A Corpus-based Study of করে (*kare*) in Bangla: Theoretical and Computational Perspectives

Priyanka Biswas[1] **Sandipan Dandapat**[2] **Kalika Bali**[2] **Monojit Choudhury**[2]
[1]Central Institute of Indian Languages, Mysore, India
biswas.priyanka@gmail.com
[2]Microsoft Research Lab India, Bangalore
{v-sandan,kalikab,monojitc}@microsoft.com

Abstract

করে (*kare*) is one of the most frequent words in Bangla corpus, which exhibits various morpho-syntactic functions. Morphologically, the word can be analyzed as a noun (meaning "hand", "tax", etc.) with a locative case-marker, a finite verb (meaning "do" or "does") as well as a non-finite form of "do" (meaning "having done"). However, owing to various functional modifications this particular lexical item can also be used as a postposition and particle. Due to this extremely variable behavior, the lexical item '*kare*' is problematic for several NLP tasks and, therefore, calls for a special treatment. In this paper we investigate the various distributions and functions of '*kare*' and identify eleven basic morpho-syntactic categories covering these various functions. On the basis of diachronic and synchronic evidence, we show how these various functions of '*kare*' can be explained by positing etymological homomorphism and/or functional diversification. Further, we propose suggestions for dealing with '*kare*' during morphological analysis, parts-of-speech tagging, chunking and other advanced NLP tasks.

1 Introduction

It is a well known fact to those who work in NLP that the most frequent lexical items of a language are also the most ambiguous ones. Appropriate disambiguation of the senses or functions of these items is often a hard task, which unless solved with a high accuracy, hampers the overall performance of the NLP system. This paper reports a theoretical study of one such highly frequent and extremely ambiguous lexical item of Bangla – করে (*kare*). Despite the fact that '*kare*' is derived from the root '*kar*' by addition of the suffix '-*e*', the word can be morphologically analyzed in the following three ways[1]:

(a) noun ('*kar*' meaning "hand", "tax", etc.) with a locative case marker ('-*e*');
(b) finite verb ('*karA*' meaning "to do") in simple present tense, third person;
(c) non-finite verb or conjunctive participle (meaning "having done")

However, as we shall see subsequently, this single lexical item can also function as an adverbializer, a particle and a postposition. The suffix '-*e*' in Bangla produces both finite and non-finite forms and is very productive with other verbal roots as well, but the extremely divergent behavior of '*kare*' is an exception.

'*kare*' is the second most frequent lexical item in the EMILLE corpus[2] consisting of about 3 million words and the fourth most frequent word in 17 million word *Anandabazar* corpus[3]. Thus, on one hand the extremely divergent function of '*kare*' renders it as one of the most ambiguous, and consequently, "harder-to-handle" lexical items in Bangla NLP; on the other hand, owing to its very high frequency, it becomes even more important to appropriately disambiguate the various functions of '*kare*' at different levels of linguistic analyses and NLP tasks. These facts lead us to believe that the lexical item '*kare*' calls for an in-depth linguistic analysis and a special treatment in NLP.

The case of '*kare*' in Bangla is comparable to that of prepositions such as 'to', 'in' and "on", and pronouns such as "it" in English. All of these lexical items are highly frequent and extremely

[1] In this paper we use the ITRANS convention (Chopde, 2001) to transcribe Bangla script.

[2] http://www.lancs.ac.uk/fass/projects/corpus/emille/

[3] Anandabazar patrika (http://www.anandabazar.com/) is a Bangla daily. The corpus was built by crawling the online articles published from 1st Jan 2001 to 31st Dec 2002

Proceedings of ICON-2008: 6th International Conference on Natural Language Processing
Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2008

ambiguous in the language. Therefore, it is not surprising that several in-depth linguistic as well as computational linguistic studies have been devoted to these special lexical items. For instance, Kelly (2002) studies the distribution of the prepositions ‘in’, ‘on’ and ‘under’ from the perspective of language acquisition, while numerous studies have been carried out on the historical development of the auxiliary function of the verb ‘do’ (see, for example, Iyeiri (2002) and references therein). In NLP, often special disambiguation modules are developed for specific lexical items. Examples include resolution between referential, existential and particle usages of ‘it’ (Bergsma et al., 2008*;* Boyd et al., 2005; Evans, 2001; Müller, 2007), inclusive and exclusive usages of ‘you’ in discourse (Gupta et al., 2007) and word by word sense disambiguation modules (Paice and Husk, 1987). Also worth noting is the fact that in the POS annotation guidelines of Penn Treebank (Santorini, 1990), the lexical item ‘to’ is annotated as ‘\TO’ regardless of whether it is a preposition or an infinitival marker. Such arrangements might help reduce errors during human and machine POS annotation, pushing the disambiguation task to subsequent NLP modules such as chunking or parsing. However, we do not know of any work on the various morpho-syntactic functions of ‘*kare*’, or for that matter any other lexical item in Bangla from a linguistic perspective, let alone from that of computational linguistics.

