the tense marker, the pronominal marker including the honorific follow the tense marker, and the entire tense-aspect construct is replaceable by modality markers.

In the following sections I explore the concatenation of these markers in details before commencing computational modeling of the markers.

2.2 Tense-Aspect-Modality (TAM) markers in Bangla Matrix Verbs

Tense, aspect and modality are operations that anchor the information expressed in a clause according to its sequential, temporal or epistemological orientation (Payne 1997, Masica 1993, Comrie 1981, Wallace 2011). Tense is associated with the sequence of events in real time, while aspect encodes for the internal structure of the event, and mode purports to the speaker's attitude or commitment to the probability of the situation. TAM markers can sometimes be difficult to tease apart especially in inflectional languages in which multiple units of meaning can be fused into a single morpheme. Due to the interrelatedness and interdeterminacy of many TAM operations, glossing often misses the mark. Hence in the current analysis I take into account the analyses that have been put forth thus far and re-analyse the TAM structure of Bangla verbs.

2.2.1 Tense Categorisation in Traditional Grammar

In traditional grammar, Bangla verbs have been classified into three tense categories, namely past, present and future. Each tense

category is then subclassified into further subcategories (Shaw 1984, Morshed 1997, Chatterji 1926). These subcategories are considered tense categories too. However, the subcategories are categories of aspect rather than tense. An example of this traditional classification system is illustrated in diagram 1(Morshed 1997) with conjugations of the verb root "likh-"(write):

		1P	2P	3P
Past	Simple	likhlam	likhle	likhlo
	Continuous	likhchilam	likhchile	likhchilo
	Perfect	likhechilam	likhechile	likhechilo
	Habitual	likhtam	likhte	likhto
Present	Simple	likhi	likho	likhe
	Continuous	likhchi	likhcho	likhche
	Perfect	likhechi	likhecho	likheche
	Habitual	likhi	likho	likhe
Future	Simple	likhbo	likhbe	likhbe
	Continuous	likhte thakbo	likhte thakbe	likhte thakbe
	Perfect	likhe thakbo	likhe thakbe	likhe thakbe
	Habitual	likhbo	likhbe	likhbe

Diagram1: Traditional classification of tense system in Bangla (Morshed 1997:327)

The traditional categories are re-analysed in the next section to tease apart tense and aspect.

2.2.2 Tense Marking in Bangla Verbs

Tense is the grammatical expression indicating the relation of the time of an event to some reference point, usually the moment the clause is uttered. Thinking of time as a line, with "now" represented by a point, moving from left to right, tense can be conceptualised in

terms of the following diagram (2) with the 'now' as the deictic center.

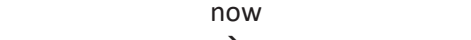

Diagram 2: (Payne 1997: 236)

Languages divide up the conceptual notion of time for the purposes of grammatical marking in various ways. The systems found cross-linguistically comprise (Payne 1997):

1. A tripartite system i.e. past present and future, which is fairly common:

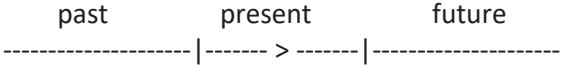

Diagram 3: (Payne 1997: 236)

2. Even more common are two-way distinctions either past/non-past or future/non-future.

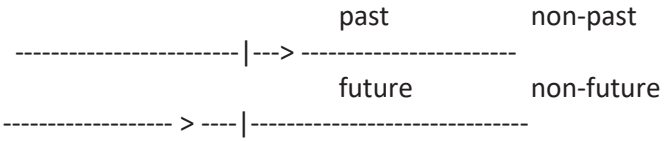

Diagram 4: (Payne 1997: 236)

3. Another possibility is a two-way distinction between present and non-present. In such a system past and future actions tend to be encoded with a single form.

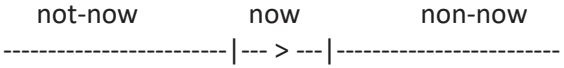

Diagram 5: (Payne 1997: 236)

Although some languages are reported to have as many as 5 tense distinctions in the future, it appears that there are never more distinctions in the future than in the past (Payne 1997).

While Bangla morphosyntactically encodes past, present and future through inflectional morphemes, however, conceptually and

historically Bangla was a non-past/past system same as other Indo-European (IE) languages. Eventually through diachronic change New Indo Aryan (NIA) languages developed a future marker (Masica 1993). Hence IE languages grammaticalise past/non-past, but for NIA we have to add future. Whether future is a tense or a mood is irrelevant in NIA. The future marker combines with aspect-unspecified stems.

The future subsystem in most NIA languages is the unspecified old present, which was crowded out of its role by newer formations, and was left with a range of vaguely future residual meanings – these are retained in most NIA languages as the contingent future or simple subjunctive. In Kashmiri and certain other northwestern languages, it came to function as future per se. In other languages expression of a definite future came to require an additional element {-g}, {-l}, usually with the further element of concord, of these {-g} is clearly an auxiliary, albeit reduced and suffixed. The future marker {-b} which prevail in eastern languages and reach as far west, as Awadhi descend from the OIA future passive Participle of Obligation or gerundive in {~tavya}. The other set represents a survival of the OIA sigmatic future itself {-sya}. An actual {-s} has survived only in Gujarati eastern Rajasthan. Although the {-b} future marker is not strictly speaking, originally, independent auxiliary elements like the {-g} and {-l}, it is convenient to treat them all simply tense and modality (T/M) markers synchronically, as part of a cross linguistic future subsystem, in which this subset of T/M markers combines with a verb stem unspecified for aspect and thus already having future implications (Masica 1993, Jain and Cardona 2007).

Thus we find {-b} as the future marker in Bangla (6).

6) kini > kinbo

Hence, tense reflects how the perception of time is grammaticalised in the language. This grammaticalisation can be expressed in three ways: lexically, morphologically or analytically.

Bangla marks time lexically and morphologically, aspect marking can be found analytically in complex predicates (7):

7) ache > hobe — future: lexical (suppletion)
kine > kinlo — past: morphological
kinlo > kine fello — completive aspect: analytic/periphrastic

The most prolific tense marking grammaticalises morphologically in Bangla, with the present form being the unmarked one, past forms are marked with {~l/~il} and future forms are marked with {~b}. As noted in §2.1, the ordering of the verbal system consists of the root followed by aspect marking followed by tense marking followed by person marking. Where a valence marker is present, namely the causative marker, it precedes the TAM marking. Modality and tense occupy the same slot in the paradigm in NIA languages and hence one replaces the other. In Bangla modality is primarily marked with complex predicates and hence briefly discussed in §2.2.4, and only the jussive marker is included in the paradigm of concatenative markers for matrix verbs.

The three tenses are exemplified through the paradigm of person marking below:

Tense	1P	2PFam	2POrd	2PHon	3POrd	3PHon
Present	kin-i	kin-ish	kin-o	kin-en	kin-e	kin-en
	kin-ch-i	kin-ch-ish	kin-ch-o	kin-ch-en	kin-ch-e	kin-ch-en
	kin-ech-i	kin-ech-ish	kin-ech-o	kin-ech-en	kin-ech-e	kin-ech-en
Past	kin-l-am	kin-l-i	kin-l-e	kin-l-en	kin-l-o	kin-l-en
	kin-ch-il-am	kin-ch-il-i	kin-ch-il-e	kin-ch-il-en	kin-ch-il-o	kin-ch-il-en
	kin-ech-il-am	kin-ech-il-i	kin-ech-il-e	kin-ech-il-en	kin-ech-il-o	kin-ech-il-en
	kin-tam	kin-t-i	kin-t-e	kin-t-en	kin-t-o	kin-t-en
Future	kin-b-o	kin-b-i	kin-b-e	kin-b-en	kin-b-e	kin-b-en

Diagram 6: Tense marking in Bangla verbs

There is some confusion regarding the boundary of some of the aspectual markers. The TAM forms {~echil-} and {~chil-} have been analysed as {~echi+l} and {~chi+l} respectively, with {~ech/echi-} and {~ch/chi-} as the allomorphs of the aspect markers and {~l} as the past tense marker (Sultana 2018). However, I re-analyse the TAM forms {~echil-} and {~chil-} as {~ech+il} and {~ch+il} respectively, with {~ech-} and {~ch-} as the aspect markers and {~il/l} as the allomorphs of the past tense marker.

The logic for this reanalysis can be found in older variants of Bangla. If we look at the same TAM paradigm in Old Bangla as exemplified in the Shadhu variant (diagram 7), we find that the past form is consistently marked with {~il} (Chatterji 1926). Hence, I can infer that the past marker has {~il} as its underlying form, with {~il} and {~l} realising as its allomorphs in SCB.

Tense	Aspect	1P	2P	3P
Past	Perfective	kin-il-am	kin-il-e	kin-il-o
	Progressive	kin-itech-il-am	kin-itech-il-e	kin-itech-il-o
	Perfect	kin-iyach-il-am	kin-iyach-il-e	kin-iyach-il-o

Diagram 7: Past marker in the Sadhu variant of Bangla

According to Masica (1993), there is a unique phenomenon of High and Low verb stems in Bangla due to a previously existing diglossic situation which has been lost over time, where the conditioning factors are no longer present.

We explore aspectual distinctions further in the following section with a full paradigm of TAM marking in the verbal system.

2.2.3 Aspect Marking in Bangla Verbs

Aspect is that which indicates the internal temporal structure of events or states. In this paper I am focusing on the aspectual distinctions that are grammaticalised morphologically in order to create a concatenative computational model. Aspectual distinctions

in Bangla that are grammaticalised analytically or periphrastically are expressed by means of complex predicates. Aspect marking in Bangla comprises the following:

2.2. 3. 1 Morphological Markers of Aspectual Distinctions

2.2.3.1.1 Imperfective

The imperfective aspect denotes a situation as an ongoing process and one that is habitual, or repeated, or incomplete at the reference point in time (Comrie 1985).

There are two types of imperfectives in Bangla with one grammaticalising morphologically:

2.2.3.1.1.1 Habitual

Habitual aspect encodes an assertion that a certain type of event regularly takes place (Payne 1997). In Bangla the habitual aspect fuses with the tense and person marker in the present form and does not grammaticalise as a separate morphological marker, hence they do not anchor dynamic events at the time of utterance as shown in the 1st, 2nd and 3rd person forms respectively (8).

8) kin-i kin-o kin-e

When the deictic center is shifted to the past the habitual is realised morphologically with the {~t-} marker (9), however there is no separate past tense marker (Masica 1993).

9) kin-t-am kin-to kin-te

2.2.3.1.1.2 Progressive

Traditionally labelled as continuous, progressive aspect marks an ongoing dynamic process (Comrie 1985). In Bangla, the progressive aspect is marked with {~ch-} in verb roots that end in a consonant followed by tense and person marking, as indicated below in the 1st, 2nd and 3rd person respectively (10). In verb roots that end with a

vowel the marker is geminated {~cch-} in order to ameliorate the juncture between the vowel and the fricative (11).

10) a) kin-ch-i kin-ch-o kin-ch-e

b) kin-ch-il-am kin-ch-il-e kin-ch-il-o
[deictic center shifted to the past]

11) {ga-cch-i} [<√ga-]

2.2.3.1.2 Perfective

The perfective aspect does not have internal temporal structure, it views the situation in its entirety independent of the tense (Comrie 1978). This is another aspectual distinction that fuses with the tense marker rather than grammaticalising as a separate morphological marker in Bangla.

12) kin-l-am kin-l-o kin-l-e

2.2.3.1.3 Pluperfect

Pluperfect refers to a combination of an aspect and a tense. Pluperfect generally combines the perfect aspect with the past tense, with the effect of this combination, resulting in a shift of the deictic center from now to some point in the past (Payne 1997, Comrie 1978).

13) kin-ech-il-am kin-ech-il-e kin-ech-il-o

2.2.3.1.4 Perfect

Perfect is an aspect that describes a currently relevant state brought about by the situation expressed by the verb (Payne 1997, Comrie 1985). The perfect aspect in Bangla is marked with {~ech-} in verb roots ending in a consonant, followed by tense and person marking, as indicated below.

14) kin-ech-i kin-ech-o kin-ech-e

In verb roots with a vocalic ending we often see allomorphy of the root, and the juncture between the two vowels ameliorated with an approximant (15):

15) cheyechi < {che [<√cha] + ech +i}

As noted in §2.2.2, the future form combines with verb stems unspecified for aspect. However progressive aspect in the future can be marked analytically through the combination of a participial on the pole verb and an aspect marking light verb (16):

16) kinte thakbo

2.2. 3. 1 Periphrastic/analytic markers for aspectual distinctions:

Bangla also expresses aspectual distinctions analytically through complex predicates by means of specific light verbs (17).

17)	kine fellam	completive
	kinte thakbo	progressive with the deictic center shifted to the future
	kine thakbe	future perfect

As we are dealing mainly with the concatenative factor of matrix verbs, I will primarily gloss over the periphrastic aspectual operations and mark it as a future endeavour.

2.2.4 Modality in Bangla verbs:

Mode describes the speaker's attitude towards the event or state, including the speaker's belief, and the probability of its reality (Payne 1997). Often it describes the speaker's estimation of the relevance of the situation with respect to themselves. The terms mood, mode and modality can be used interchangeably. The highest-level distinction in mode operations is between realis and irrealis, however, these terms describe a continuum rather than a dichotomy. A prototypical realis mode strongly asserts that a specific event or state of affairs has actually happened and holds true. An irrealis mode does not make any assertions regarding the

actuality of an event or situation. Mode interacts significantly with aspect and tense (Wallace 2011).

In NIA languages, tense and mode together constitute a category (Masica 1993, Jain and Cardona 2007). In terms of mutual substitutability, tense is substituted by mode in a particular slot in the paradigm. A verbal expression, may have tense or mode, but not both while either combines more or less freely with aspect (Masica 1993).

18) Aspect+T/M, with $^{\text{CONCORD}}$

Mode in Bangla is primarily encoded periphrastically with presumptive, subjunctive and contrafactive marked with the use of particular light verbs (Masica 1993).

19)	kine thakbe	presumptive [the presumptive subsumes the future]
	kine thake	subjunctive
	kine thakto	contrafactive

As this paper is focusing on the combinatorial properties of matrix verbs instead of V1+V2 complex predicates, we will limit the discussion of modality to morphological markers.

The only modes that are grammaticalised as a morphological marker in Bangla are the imperative and the jussive, which falls into the irrealis category. However, Payne (1997) denotes the imperative not as a modality marker but rather a pragmatically marked verb form.