In this paper we investigate the various distributions and functions of ‘*kare*’ in Bangla and based on our observations make specific recommendations for the treatment of the word during some basic NLP tasks. More precisely, the aims of the present work are:

- identification and classification of the distinct morpho-syntactic functions of the word ‘*kare*’ through a corpus based analysis;
- linguistic explanations for the observed divergent functionalities of the word based on diachronic and synchronic evidence;
- recommendations for the treatment of ‘*kare*’ during morphological analysis, POS annotation, shallow parsing and parsing.

Note that ‘*kare*’ has multiple senses, as many as 53 have been reported in the literature (Bandyopadhyay, 1966), both as a verb and a noun. The scope of the present work, however, is only limited to a morpho-syntactic analysis of ‘*kare*’. The semantic analysis and automatic sense disambiguation of the word are equally challenging and interesting problems that we plan to address in the future.

The rest of the paper is organized as follows. Section 2 enumerates different types of distribution and function of ‘*kare*’ as observed in the corpus. Linguistic explanations of the various functions are presented in Section 3. Section 4 discusses the recommendations for the treatment of ‘*kare*’ during various NLP tasks. Section 5 concludes the paper by summarizing the work and discussing future steps.

2 Functions and Distributions of ‘*kare*’

In order to study the various distributions of ‘*kare*’, we refer to two distinct sources – dictionaries and text corpora. Dictionaries are similar to secondary data that provide us the traditional classification of the morpho-syntax and semantics of the word, often backed by etymological evidence. Text corpora, on the other hand, are similar to primary data and are especially useful from the perspective of NLP. Although this work is based on primary evidence obtained from the corpus, we begin the discussion with the lexicographers’ perspective on ‘*kare*’.

2.1 ‘*kare*’ in Traditional Bangla Lexicon

The *BangiYa shabdakoSha* (Bandyopadhyay, 1966) has no entry dedicated to ‘*kare*’. However, it enlists 53 different meanings of the verb root ‘*karA*’ on the basis of different complex predicates. Thus, some of the non-verbal morpho-syntactic functions of the word are explained by postulating appropriate semantic senses. For instance, the use of ‘*kare*’ as a postposition in the sentence ‘*gA.Di kare gelAma*’ (went by a/the car) is accounted for by positing that ‘*karA*’ can also mean “to ride”.

The *Samsad Bangla dictionary* (Biswas, 1957), however, has a specific entry for ‘*kare*’ where the various non-verbal usages of the word are also listed. Apart from the conjunctive participle (non-finite verb) sense of the word, the dictionary mentions that ‘kare’ can also be used as an *avyaya* (indeclinable) in the following four distinct ways (senses):

(a) by, with the help of; e.g., *hAte kare* (by hand), *mukhe kare* (by mouth)

(b) think, analyze; e.g., *eka TAkA kare chA.MdA* (one rupee donation [per head]), *doSha kama kare dekhi* ([I] consider the mistakes to be less serious)

(c) in the way of; e.g., *ki kare e kAja karale* (in what way [how] have you done this?), *bhAlo kare khAo* (eat in a good way)

(d) one after another, sequence; e.g., *ekajana-ekajana kare yAo* (go one by one).

As we shall discuss shortly, the usage listed in (a) and (c) above are similar to that of a postposition and adverbializer respectively; in (d) and the first example of (b), ‘*kare*’ has been used as a particle which bears the sense of iterative/distributive action. The second example of (b) also uses ‘*kare*’ as a particle, but with a sense of *hedging* (approximation). Therefore, we are at a loss in explaining why in Biswas (1957) the particle use of ‘*kare*’ for two very different senses (approximation and iterative/distributive) has been clubbed together under the banner of ি◌বেবচনা(‘*bibechanA*’ meaning “think” or “analyze”). In our opinion, the hedging or approximation can be equated to a subjective judgment arrived at through “thinking” or “analysis”, which, possibly is the link between the second example in (b) and the word ‘*bibechanA*’. Nevertheless, as we shall argue in Sec. 2.12, the first example of (b) refers to a different usage of ‘*kare*’ which is more similar to (d) and cannot be analyzed as ‘*bibechanA*’.

Note that in both the dictionaries ‘*kara*’ also has entries as noun with several senses (e.g., hand, ray, trunk of an elephant, tax and fee) and suffix.

2.2 ‘*kare*’ in the Corpus

Initial manual inspection of a part of a Bangla corpus revealed that ‘*kare*’ has the following three orthographic variants:

- ভালো করে (*bhAlo* [good] *kare*) – the most frequent form used.
- ভালো ক’রে (*bhAlo* [good] *ka’re*) – this form is used for the non-finite verb to distinguish it from the noun and finite verb forms (which is করে). This convention is rarely followed these days.
- ভালোকরে (*bhAlokare* [well]) – the non-finite form of ‘*kare*’ is sometimes (very rarely and in very specific cases) written together with the previous noun or adjective. As we will discuss in Sec. 2.8, use of non-finite ‘*kare*’ as a suffix is only possible when it is used as an adverbializer.