2.2.4.1 Imperative

In Bangla, as the imperative marker patterns with the 2nd person imperfective in the present tense or the future form, we choose to merge the former with the latter in in the finite transducer model for the sake of disambiguation. The imperative forms in the 2nd

person across the honorific spectrum, in present and future forms respectively, with the future marked with {~b-}, are given below (20):

20) kin-o [ordinary]
kin-b-e

kin-en [honorific]
kin-b-en

kin [familiar]
kin-b-i

The imperative in the second person familiar in the present tense is of interest as it is the only form that patterns with the root form and does not require any inflection for the syntactic environment (21).

21) √kin > kin [buy.imp.2Pfam]

2.2.4.2 Jussive

The jussive mode falls in the imperative-hortative system of modalities and involves the 1st and 3rd person in terms of a command relationship (Zanuttini 2012, Saetbyol 2017). It is subsumed in the larger category of subjunctive mode and is also inclined towards the irrealis end of the realis-irrealis continuum (Payne 1997).

In Bangla the jussive is realised as the morphological marker {~uk} when preceded by a consonant and {~k} when preceded by a vowel, and indicates a command by the speaker or 1st person but covers the 3rd person (ordinary and honorific) instead of the 2nd person, both in the singular and plural forms.

The modality marker is in complementary distribution with tense markers and replaces the entire tense-aspect construct (Masica 1993).

22) she/tini/tara/tãra/ kin-uk
3PSg.Ord/3P.Sg.Hon/3.Pl.Ord/3P.Pl.Hon buy.jus

"Let him/her/them buy."

23) she kha-k
3P.Sg.Ord eat.Jus

"Let him/her eat."

2.2.5 Pronominal Markers in Bangla Verbs

The category of person distinguishes between speech act participants: the speaker or first person, and the hearer or second person; and non-speech act participants: the third person (Payne 1997).

The pronominal system in Bangla has agreement between full form pronouns and verbal predicates with anaphoric clitics that cliticise to the verb due to which Bangla is considered a pro-drop language. The pronoun/anaphoric clitic paradigm has distinctions across the categories of person and honorifics in Bangla.

2.2.5.1 Full Form Pronouns

Full form pronouns in Bangla are shown in diagram (8) below. These pronouns have agreement with the verbal systems by means of anaphoric clitics. Often different pronouns and anaphoric clitics are used depending on the relative social statuses of the speech of participants (Payne 1997, Jain and Cardona 2008). The Bangla pronominal system has further distinctions in the marking of honorifics with verbal agreement, with a three-way distinction in the in the 2nd person, and a two-way distinction in the 3rd person.

	1P	2P Honorific	2P Ordinary	2P Familiar	3P Honorific	3P Ordinary
Sg	ami	apni	tumi	tui	tini	she
Pl	amra	apnara	tomra	tora	tāra	tara

Diagram 8: Full form pronouns in Bangla

2.2.5.2 Anaphoric Clitics and Number Marking

Although there are number distinctions in the pronominal full forms across the person values, the number distinctions do not show up morphologically in the anaphoric clitics.

	1P	2P Honorific	2P Ordinary	2P Familiar	3P Honorific	3P Ordinary
Sg	ami kini	apni kinen	tumi kino	tui kinish	tini kinen	she kine
Pl	amra kini	apnara kinen	tomra kino	tora kinish	tāra kinen	tara kinen

Diagram 9: Absence of number agreement on verbs

2.2.5.3 Anaphoric Clitics and TAM Marking:

Bangla draws the distinction between first, second and third person with anaphoric clitics of agreement on the verbal system, with further distinctions in the 2nd and 3rd person in terms of honorifics.

The first-person markers are {~i} in the present form, {~am} in the past form, and {~o} in the future form. The second person has a three ways distinction that varies over the tense forms, with the marker {~en} for all three tense forms in the honorific, the markers {~o} for present and {~e} for past and future forms in the ordinary, and for the familiar form we have the markers {~ish} in the present and {~i} in the past and future forms. The third person honorific patterns the same as the second person honorific with the marker {~en} across all tense forms. For the third person ordinary we have the marker {~e} for the present and future form, and {~o} for the past form.

2.2.6 Valence Marking in Bangla Verbs

Valence is considered to be a combination of a semantic as well as a syntactic notion (Miyagawa 2017, Dixon 200). Semantic valence deals with the number of participants on the discourse stage, while syntactic valence is concerned with the number of arguments in the

clause. Valence is of relevance to the discussion of the morphosyntax of Bangla verbs due to a valence adjusting operation, namely the causative, grammaticalising as a morphological marker.

2.2.6.1 Causative Markers

Causatives are a valence increasing operation. Causative constructions are considered to be the linguistic instantiations of the conceptual notion of causation (Payne 1997). In Bangla the causative grammaticalises as the morphological marker {~a-}and {~i-} and concatenates between the verb root and the TAM marking. The vowel junctures are ameliorated by the approximants /w/ and /y/ respectively. It should be mentioned the underlying marker is {~a-} which assimilates to {~i-} when followed by a front vowel, otherwise it retains its underlying shape.

2.3 A Complete Paradigm of a Regular Bangla Verb with TAM and Pronominal Markers

Tense	Aspect	1P	2PFam	2POrd	2PHon	3POrd	3PHon
Present	Imperfective	kin-i	kin-ish	kin-o	kin-en	kin-e	kin-en
	Progressive	kin-ch-i	kin-ch-ish	kin-ch-o	kin-ch-en	kin-ch-e	kin-ch-en
	Perfect	kin-ech-i	kin-ech-ish	kin-ech-o	kin-ech-en	kin-ech-e	kin-ech-en
Past	Perfective	kin-l-am	kin-l-i	kin-l-e	kin-l-en	kin-l-o	kin-l-en
	Progressive	kin-ch-il-am	kin-ch-il-i	kin-ch-il-e	kin-ch-il-en	kin-ch-il-o	kin-ch-il-en
	Perfect	kin-ech-il-am	kin-ech-il-i	kin-ech-il-e	kin-ech-il-en	kin-ech-il-o	kin-ech-il-en
	Habitual	kin-tam	kin-t-i	kin-t-e	kin-t-en	kin-t-o	kin-t-en
Future	N/A	kin-b-o	kin-b-i	kin-b-e	kin-b-en	kin-b-e	kin-b-en

Diagram 10: Paradigm of Bangla TAM and pronominal markers

24) Summary of Markers:

Verb root = √kin

TAM Markers:
Present imperfective: 1P:{~i}, 2P: {~ish, ~o, ~en}, 3P: {~e, ~en}
Progressive: {~(c)ch-}

Perfect: {~ech-}
Past perfective: {~l-, ~il-}
Past habitual: {~t-}
Future: {~b-}
Pronominal Markers:
1P: Pres- {~i} Past {~am}, Fut-{~o}
2P Fam: Pres-{~ish}, Past+Fut {~i}
2P Ord: Pres-{~o}, Past+Fut-{~e}
2P Hon: {~en}
3P Ord: Pres+Fut-{~e}, Past-{~o}
3P Hon: {~en}