We automatically extracted sentences containing the occurrences of ‘*kare*’ and its orthographic variants, including the suffixed form, from the Anandabazar Patrika corpus and the EMILLE Bangla corpus. The extracted sentences were manually analyzed and we tried to come up with a few basic morpho-syntactic classes under which all occurrences of ‘*kare*’ can be categorized. In order to define these basic classes in an unambiguous fashion, we used several contrastive and comparative evidence from Bangla as well as other languages, as described below.

- Syntactic function as perceived by native speaker intuition as well as linguistic analysis;
- distribution based on the lexical categories of the preceding and following words within a sentence;
- list of other lexical items in Bangla that can replace ‘*kare*’ in that particular usage;
- translation equivalents of ‘*kare*’ in Hindi and English.

Our analysis revealed that including the noun, finite verb form and a special category of frozen expressions, ‘*kare*’ has 11 distinct morpho-syntactic functions in Bangla. Thus, the non-finite form alone has 8 distinct functions. Table 1 summarizes these 11 functional categories, which we will refer to as **Types,** along with examples and supporting evidence. In the rest of this section, we will discuss the nature of each of these types and argue for their identity as a distinct morpho-syntactic function.

2.3 Type 0 (Noun)

As illustrated by the following example, ‘*kare*’ as noun has a distribution and function similar to other nouns with locative case-marker.

(1)আয়-করে ১০% বৃদ্ধি হয়েছে।
AYa-kare 10% bRRiddhi haYeChe
income-tax 10% increase is-3sgPr
Income tax has increased by 10%.

Type	Example ##	Function	Distribution	Replacement in Bangla	Equivalent in Hindi*	Equivalent in English*
0	(1)	Noun with locative casemarker /-e/	Similar to *Nouns*	synonyms	Equivalent nouns	Equivalent Noun
1	(2)	Finite Verb	Similar to finite verbs	-	Finite forms of *'karanA'*	Equivalent finite verb
2	(3)	Conjunctive or perfective Participle	Noun *kare* Adjective *kare*	*karaara pare/ karaara para*	*karake/-kara*	After V+ing/ Having V-ppl
3	(5)	Adverbial Participle	Noun *kare*	-	*-kara/-te hue*	while V-ing
4	(8)	Action Instrumentalized	Noun *kare*	Noun-*era dvArA*	Noun + V-*kara*	by V-ing
5	(9)	Adverbializer / Adverbial Participle	Adjective *kare* Adverb *kare*	*bhAbe*	*se*/*taraha se*	Manner adverb
6	(10), (11)	Postposition	Noun(-e) *kare*	*madhye*/ *diye*	*meM*/*se*	in/ by
7	(12), (13)	Hedged Adverb	Quantifier *kare*	-	*sA*	-
8	(16)	Hedged Adjective	Attributive Adjective *kare*	*mato*/ *matan*	*sA*	Adjective-*ish*
9	(17), (18)	Distributive or iterative event	Cardinal-*TA kare*	-	reduplication of number	each
10	(20), (21)	Frozen Expression (Manner Adverbs)	-	-	-	-

Table 1 The distinct morpho-syntactic functions and distributions of '*kare*'. *The equivalent forms in Hindi and English are based on generalizations and may not always lead to fluent translations

2.4 Type 1 (Finite Verb)

Verb root '*karA*' in 3^{rd} person present tense simple aspect inflects as '*kare*' in Bangla. This is illustrated in (2). We also note that the verb '*karA*' forms a large number of complex predicates with nouns and adjectives, such as '*paChanda karA*' (to like) and '*sojA karA*' (to straighten).

(2)রাম রোজ একই ভুল করে
rAma roja eka;i bhula kare
ram everyday same mistake do-3sgPr
Ram does the same mistake every day.

2.5 Type 2 (Perfective Participle)

The suffix '*-e*' is used to derive the perfective or conjunctive participle form of a verb in Bangla, which motivates a sequential reading of the verb groups in a sentence. The same form also denotes the completion of the event, and hence, behaves as a perfective participle. In (3), the sentence contains two events – doing homework and going to school. Completion of the first event, i.e., 'doing homework' is followed by the second event of 'going to school'.

(3) রাম হোমওয়ার্ক করে স্কুলে গেলো
rAma homaoYArka kare skule gelo
After doing the homework, Ram went to the school.

'*kare*' can be the main verb of a finite verb group, but yet in the non-finite form if it is followed by auxiliary or vector verbs. Thus, in the following example, where '*kare*' is followed by the vector '*chalA*', the verb group is interpreted as an ongoing event and not sequential as seen in the previous example.

(4) কাজ করে চলো, ফলের আশা করো না
kAja kare chalo, phalera AshA karo na
work do-nf go, result expectation do not
Keep on working, don't expect the result.

2.6 Type 3 (Adverbial Participle)

The non-finite form of the verb can also denote simultaneity of the events and thus, behave as an adverbial participle. For instance, '*chupa karA*' is a complex predicate that means "to keep quiet". In (5), the non-finite form of this complex predicate act as an adverbial participle, meaning "quietly", rather than a perfective participle.