2.3.1 A Verb Paradigm with TAM and Causative Marking

Tense	Aspect	1P	2PFam	2POrd	2PHon	3POrd	3PHon
Present	Imperfective	kin-a-i	kin-a-sh	kin-a-o	kin-a-n	kin-a-e	kin-a-n
	Progressive	kin-a-cch-i	kin-a-cch-ish	kin-a-cch-o	kin-a-cch-en	kin-a-cch-e	kin-a-ach-en
	Perfect	kin-i-ech-i	kin-i-ech-ish	kin-i-ech-o	kin-i-ech-en	kin-i-ech-e	kin-i-ech-en
	Perfective	kin-a-l-am	kin-a-l-i	kin-a-l-e	kin-a-l-en	kin-a-l-o	kin-a-l-en
Past	Progressive	kin-a-cch-il-am	kin-a-cch-il-i	kin-a-cch-il-e	kin-a-cch-il-en	kin-a-cch-il-o	kin-a-cch-il-en
	Perfect	kin-i-ech-il-am	kin-i-ech-il-i	kin-i-ech-il-e	kin-i-ech-il-en	kin-i-ech-il-o	kin-i-ech-il-en
	Habitual	kin-a-tam	kin-a-t-i	kin-a-t-e	kin-a-t-en	kin-a-t-o	kin-a-t-en
Future	N/A	kin-a-b-o	kin-a-b-i	kin-a-b-e	kin-a-b-en	kin-a-b-e	kin-a-b-en

Diagram 11: Paradigm of Bangla TAM and pronominal markers for causatives

25) Causative marker: {~(w)a-}, {~i(y)-}

2.3.2 A verb paradigm with modal marking

2.3.2.1 Imperative marking

Tense	2P Familiar	2P Ordinary	2P Honorific
Present	kin	kin-o	kin-en
Future	kin-b-i	kin-b-e	kin-b-en

Diagram 12: Bangla verbform with imperative markers

As the imperative markers pattern with the 2nd person imperfective markers in the present and future forms, we will merge the two in an effort to avoid ambiguities in computational implementation. The verbform of interest in this case is the 2P familiar present tense form, as it is the only instance of the bare root form of the verb existing freely in a syntactic environment.

2.3.2.2 Jussive marking

Tense	3P Familiar/Ordinary/Honorific	
Present	kin-uk	verb root ending in consonant
	kha-k	verb root ending in vowel

Diagram 13: Bangla verbform with jussive marking

26) Jussive marker: {~(u)k}

3. The Computational Perspective

In broad strokes, computational linguistics or natural language processing (NLP) studies how to build computer programs that can understand, generate, and learn natural language (Jurafsky & Martin 2023, Bird et. al 2009). A subfield of linguistics, computer science, and artificial intelligence -computational linguistics as it is known to linguists and NLP as it is known to computer scientists, focus on getting computers to understand, interpret, generate, and respond to human language. While computational linguistics is considered to be a subfield of linguistics and draws on its theoretical analyses, the computational perspective is vastly different from purely theoretical approaches in analysing languages. Computational linguists focus on building systems that can understand, generate, or translate human language, leveraging data-driven methods and statistical models. In contrast, a pure linguistic theoretic perspective, especially in the generative

tradition, seeks to uncover the underlying rules and structures that govern all human languages, emphasising formal rigor and cognitive plausibility. Theoretical linguistics focuses on understanding language through abstract principles and formal models using primarily rule-based frameworks. Computational linguistics Uses data, algorithms, and statistical models to analyse language focusing on practical output and optimisation even at the cost of theoretical completeness. This results in various approaches that can be at loggerheads with each other in terms of theoretic plausibility and do not port over from one to the other. In this paper, I bridge the gap by proposing a finite state transducer, modeled on the pure theoretic account detailed in the previous sections. I have analysed Bangla verb morphology in terms of TAM, valence, pronominal and honorific markers from a linguistic theoretic perspective and will extend the analysis by modeling a finite state transducer of said verb morphology.

3.1 Finite State Transducers and Morphological Parsing

In order to computationally model the verb morphology of the Bangla matrix verb analysed thus far, we will turn to morphological parsing via a two-level morphological model proposed by Koskenniemi (1983). Two-level morphology represents a word as a correspondence between a lexical level, which represents a simple concatenation of morphemes forming a word, and the surface level, which represents the actual orthographic form of the output word. Morphological parsing is implemented by building mapping rules that map letter sequences on the surface level onto morphemes and feature sequences on the lexical level.

The automaton that we use for performing the mapping between these two levels is the finite state transducer (FST). A finite-state transducer is a finite automaton in which each transition is labeled with both an input and an output symbol and defines a relation between input and output strings (Jurafsky & Martin 2023). An FST

is an abstract computational model used to represent a relation between two sets of symbols, typically used for modeling morphological analysis, phonological rules, and translation tasks in computational linguistics and natural language processing. An FST extends the concept of a finite state automaton (FSA) by adding an output to each transition, allowing it to map between strings rather than just accept or reject them (Beesley & Karttunen 2003). An FST consists of:

- A finite set of states
- An input alphabet and an output alphabet
- A set of transitions (state changes triggered by input symbols with corresponding output symbols)
- One or more start and accept states

Each transition takes an input symbol and produces an output symbol, which can be a string, allowing transformation of inputs into outputs (e.g., mapping a word's surface form to its root form).

A formal definition an FST is given below, based on the Mealy machine extension to a simple finite state automaton FSA (Jurafsky & Martin 2023):

- Q: a finite set of N states, q0, q1,…, qN
- $\sum$: a finite alphabet of complex symbols. Each complex symbol is composed of an input-output pair *i:o*; one symbol *i* from an input alphabet *I*, and one symbol *o* from an output alphabet O, thus $\sum \subseteq IxO$, *I* and *O* may each also include the epsilon symbol ε.
- q0: the start state
- F: the set of final states $F \subseteq Q$
- δ (*q,i:o*): the transition function between states. Given a state $q \in Q$ and complex symbol $i:o \in \sum$, δ (q, i:o) returns a new state $q' \in Q$. δ is thus a relation from $Qx\sum$ to Q

Where an FSA accepts a language stated over a finite alphabet of single symbols, an FST accepts a language stated over pairs of symbols. In two-level morphology, the pairs of symbols are also called feasible pairs.

As mentioned earlier, the two-level morphology views an FST as having two tapes. The upper or lexical tape is composed of characters from the left side of the pairs; the lower or surface tape, is composed of characters from the right side of the pairs. Hence, each symbol in the transducer alphabet expresses how the symbol from one tape is mapped to the symbol on another tape.

3.1.1 An FST for Bangla Matrix Verbs

We are now ready to build an FST morphological parser out of our earlier, morphosyntactic modeling of the Bangla verb or simplex predicate. In order to do this, we will have the two-level tapes, the lexical tape and the surface tape with morphological features that correspond to the morphological markers we have delineated thus far. The proposed morphological parser needs to map multiple forms, this will be done by cascading the lexicon. Cascading two automata means running them in series with the output of the first feeding the input to the second (Beesley & Karttunen 2003, Jurafsky & Martin 2023).

3.1.2 Parse tables for Bangla Matrix Verbs

Based on the morphological markers we have analysed in the previous sections, we can create the following parse tables for our verbforms according to person:

1. 1P

Input	Markers	Morphological Parse	Gloss
kini	kin-i	kin+V+1Ppres	buy.Pres.1P
kinchi	kin-ch-i	kin+V+Prog+1Ppres	buy.Prog.Pres.1P
kinechi	kin-ech-i	kin+V+Perf+1Ppres	buy.Perf.Pres.1P
kinlam	kin-l-am	kin+V+Past+1Ppast	buy.Past.1P
kinchilam	kin-ch-il-am	kin+V+Prog+Past+1Ppast	buy.Prog.Past.1P
kinechilam	kin-ech-il-am	kin+V+Perf+Past+1Ppast	buy.Perf.Past.1P
kintam	kin-t-am	kin+V+PastHab+1Ppast	buy.Hab.1P
kinbo	kin-b-o	kin+V+Fut+1Pfut	buy.Fut.1P

Diagram 14: Parse table for 1P matrix verbs

2. 1P Causative

Input	Markers	Morphological Parse	Gloss
kinai	kin-a-i	kin+V+Caus+1pres	buy.Caus.Pres.1P
kinacchi	kin-a-cch-i	kin+V+Caus+Prog+1Ppres	buy. Caus.Prog.Pres.1P
kiniechi	kin-i-ech-i	kin+V+ Caus+Perf+1Ppres	buy. Caus.Perf.Pres.1P
kinalam	kin-a-l-am	kin+V+ Caus+Past+1Ppast	buy. Caus.Past.1P
kinacchilam	kin-a-cch-i-lam	kin+V+Caus+Prog+1Ppast	buy. Caus.Prog.Past.1P
kiniechilam	kin-i-ech-il-am	kin+V+ Caus+Perf+Past+1Ppast	buy. Caus.Perf.Past.1P
kinatam	kin-a-t-am	kin+V+ Caus+PastHab+1Ppast	buy. Caus.Hab.1P
kinabo	kin-a-b-o	kin+V+ Caus+Fut+1Pfut	buy. Caus.Fut.1P

Diagram 15: Parse table for 1P causative verbs

3. 2P Ordinary

Input	Markers	Morphological Parse	Gloss
kino	kin-o	kin+V+2PPresOrd	buy.Pres.2POrd
kincho	kin-ch-o	kin+V+Prog+2PpresOrd	buy.Prog.Pres.2POrd
kinecho	kin-ech-o	kin+V+Perf+2PpresOrd	buy.Perf.Pres.2POrd
kinle	kin-l-e	kin+V+Past+2PpastOrd	buy.Past.2POrd
kinchile	kin-ch-il-e	kin+V+Prog+Past+2PpastOrd	buy.Prog.Past.2POrd
kinechile	kin-ech-il-e	kin+V+Perf+Past+2PpastOrd	buy.Perf.Past.2POrd
kinte	kin-t-e	kin+V+PastHab+2PpastOrd	buy.Hab.2POrd
kinbe	kin-b-e	kin+V+Fut+2PfutOrd	buy.Fut.2POrd

Diagram 16: Parse table for 2P ordinary verbs

4. 2P Ordinary Causative

Input	Markers	Morphological Parse	Gloss
kinao	kin-a-o	kin+V+Caus+2PPresOrd	buy.Caus.Pres.2POrd
kinaccho	kin-a-cch-o	kin+V+Caus+Prog+2PpresOrd	buy.Caus.Prog.Pres.2POrd

kiniecho	kin-i-ech-o	kin+V+Caus+Perf+2PpresOrd	buy.Caus.Perf.Pres.2POrd
kinale	kin-a-l-e	kin+V+Caus+Past+2PpastOrd	buy.Caus.Past.2POrd
kinacchile	kin-a-cch-il-e	kin+V+Caus+Prog+Past+2PpastOrd	buy.Caus.Prog.Past.2POrd
kiniechile	kin-i-ech-il-e	kin+V+Caus+Perf+Past+2PpastOrd	buy.Caus.Perf.Past.2POrd
kinate	kin-a-t-e	kin+V+Caus+PastHab+2PpastOrd	buy.Caus.Hab.2POrd
kinabe	kin-a-b-e	kin+V+Caus+Fut+2PfutOrd	buy.Caus.Fut.2POrd

Diagram 17: Parse table for 2P ordinary with causative

5. 2P Honorific

Input	Markers	Morphological Parse	Gloss
kinen	kin-en	kin+V+2PPresHon	buy.Pres.2PHon
kinchen	kin-ch-en	kin+V+Prog+2PpresHon	buy.Prog.Pres.2PHon
kinechen	kin-ech-en	kin+V+Perf+2PpresHon	buy.Perf.Pres.2PHon
kinlen	kin-l-en	kin+V+Past+2PpastHon	buy.Past.2PHon
kinchilen	kin-ch-il-en	kin+V+Prog+Past+2PpastHon	buy.Prog.Past.2PHon
kinechilen	kin-ech-il-en	kin+V+Perf+Past+2PpastHon	buy.Perf.Past.2PHon
kinten	kin-t-en	kin+V+PastHab+2PpastHon	buy.Hab.2PHon
kinben	kin-b-en	kin+V+Fut+2PfutHon	buy.Fut.2PHon

Diagram 18: Parse table for 2P honorific verbs

6. 2P Honorific Causative

Input	Markers	Morphological Parse	Gloss
kinan	kin-a-n	kin+V+Caus+2PPresHon	buy.Caus.Pres.2PHon
kinacchen	kin-a-cch-en	kin+V+Caus+Prog+2PpresHon	buy.Caus.Prog.Pres.2PHon
kinieche n	kin-i-ech-en	kin+V+Caus+Perf+2PpresHon	buy.Caus.Perf.Pres.2PHon
kinalen	kin-a-l-en	kin+V+Caus+Past+2PpastHon	buy.Caus.Past.2PHon

kinacchilen	kin-a-cch-il-en	kin+V+Caus+Prog+Past+2PpastHon	buy.Caus.Prog.Past.2PHon
kiniechilen	kin-i-ech-il-en	kin+V+Caus+Perf+Past+2PpastHon	buy.Caus.Perf.Past.2PHon
kinaten	kin-a-t-en	kin+V+Caus+PastHab+2PpastHon	buy.Caus.Hab.2PHon
kinaben	kin-a-b-en	kin+V+Caus+Fut+2PfutHon	buy.Caus.Fut.2PHon

Diagram 19: Parse table for 2P honorific with causative

7. 2P Familiar

Input	Markers	Morphological Parse	Gloss
kinish	kin-ish	kin+V+2PPresFam	buy.Pres.2PFam
kinchish	kin-ch-ish	kin+V+Prog+2PpresFam	buy.Prog.Pres.2PFam
kinechish	kin-ech-ish	kin+V+Perf+2PpresFam	buy.Perf.Pres.2PFam
kinli	kin-l-i	kin+V+Past+2PpastFam	buy.Past.2PFam
kinchili	kin-ch-il-i	kin+V+Prog+Past+2PpastFam	buy.Prog.Past.2PFam
kinechili	kin-ech-il-i	kin+V+Perf+Past+2PpastFam	buy.Perf.Past.2PFam
kinti	kin-t-i	kin+V+PastHab+2PpastFam	buy.Hab.2PFam
kinbi	kin-b-i	kin+V+Fut+2PfutFam	buy.Fut.2PFam