(5) চুপ করে শোনো
chupa kare shono
quiet do-nf listen
Listen quietly.

The distinction between the perfective and adverbial participle is clearer when we look at the equivalent translations of (3) and (5) in Hindi presented in (6) and (7) respectively. While the perfective participle gets translated as '*karake*', the adverbial participle is translated as an adverb.

(6) *rAma homawarka karake skula gayA*
Ram homework do (perf part) school went
(7) *chupachApa suno*
quietly listen

2.7 Type 4 (Action Instrumentalized)

As pointed out in (Kawtrakul et al., 2006; Choudhury et al., 2007), the non-finite form of a verb can denote the fact that the action is used as an instrument for execution of the main event of the sentence. The authors refer to this as *action instrumentalization*. For example, in the following sentence '*bala praYoga kare*' is used as an instrument of the event of "opening the door". This becomes apparent from the English translation of this phrase – "by applying force".

(8)সে বল প্রয়োগ করে দরজা খুলল
se bala praYoga kare darajA khulalo
he force application do-nf door opened
He opened the door by applying force.

2.8 Type 5 (Adverbializer)

Till now we have seen that in its various functions as a non-finite verb, '*kare*' retains its sense of "to do" or the corresponding complex predicate of which it is a part. Thus, in (8) '*praYoga kare*' means "by applying", which indeed is the sense of the underlying complex predicate '*praYoga karA*' meaning "to apply". The remaining functions of the non-finite '*kare*' (types 5 to 10 in Table 1) differ significantly from the earlier cases (types 1 to 4) in the sense that '*kare*' does not retain the meaning of "doing" or the underlying complex predicate. In fact, in these cases, if '*kare*' is assumed to be a verb, then it creates difficulties in parsing the sentence. Consider the following example.

(9)মাটির সঙ্গে ভালো করে মিশিয়ে দিন
mATira sa~Nge bhAlo kare mishiYe dina
soil-gen with good do-nf mix-nf give
Mix well with soil.

The adverbial participle phrase in the above sentence is a combination of adjective ('*bhAlo*') and '*kare'*. '*bhAlo kare*' means "well", which is not the same as the meaning of the complex predicate '*bhAlo karA*' which means "to do good". Rather, in this context, '*kare*' can be replaced by '*bhAbe*' as in '*bhAlo bhAbe*' (meaning: in a good manner). Equivalent Hindi example would be '*achChi* [good] *taraha* [manner] *se* [by] *milAo* [mix]' ("mix well"). Hence, in these cases '*kare*' behaves as a particle with a function similar to that of an adverbializer because it marks adverbial phrases.

It is worth noting that '*kare*' in the adverbializer forms are sometimes written together with the adjective (e.g,. '*bhAlokare*') as if it was a suffix. This phenomenon is never observed for types 1 to 4.

2.9 Type 6 (Postposition)

'*kare*' can act as a post position in a few specific cases as illustrated by the following sentences.

(10) থালাতে করে খাবারটা নিয়ে এসো
thAlAte kare khAbAraTA niYe eso
Plate-loc do-nf the-food bring-nf come
Bring the food in a plate.

(11) গাড়ি-(তে) করে বাড়ি গেলাম
gA.Di-(*te*) *kare bA.DI gelama*
car-(loc) do-nf home went
I went home in/by a car

Kawtrakul et al., (2006) and Choudhury et al. (2007) argue that '*kare*' as a postposition marks the instrumental case where the instrument is either

(a) means of transport as in (11), or
(b) a body part, such as '*hAte kare*' meaning "by hand", or
(c) a container as in (10).

Although it is difficult to justify that '*kare*' is always used as an *instrumental* postposition, the fact that '*kare*' does behave as a postposition in certain contexts (in 10, 11) can hardly be contested.

Further evidence comes from the facts that equivalent Hindi/English examples would always use a locative/instrumental case marker with the noun ('*meM*' ["in"] and '*se*' ["from/by"] in Hindi, 'in' and 'by' in English). Similarly, equivalent replaceable instances in Bangla would contain an appropriate case-marker ('*-e*') or suffix ('*diYe*' or '*madhye*').

2.10 Type 7 (Hedged Adverbs)

Phrases formed by non-numeral quantifiers (e.g., '*alpa*' "little"; '*kama*' "less"; '*beshi*' "lot") or intensifiers (e.g., '*khuba*' "very"; '*AchChA*' "quite") followed by '*kare*' act as *hedged adverbs* in Bangla. For example, in

(12) অল্প করে ভাত দিন
alpa kare bhAta dina
little do-nf rice give
Give rice a little.

even though '*alpa kare*' seems to modify the noun "rice", we claim that actually it modifies the event of "giving" and hence, is an adverbial phrase. The justification is as follows. Since Bangla is a relatively free word-order language, scrambling of syntactic constituents is frequent. However, while phrases like '*alpa kare*' can be placed anywhere in the sentence far from the noun '*bhAta*' (which is also a characteristic of adverbial phrases), the same is not possible for '*alpa'* (an adjective) in the phrase '*alpa bhAta*'. Consider the following sets of sentences.

(13) Drink a lot of water.
(a) বেশী করে জল খাও {*beshI kare jala khAo*}
(b) জল বেশী করে খাও {*jala beshI kare khAo*}
(c) জল খাও বেশী করে {*jala khAo beshI kare*}
(14) Drink a little amount of tea.
(a) কম চা খাও {*kama chA khAo*}
(b) *চা কম খাও {*chA kama khAo*}
(c) *চা খাও কম {*chA khAo kama*}

The unacceptable sentences (14b) and (14c) and their acceptable counterparts (13b) and (13c) show that the adverbial nature of '*beshI kare*' allows it to be adjoined to the VP but the NP '*kama chA*' cannot be adjoined to the VP in (14b) and (14c).[4] The fact that these adverbial phrases carry a sense of *hedging* can be understood by comparing (12) with

(15) অল্প ভাত দিন
alpa bhAta dina
Give a little rice.

While the former means "somewhat little rice", the latter specifically says "little rice". In fact, this distinction is clearly marked in the equivalent Hindi constructs '*tho.DA* [little] *sA*[like] *chAvala* [rice]' ("a little bit of rice") for (12) vs. '*tho.DA* [little] *chAvala* [rice]' ("little rice") for (15). Thus, '*kare*' in this context is replaceable by the particle '*sA*' in Hindi.

2.11 Type 8 (Hedged Adjectives)

While with non-numeral quantifiers and intensifiers '*kare*' forms hedged adverbs, with attributive adjectives it can form *hedged adjectives*. This is illustrated in the following example.

(16) লম্বা করে ছেলেটা
lambA kare CheleTA
tall do-nf (the) boy
The tallish boy

In this context, '*kare*' can be replaced by the Bangla particles '*mato*' or '*matana*' and the equivalent Hindi translation would contain the particle '*sA*' as in '*lambA* [tall] *sA* [like] *la.DakA* [boy]' ("the tallish boy"). Thus, these constructs are very similar to that of Type 7 except for the fact that unlike Type 7, here the resultant phrase functions like an adjective. Deeper analysis reveals that '*kare*' can yield hedged adjectives only with a very specific category of attributive adjectives that have other interesting morpho-syntactic features. However, due to paucity of space we are unable to include a detailed discussion on this issue.

2.12 Type 9 (Distributive or Iterative Event)

'*kare*' bears the sense of a distributive or iterative event when it follows a numeral quantifier (i.e., a number). We illustrate this fact with the following examples.

(17) প্রত্যেক ক্ষেত্রে দুটি করে বিকল্প তৈরি হয়েছে
pratyeka kshetre duTi kare bikalpa tairi haYeche
every case-loc two do-nf option create is
There are two options in each of the cases.

(18) অনেকগুলি ক্ষেত্রে দুটি করে বিকল্প তৈরি হয়েছে
anekaguli kShetre duTi kare bikalpa tairi haYechhe
many case-loc two do-nf option create is
There are two options in many of the cases.

(19) একটি ক্ষেত্রে দুটি (*করে) বিকল্প তৈরি হয়েছে
*ekaTi kShetre duTi (*kare) bikalpa tairi haYeChe*
one case-loc two (*do-nf) option create is
There are two options in one of the cases.

[4] (14b) and (14c) are not ungrammatical. However, it does not have the same meaning as (14a).

‘*duTi kare*’ is unnatural in (19) because it is not possible to distribute the fact of two options being created over the elements of a singleton set (‘*ekaTi kShetre*’ - single case). Thus, in this context ‘*kare*’ acts as a particle carrying the sense of a distributive or iterative event. Note that through arguments based on scrambling one can show that like Type 7, phrases such as ‘*duTi kare*’ are also adverbs, whereas ‘*duTi*’ is an adjective.

2.13 Type 10 (Frozen Expressions)

There are a few frozen expressions such as ‘*aneka kare*’ and ‘*kata kare*’ that cannot be classified under any of the previous 10 categories. We illustrate two such cases, though possibly there are many more frozen expressions with ‘*kare*’.

(20) অনেক করে থাকতে বলেছেন
aneka kare thAkate balechena
many do-nf stay-inf said
(He) requested several times to stay back.

(21) কত করে থাকতে বললেন
kata kare thAkate balalena
how-much do-nf stay-inf said
Requested so many times to stay back

2.14 Reduplication: ‘*kare kare*’

Reduplication of ‘*kare*’ refers to iteration of the same event. Although this has not been listed as a morpho-syntactic type, we mention this usage for the sake of completeness.

(22) একই কাজ করে করে ক্লান্ত হয়ে গেছি
Ekai kAja kare kare klAnta haYe gechhi
One work do-nf do-nf tired is went
(I am) tired of doing the same thing again and again.