Diagram 20: Parse table for 2P familiar

8. 2P Familiar Causative

Input	Markers	Morphological Parse	Gloss
kinash	kin-a-sh	kin+V+Caus+2PPresFam	buy.Caus.Pres.2PFam
kinacchish	kin-a-cch-ish	kin+V+Caus+Prog+2PpresFam	buy.Caus.Prog.Pres.2PFam
kiniechish	kin-i-ech-ish	kin+V+Caus+Perf+2PpresFam	buy.Caus.Perf.Pres.2PFam
kinali	kin-a-l-i	kin+V+Caus+Past+2PpastFam	buy.Caus.Past.2PFam
kinacchili	kin-a-cch-il-i	kin+V+Caus+Prog+Past+2PpastFam	buy.Caus.Prog.Past.2PFam
kiniechili	kin-i-ech-il-i	kin+V+Caus+Perf+Past+2PpastFam	buy.Caus.Perf.Past.2PFam
kinati	kin-a-t-	kin+V+Caus+PastHab+2PpastFam	buy.Caus.Hab.2PFam

	i		
kinabi	kin-a-b-i	kin+V+Caus+Fut+2PfutFam	buy.Caus.Fut.2PFam

Diagram 21: Parse table for 2P familiar with causative

9. 2P Familiar Imperative

Input	Markers	Morphological Parse	Gloss
kin	kin	kin+V+Imp+2PFam	buy.Imp.Pres.2PFam

Diagram 22: Parse table for 2P familiar with imperative

10. 3P Ordinary

Input	Markers	Morphological Parse	Gloss
kine	kin-e	kin+V+3Ppres	buy.Pres.3POrd
kinche	kin-ch-e	kin+V+Prog+3Ppres	buy.Prog.Pres. 3POrd
kineche	kin-ech-e	kin+V+Perf+3Ppres	buy.Perf.Pres. 3POrd
kinlo	kin-l-o	kin+V+Past+3Ppast	buy.Past. 3POrd
kinchilo	kin-ch-il-o	kin+V+Prog+Past+3Ppast	buy.Prog.Past. 3POrd
kinechilo	kin-ech-il-o	kin+V+Perf+Past+3Ppast	buy.Perf.Past. 3POrd
kinto	kin-t-o	kin+V+PastHab+3Ppast	buy.Hab. 3POrd
kinbe	kin-b-e	kin+V+Fut+3Pfut	buy.Fut. 3POrd

Diagram 23: Parse table for 3P ordinary

11. 3P Ordinary Causative

Input	Markers	Morphological Parse	Gloss
kinae	kin-a-e	kin+V+Caus+3Ppres	buy.Caus.Pres.3POrd
kinieche	kin-i-ech-e	kin+V+Caus+Perf+3Ppres	buy.Caus.Perf.Pres. 3POrd
kinalo	kin-a-l-o	kin+V+Caus+Past+3Ppast	buy.Caus.Past. 3POrd
kiniechilo	kin-i-ech-il-o	kin+V+Caus+Perf+Past+3Ppast	buy.Caus.Perf.Past. 3POrd
kinato	kin-a-t-o	kin+V+Caus+PastHab+3Ppast	buy.Caus.Hab. 3POrd
kinabe	kin-a-b-e	kin+V+Caus+Fut+3Pfut	buy.Caus.Fut. 3POrd

Diagram 24: Parse table for 3P ordinary with causative

12. 3P Honorific

Input	Markers	Morphological Parse	Gloss
kinen	kin-en	kin+V+3PHon	buy.Pres.3PHon
kinchen	kin-ch-en	kin+V+Prog+3PHon	buy.Prog.Pres.3PHon
kinechen	kin-ech-en	kin+V+Perf+3PHon	buy.Perf.Pres.3PHon
kinlen	kin-l-en	kin+V+Past+3PHon	buy.Past.3PHon
kinchilen	kin-ch-il-en	kin+V+Prog+Past+3PHon	buy.Prog.Past.3PHon
kinechilen	kin-ech-il-en	kin+V+Perf+Past+3PHon	buy.Perf.Past.3PHon
kinten	kin-t-en	kin+V+PastHab+3PHon	buy.Hab.3PHon
kinben	kin-b-en	kin+V+Fut+3PHon	buy.Fut.3PHon

Diagram 25: Parse table for 3P honorific causative

13. 3P Honorific Causative

Input	Markers	Morphological Parse	Gloss
kinan	kin-a-n	kin+V+Caus+3PHon	buy.Caus.Pres.3PHon
kiniechen	kin-i-ech-en	kin+V+Caus+Perf+3PHon	buy.Caus .Perf.Pres.3PHon
kinalen	kin-a-l-en	kin+V+Caus+Past+3PHon	buy.Caus.Past.3PHon
kiniechilen	kin-i-ech-il-en	kin+V+Caus+Perf+Past+3PHon	buy.Caus.Perf.Past.3PHon
kinaten	kin-a-t-en	kin+V+Caus+PastHab+3PHon	buy.Caus.Hab.3PHon
kinaben	kin-a-b-en	kin+V+Caus+Fut+3PHon	buy.Caus.Fut.3PHon

Diagram 26: Parse table for 3P honorific with causative

14. 3P Jussive

Input	Markers	Morphological Parse	Gloss
kinuk	kin-uk	kin+V+Jus	buy.Jus.3P

Diagram 27: Parse table for 3P Jussive

Here, the tags for morphological markers are in a 1:1 relationship in order to avoid redundancies. As there are overt morphological markers for the past and future tense there are overt tense tags, however, as

the present tense is the unmarked form there is no extra tag. If a past or future tag is absent then the verbform is in the present form. For pronominal markers in the same person category that change according to tense I have kept tense information in the person marker tag. Since the honorific forms in the second and third person remain the same across the tense forms, a single uniform tag has been kept for all forms. The verb tag i.e. '+V' maps to an empty parse marked by an epsilon as shown in the next section.

3.1.3 A Sample Parse

In this section we build a Finite State Transducer (FST) morphological parser and exemplify a sample parse of the verbform 'kinechilen' [buy.Perf.Past.3P.Hon], based on the morphological markers analysed thus far along with the parse tables in §3.1.2.

To build a FST that maps the morphological parse [kin+V+Perf+Past+3PHon] to the surface form [kin+ech+il+en], we need to define how the abstract morphological features are realised as surface morphemes (§3.1.3.1).

3.1.3.1 Conceptual Representation of a Finite State Transducer (FST)

The FST comprises a lexical tape and the appropriate morphological features that correspond to each morpheme and map to the surface form. These features also map to the empty string ε (diagram 25) since there is no segment corresponding to them. We can represent the mapping in diagram using a series of transitions. A simplified version of said transducer is given below:

27) Lexical Tape → Surface Tape Mapping

Lexical Morpheme	Surface Form
kin	kin
+V	ε (not realised)
+Perf	ech
+Past	il
+3PHon	en

Diagram 28: lexical tape to surface mapping of "kinechilen"

Diagram-28 presents a morpheme-by-morpheme correspondence, where the final surface string kin+ech+il+en reflects sequential realisation of the features.

- Lexical tape: kin+V+Perf+Past+3PHon
- Surface tape: kinechilen

Here:

- ε denotes deletion (i.e. the symbol is not pronounced).
- echilen is the realisation of the morphological features.