3 Theoretical analysis

The extremely divergent function of ‘*kare*’ poses a very interesting question: how can one explain the various morpho-syntactic roles of ‘*kare*’, when morphologically it can have only three possible analyses? This question is of significant interest to not only theoretical and historical linguists, but also to NLP researchers because a deeper understanding of the evolution of these divergent functions and their inter-connections, if any, could help develop a theory for a comprehensive treatment of ‘*kare*’ in NLP.

In this section, we show how these apparently divergent functions of ‘*kare*’ can be attributed to two very common linguistic phenomena - morpho-phonemic change and functional extensions.

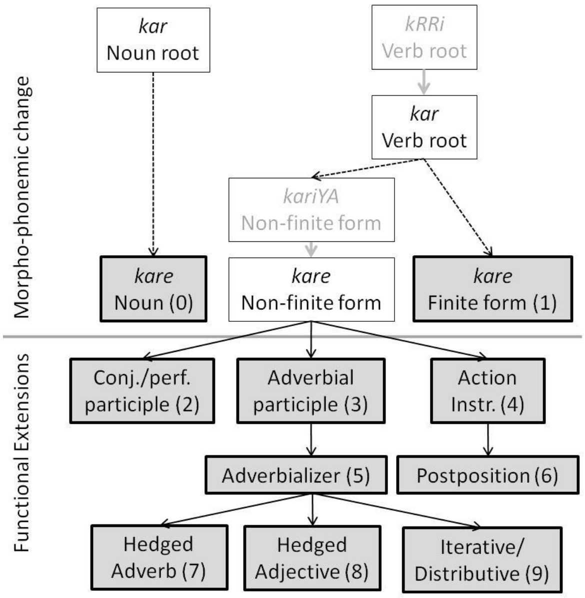

Figure 1 Schematic for evolution and interrelations of the various morpho-syntactic functions of ‘*kare*’. Type 10 (frozen expressions) is not included. Legend: Black font – current forms, gray font – historical forms, broken arrow – inflection, gray arrow – morpho-phonemic change, black arrow – functional extensions, number in parenthesis – the type (refer to Table 1).

We use both synchronic and diachronic arguments and come up with a unified framework that captures the evolution as well as the interrelations of the 10 types (excluding type 10, i.e., frozen expressions) described in Table 1. A schematic of this framework is shown in Fig. 1.

3.1 Morpho-phonemic Change

The three underlying morphological structures of ‘*kare*’ are derived from three different historical forms. As described in (Chatterji, 1926), the nominal root ‘*kar*’ and its inflected form ‘*kare*’ had the same forms in Old Indo Aryan (i.e., Vedic Sanskrit). Thus, the forms remained unchanged over a long period. However, the verb root ‘*kar*’ and its inflected forms have been derived from the Sanskrit ‘*kRRi*’ through a series of morpho-phonemic changes. The finite verb form ‘*kare*’ is also found in Classical Bangla, which was spoken during 12-15th century (Chatterji, 1926). Nevertheless, the non-finite form ‘*kare*’ is a rather recent entry that has been derived from the corresponding Classical Bangla form ‘*kariYA*’, again through a series of phonological changes.

The non-finite ‘kare’ is pronounced as /kore/, whereas the other two are pronounced as /kɔre/. Based on this very simple clue, one can claim that the types 2 to 10 have the non-finite ‘*kare*’ as their underlying form. In the next subsection, we try to argue that these various morpho-syntactic functions have been derived from the non-finite form as a consequence of functional extensions.

3.2 Functional Extensions

In Bangla, the non-finite form of a verb obtained through suffixation of ‘*-e*’ serves three basic functionalities: (a) conjunctive or perfective participle, (b) adverbial participle, and (c) action instrumentalization. While the conjunctive participle usage of the non-finite form is well accepted and needs no further justification, we illustrate the other two usages through examples (23) and (24).

(23) সে হেসে দরজা খুললো
se hese darajA khulalo
‘He/she opened the door with a smile’

(24) সে লাফিয়ে ঘরের মধ্যে ঢুকলো
se lAphiYe gharera madhye Dhukalo
‘He/she entered the room by jumping into it’

In (23), ‘*hese*’, which is the non-finite form of the verb ‘*hAsA*’ (to smile/laugh), signifies the simultaneity of the event of “smiling” while “opening the door”. Thus, clearly it is an adverbial participle. Similarly, in (24) ‘*lAphiYe*’, which is the non-finite form of the verb ‘*lAphAno*’ (to jump), signifies the manner in which the event of “entering into the room” took place. Hence, it is also an adverbial participle. However, since the act of “jumping” facilitates the event of “entering”, ‘*lAphiYe*’ can as well be considered as the instrumentalized form of the verb ‘*lAphAno*’ (Kawtrakul, 2006; Choudhury, 2007). Note that it is possible to argue that action instrumentalization is essentially a special case of adverbial participle. However, here we choose to maintain this subtle difference because, as we shall see, this helps us in understanding the functional extensions.