3.1.3.4 Conceptual Transitions for FST

A sequence of states with transitions that output the surface string from the lexical input is conceptualised in this section. The state transitions of the verb-form “kinechilen” is given below:

- State 0 → 1: k/k
- State 1 → 2: i/i
- State 2 → 3: n/n
- State 3 (self-loop): +V/ε
- State 3 → 4: +Perf/ech
- State 4 → 5: +Past/il
- State 5 → 6: +3PHon/en
- State 6: Final state

In summary, this FST maps:

- Lexical form: kin+V+Perf+Past+3PHon
- Surface form: kinechilen

by:

- Keeping ‘kin’ unchanged
- Fusing morphological tags into the string ‘echilen’

The proposed morphological parser needs to map surface forms to lexical forms with the features delineated. This can be done by cascading two automata i.e. running them in series with the output of the first feeding the input to the second. Building on Jurafsky Martin’s (2023) transducer for nominal forms, we would first represent the lexicon of Bangla verb stems using the FST T_{STEMS}

(diagram 26), the output of this FST would then feed the TAM automaton T_{TAM}.

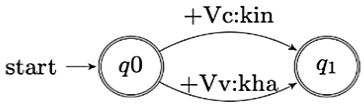

Diagram 29: A sample FST for T_{STEMS} that maps verb-roots with consonantal or vocalic endings.

The two transducers can also be composed with a composition algorithm i.e. taking a cascade of two transducers with many different level inputs and outputs and converting them into a single two-level transducer with one input tape and one output tape. The resulting composed automaton will be $T_{VERB} = T_{TAM}$ o T_{STEMS}. This transducer will map the verb root and the tam markers, thus a surface form of 'kinechilen' will map to 'kin+V+Perf+Past+3Phon' as follows (28):

28) kin:kin +V:ɛ +Perf:ech +Past:il +3Phon:en#

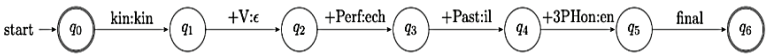

Diagram 30: A simple FST for the verbform 'kinechilen'

The architecture of the FST is a two-level cascade of transducers i.e. a set of transducers in series with the output of one transducer acting as the input of another transducer (Jurafsky & Martin 2023). Cascades can be of arbitrary depth allowing for intermediate levels, with each level built out of many individual transducers, a feature of the architecture which can be utilised in the inclusion of parses based on Bangla orthographic rules as explored in the following section.

3.2 Bangla Orthography and Finite State Transducers

The analysis thus far has been neatly concatenative as it is based on the phonological makeup of Bangla as shown in the Romanised representations, which is bears the concatenative nature of the

Latin alphabet itself. However, the Bangla script, a descendant of Brahmi is an alphasyllabary or abugida orthographic system (Ethnologue 2025, Daniels 2008) and comprises vowel diacritics that can precede or follow the consonant in script form (29).

29) কিনেছিলেন > কি (i+k=ki) নে (e+n=ne) ছি (i+ch=chi) লে (e+l=le) ন (n)

However, in the computational representation of Bangla, often it is typed in concatenative CV mode according to the phonological makeup and then rendered into the correct ordering. Hence, an unrendered intermediate layer can serve to map to the concatenative input form in a CV pattern (30-31).

30) কিনেছিলেন > ক-ি-ন-ে-ছ-ি-ল-ে-ন
31) ক (k) + ি (i) + ন (n) + ে (e) + ছ (ch) + ি (i) + ল (l) + ে (e) + ন (n) > k+i+n+e+ch+i+l+e+n > kinechilen

3.2.1 Orthographic Rules:

A simplified example of sample orthographic rules for the aforementioned verb is given below:

Name	Description of Rules	Example
Consonant followed by ‘ি’ vowel diacritic	If a consonant is followed by a ‘ি’ vowel diacritic with delimiters on both sides the vowel sign will render to the left of the consonant	ক-ি > কি
Consonant followed by ‘ে’ vowel diacritic	If a consonant is followed by a ‘ে’ vowel diacritic with delimiters on both sides the vowel sign will render to the left of the consonant	ন-ে > নে
Consonant followed by ‘া’ vowel diacritic	If a consonant is followed by a ‘া’ vowel diacritic with delimiters on both sides the vowel sign will render to the right of the consonant	খ-া > খা

Diagram 31: simplified orthographic rules for FST

Taking this into account, the FST can have the following levels:

32) Lexical tape: kin+V+Perf+Past+3PHon

Intermediate tape 1: kin+ech+il+en

Intermediate tape 2: ক-ি-ন+ে-ছ+ি-ল+ে-ন

Surface tape: কিনেছিলেন

4. Limitations and Future Work

Conceptually modeling a finite state transducer (FST) for Bangla verbs provides a valuable abstraction of the morphosyntax of the language. However, such modeling has inherent limitations without hands-on implementation. The conceptual modeling proposed here is for matrix verbs and can be extended to light verbs of complex predicates. However, without implementation on balanced corpora that account for all forms, it is difficult to test the coverage, accuracy, or scalability of the model across verb classes and exceptions. Manual modeling also risks oversimplifying the data without taking into account the linguistic irregularities of structure. Having acknowledged these limitations, I contend that the conceptual modeling provided in this paper will serve as an integral steppingstone to the implementation of morphological parsers grounded in the rigors of linguistic theory. Computational analysis and implementation often gloss over the integrity of linguistic theoretic analyses. This paper bridges the gap by grounding the conceptual model of the FST in rigorous morphosyntactic modeling. Finite state transducers remain essential tools in natural language processing, especially for tasks like morphological analysis, parsing, generation, and spell-checking. Their efficiency, formal rigor, and compatibility with regular language operations make them ideal for modeling morphologies like Bangla. Future work will involve implementing the transducer using tools which will allow for systematic evaluation, refinement, and integration into practical NLP systems.

5. Conclusion:

One of the most widely spoken languages in the world, Bangla is also one of the most digitally under-resourced languages. Due to its rich morphology, Bangla comprises simplex and complex predicates with concatenative inflections, making it well-suited for computational modeling. This paper provides a detailed examination of Bangla verb morphology from a computational perspective. We first present an overview of the structural characteristics of Bangla verb forms, highlighting key inflectional patterns, analysing the plethora of tense, aspect, mode, valence and pronominal markers as well as their combinatorial ordering. We then created parse tables based on the morphological markers identified and proceed to create a conceptual finite state transducer for morphological parsing of Bangla verbs. This paper presents a detailed computational analysis of Bangla matrix verbs or simplex predicates, aiming to bridge the gap between theoretical linguistics and practical language technology. We propose a computational framework that integrates linguistic insights with algorithmic strategies to improve morphological analysis and generation for Bangla verbs. The aim of this research work contributes to both the theoretical understanding of Bangla morphology and the practical advancement of NLP tools for low-resource languages. The goal of such an endeavour is to develop a more accurate and linguistically informed model of Bangla predicate morphology that can serve both descriptive and applied purposes.

References

Abdullah, M. M., Islam, M. Z., & Khan, M. (2007). Error-tolerant finite-state recognizer and string pattern similarity based spelling-checker for Bangla. In *Proceeding of 5th international conference on natural language processing (ICON)*.

Alam, F., Chowdhury,S.A. and Noori, S.R.H.(2016). Bidirectional LSTMs—CRFs networks for Bangla POS tagging. In 2016 19th International Conference on Computer and Information Technology (ICCIT). IEEE, 377–382.

Alam, F., Nath,P.K and Khan,M. (2007). Text to speech for Bangla language using festival. In Proc.1st Intl. Conf. on Digital Comm. and Computer Applications, Vol. 1. 853–859.

Alam,M., UzZaman,N. and Khan,M. (2007). N-gram based statistical grammar checker for Bangla and English. Technical Report. BRAC University.

Alam, T., Khan, A., and Alam, F. (2020). Bangla Text Classification using Transformers. arXiv preprint arXiv:2011.04446 (2020).

Beesley, K. R., & Karttunen, L. (2003). *Finite-State Morphology*. CSLI Publications.

Bhattacharjee, A., Hasan,T., Samin,K., Rahman,M.S., Iqbal, A and Shahriyar, R. (2021). BanglaBERT: Combating Embedding Barrier for Low-Resource Language Understanding. arXiv preprint arXiv:2101.00204 (2021). arXiv:cs.CL/2101.00204

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

Butt, M. (2010). The light verb jungle: Still hacking away. *Complex predicates: Cross-linguistic perspectives on event structure*, 48-78.

Chatterji, S. K. (2024). *The Origin and Development of the Bengali Language: Volume One*. Routledge.

Chatterji, S. K. (2024). *The Origin and Development of the Bengali Language: Volume Two*. Taylor & Francis.

Chatterji, S. K. (1926). *The Origin and Development of the Bengali Language: Part Two*. Calcutta University Press.

Chaudhuri,B.B., and Pal, U. (1998). A complete printed Bangla OCR system. Pattern recognition 31, 5 (1998), 531–549.

Chowdhury,S.A., Alam,F., and Khan,N (2018). Towards Bangla named entity recognition. In 2018 21st International Conference of Computer and Information Technology (ICCIT). IEEE, 1–7.

Comrie, B. (1985). *Tense* (Vol. 17). Cambridge university press.

Comrie, B. (1981). Aspect and voice: some reflections on perfect and passive. In *Tense and aspect* (pp. 65-78). Tedeschi, P., & Zaenen, A. (Eds.). (13 Jan. 2020). Leiden, The Netherlands: Brill.

Daniels, P. T. (2008). Writing systems of major and minor languages. In B. B. Kachru, Y. Kachru, & S. N. Sridhar (Eds.), *Language in South Asia* (pp. 285–308). chapter, Cambridge: Cambridge University Press.

Das, A. and Bandyopadhyay, S.(2010). SentiWordNet for Bangla. Knowledge Sharing Event-4: Task 2 (2010), 1–8.

Das, D., Pal, S., Mondal, T., Chakraborty, T., & Bandyopadhyay, S. (2010, August). Automatic extraction of complex predicates in Bengali. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications* (pp. 37-45).

Dasgupta, Probal. (1977) "The internal grammar of compound verbs in Bangla." *Indian linguistics* 38, no. 2 (1977): 68-85.

Dasgupta, S. and Khan, M. (2004). Morphological parsing of Bangla words using PC-KIMMO.

Dasgupta, S., Khan, N., Sarkar, A.I., Pavel, D.H.S. and Khan, M. (2005). Morphological analysis of inflecting compound words in Bangla.

Dixon, R. M. (2000). A typology of causatives: form, syntax and meaning. *Changing valency: Case studies in transitivity*, *30*, 83.

Hasan, A., Alam,F., Chowdhury,S.A., and Khan,N. (2019). Neural vs Statistical Machine Translation: Revisiting the Bangla-English Language Pair. In 2nd International Conference on Bangla Speech and Language Processing (ICBSLP) (2019-01-01).

Hasan,A., Tajrin,J., Chowdhury,S.A., and Alam, F. (2020). Sentiment Classification in Bangla Textual Content: A Comparative Study. In 2020 23rd International Conference on Computer and Information Technology (ICCIT). 1–6.

Hasan, K.M.A, Islam,S. Elahi,G.M.M. and Izhar, M.N. (2013). Sentiment recognition from Bangla text. In Technical Challenges and Design Issues in Bangla Language Processing. IGI Global, 315–327.

Hasnat, M.A., Habib,S.M.M and Khan,M. (2008). A high performance domain specific OCR for Bangla script. In Novel algorithms and techniques in telecommunications, automation and industrial electronics. Springer,174–178.

Islam, M.S. (2009). Research on Bangla language processing in Bangladesh: progress and challenges. In 8th international language & development conference. 23–25.

Jain, D., & Cardona, G. (2007). *The Indo-Aryan Languages*. Routledge.

Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing* (3rd ed. draft). Stanford University.

Koskenniemi, K. (1983). Two-level model for morphological analysis. In *IJCAI* (Vol. 83, pp. 683-685).

Kachru, Y., & Pandharipande, R. (1980). Towards a typology of compound verbs in South Asian languages. *Studies in the Linguistic Sciences*, *10*(1), 113-124.

Masica, C. P. (1993). *The indo-aryan languages*. Cambridge University Press.

Miyagawa, S. (2017). Causatives. *The handbook of Japanese linguistics*, 236-268.

Morshed, A. K. M. (2004). *Adhunik bhashatattva*. Dhaka: Mowla Brothers.

Mosaddeque, A.B. and Haque,N. (2004). Context-Free Grammar for Bangla. Technical Report. BRAC University.

Paul, A.K., Das, D. and Kamal,M.M. (2009). Bangla speech recognition system using LPC and ANN. In 2009 Seventh International Conference on Advances in pattern recognition. IEEE, 171–174.

Paul, S. (2003). *Composition of compound verbs in Bangla*. In D. Beermann & L. Hellan (Eds.), *Proceedings of the Workshop on Multi-Verb Constructions* (pp. 1–10). Trondheim: Norwegian University of Science and Technology.

Paul, S. (2008). The Semantics of Bangla Compound Verbs. In *2004* (pp. 101-112). Berlin, New York: De Gruyter Mouton. https://doi.org/10.1515/9783110179897.101

Payne, T. E. (1997). *Describing morphosyntax: A guide for field linguists*. Cambridge University Press.

Saetbyol,S. (2017). *(The) Syntax of Jussives* (Doctoral dissertation, 서울대학교 대학원).

Sengupta, P. (1993). On lexical and syntactic processing of Bangla language by computer. Ph.D. Dissertation. Indian Statistical Institute, Kolkata.

Sengupta, P. and Chaudhuri, B.B. (1993). A morpho-syntactic analysis based lexical subsystem. International journal of pattern recognition and artificial intelligence 7, 03 (1993), 595–619.

Sennrich, R, Haddow, B and Birch, A. 2016. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 1715–1725.

Shaw, R. (1984). *সাধারণ ভাষাবিজ্ঞান ও বাংলা ভাষা* [*Sadharan Bhasabigyan O Bangla Bhasha*]. Kolkata: Pustak Bipani.

Sultana, A. (2016). Description of verb morphology in colloquial Bangla. BRAC University Journal, 11(1), 69–77.

Wallace, S. (1982). Figure and Ground: The Interrelationships of Linguistic Categories. In P. Hopper (Ed.), *Tense-Aspect: Between semantics & pragmatics* (pp. 201-224). John Benjamins Publishing Company.

Zanuttini, R., Pak, M., & Portner, P. (2012). A syntactic analysis of interpretive restrictions on imperative, promissive, and exhortative subjects. *Natural Language & Linguistic Theory*, *30*(4), 1231-1274.