Thus, type 2, 3 and 4 are three basic morpho-syntactic functions of ‘*kare*’, which are also exhibited by the non-finite forms of the other Bangla verbs. As shown in Fig. 2, this makes the first layer of functional extensions of the non-finite ‘*kare*’. It is also worth mentioning that for these three types, ‘*kare*’ retains its verbal function, a consequence of which is the fact that the phrase formed by ‘*kare*’ and its preceding word (say W) in these contexts has the same meaning as the complex predicate ‘W *karA*’.

The remaining six functions of ‘*kare*’ are non-verbal even though the underlying form is the same non-finite verb. We argue that type 5 (adverbializer) is an extension of the adverbial participle function of ‘*kare*’ (i.e., type 3). These two functions are similar in the sense that the preceding word(s) and ‘*kare*’ together form an adverbial phrase, or more precisely, manner adverbs. However, the difference, as mentioned in Sec. 2.6, lies in the fact that the complex predicate semantics is valid for type 3, but not for type 5. Moreover, in type 3 nouns precede ‘*kare*’, whereas in type 5 adjectives or adverbs precede ‘*kare*’. We hypothesize that this is a consequence of a generalization of the semantics of ‘W *kare*’ from the complex predicate ‘W *karA*’ for Ws that are frequently occurring adjectives and adverbs.

Noting that ‘*kare*’ is primarily used as an instrumental postposition, we claim that type 6 is an extension of the action instrumentalization usage (type 4). Thus, while in (8), ‘*praYoga kare’* can be analyzed as a complex predicate, an alternative analysis where ‘*praYoga*’ (application) is a noun and ‘*kare*’ is an instrumental postposition similar to “by” would lead to very similar semantics. We claim that this alternative analysis has been extended to usages such as ‘*jAhAje* [ship-loc] *kare*’ (by/in a ship), where the complex predicate analysis is no longer possible.

The link of the non-finite ‘*kare*’ to the types 7, 8 and 9 is least clear. However, as discussed in Sections 2.10 and 2.12, types 7 and 9 should be analyzed as adverbial phrases, which leads us to believe that these usages are functional extensions of type 5 (adverbializer). One could also argue that type 7 and 9 are direct extensions of type 3 and not type 5. However, we feel that it is difficult to establish a direct connection between the adverbial participle and these functions because the complex predicate analysis, which is valid for the adverbial participles, is absolutely incorrect for types 7 and 9. However, it seems plausible to semantically stretch the meaning of the complex predicates a little to fit similar analysis for type 5. Thus, it can be argued that type 5 is a functional link between adverbial participles and its extensions as hedged and distributive adverbs.

We also observe that types 7 and 8 are linked by the fact that in both the cases, ‘*kare*’ bears the sense of hedging. Presently, we are unsure whether type 8 is a direct extension of type 5 and, thus, a sibling of type 7 in the functional extension tree (Fig. 1), or a further extension of type 7. In the absence of any other clue, we resort to the principle of Occam's razor and choose the former alternative because it limits the depth of the functional extension tree (Fig. 1) to three. However, we are aware that these choices are nothing more than educated guesses, and deeper linguistic analysis is required to resolve these issues in a systematic manner.

4 Treatment of '*kare*' in NLP

We have seen that the lexical item '*kare*' is ambiguous at every level of linguistic analyses. It has two possible pronunciations that must be resolved for speech applications, three possible morphological analyses, 11 distinct morpho-syntactic functions that need to be resolved during POS-tagging, chunking and/or parsing, and 53 different senses that should be resolved during semantic analysis. While sense resolution is beyond the scope of the current work, pronunciation resolution is trivial if the morpho-syntactic function of '*kare*' is resolved. Therefore, here we make some recommendations on how '*kare*' should be handled at various levels of morpho-syntactic analysis.

4.1 Morphological Analysis

As discussed earlier, '*kare*' has three distinct morphological analyses. Therefore, it is recommended that a morphological analyzer should analyze the form as (a) *kara* (Noun) + *e* (locative case-marker), (b) *karA* in finite verb form (3rd person Present tense simple/ habitual aspect), (c) *karA* in non-finite form, and (d) particle (which includes postposition). Further disambiguation and differentiation between these types should be handled by the subsequent phases of analysis.

When '*kare*' is written as a suffix (as in *'bhAlokare*') the morphological analyzer should be able to recognize it and parse the whole word as an adverb. However, in compound words such as '*Ayakare*' (income tax-loc) the analysis should be *Aya+kara+e* or *Ayakara+e*, but never *Aya+kare*.

4.2 Parts-of-Speech Tagging

We came across three different POS annotation schemes for Bangla – Anncorra (Bharati et al., 2006), IL-POSTS (Baskaran et al., 2008) and Sankalan (Choudhury et al., 2004), which recommend slightly different ways of annotating '*kare*'. Table 2 compares these annotation schemes against the classification proposed here.

Type	Anncorra	IL-Posts	Sankalan
0	Common Noun (NN)	Noun common (NC)	Noun common (NN)
1	Verb Main (VM)	Verb main fn- marked (VM.fn)	Finite verb (VF)
2-4		Verb main nfn-marked (VM.nfn)	Nonfinite verb (VN)
5-10	Post-position (PP)	Post-position (PP)	Post-position (PPI)

Table 2 Annotation conventions for '*kare*' in 3 POS tagsets. Legend: (n)fn – (non)finite

We see that all the tagsets agree on annotation of type 0 and type 5 to 10. However, due to basic differences in the treatment of verbs in the annotation schemes, types 1 to 4 are handled differently in the tagsets. However, based on our analysis we propose the following annotation scheme:

- Type 0: *Noun* with locative case-marker
- Type 1: *Finite verb*
- Type 2 and 3: *Non-finite verb*, because it is not possible to disambiguate between these types from syntactic context alone. One requires deep semantic information.
- Type 4: *Non-finite verb* or *Postposition*: Since this type will always surface as prepositional phrases in many languages, it would help during machine translation if Type 4 is annotated as a postposition. However, as discussed earlier, it is also possible to treat this type as an adverbial participle, in which case it should be annotated as a non-finite verb. We believe that the latter scheme will lead to more accurate POS taggers.
- Type 5, 7-10: *Particle*, because it is easy to identify these types due to the preceding word always being an adjective, adverb or quantifier. We think that further disambiguation between these types is not important at the level of POS-tagging.
- Type 6: *Postposition*, because identification of this type is not very difficult (the preceding word comes from a restricted class, such as means of transportation, body part, container etc., and sometimes the presence of the suffix '*-e*' and/or postposition '*madhye*'), and disambiguation helps in further analysis.

4.3 Chunking and Parsing

We recommend that ‘*kare*’ as a postposition or particle (types 5 to 10) should be chunked with the previous word, whereas ‘*kare*’ as a noun (type 0) is to be treated similar to other nouns. Types 1 to 4 bring in some confusion because ideally we would like to identify the complex predicate as a single chunk. Nevertheless, this might turn out to be a formidable task unless one has a list of all the complex predicates with ‘*karA*’. We also recommend that type 3 should be parsed as an adverbial phrase, while type 4 can be parsed either as an adverbial phrase or a noun followed by postposition ‘*kare*’ depending on the POS annotation scheme.

5 Conclusion

In this paper, we have analyzed the various morpho-syntactic functions of the lexical item ‘*kare*’ in Bangla, established the interconnections and evolution of these apparently divergent functionalities, and based on this analysis made some recommendations for treatment of ‘*kare*’ in NLP at different levels of morpho-syntactic analysis. Due to the paucity of space, it has not been possible to include an in-depth discussion on the difficulties that one might face during POS annotation or parsing of ‘*kare*’. As an extension of this work, we plan to strengthen the theoretical framework by deeper analysis of types 7, 8 and 9. Furthermore, we plan to build a ‘*kare*’ bank consisting of annotated examples of the different functions of the word and train learning algorithms for disambiguation.

Acknowledgments

We would like to thank Prof Pushpak Bhattacharyya, Prof Tanmoy Bhattacharya and Prof Jayshree Chakraborty for their valuable inputs.

References

Haricharan Bandyopadhyay. 1966. বঙ্গীয় শব্দকোষ, প্রথম খন্ড । New Delhi: Sahitya Akademi

S. Baskaran et al. 2008. A Common Parts-of-Speech Tagset Framework for Indian Languages. *LREC, 2008*. Morocco.

S. Bergsma, D. Lin, and R. Goebel. 2008. Distributional Identification of Non-Referential Pronouns. *ACL-08: HLT*, Columbus, Ohio, USA. pp.10-18.

A. Bharati, D. M. Sharma, L. Bai and R. Sangal. 2006. AnnCorra: Annotating Corpora Guidelines for POS and Chunk Annotation for Indian Languages. http://ltrc.iiit.ac.in/winterschool08/

Shailendra Biswas. 1957 (2006). *Samsad Bangla AbhidhAna*. SAhitya samsad, Kolkata.

A. Boyd, W. Gegg-Harrison, and D. Byron. 2005. Identifying non-referential it: a machine learning approach incorporating linguistically motivated patterns. In *ACL Workshop on Feature Engineering for Machine Learning in NLP*. pp. 40-47.

S. K. Chatterji. 1926. *The Origin and Development of the Bengali Language*. Rupa & Co., New Delhi.

Avinash Chopde. 2001. ITRANS Version 5.30. http://www.aczone.com/itrans/

M. Choudhury, P. R. Ray, S. Sarkar and A. Basu. 2004. Hindi and Bengali Tagset and Guidelines for Manual Tagging of the Multilingual Parallel Corpus. http://www.mla.iitkgp.ernet.in/Tag.html

M. Choudhury, E. Murguia, S. Sarkar, V. Moriceau, A. Kawtrakul and P. Saint-Dizier. 2007. Generating Instrumental Expressions in a Multilingual Question-Answering System. *IJCAI-07 Workshop on Cross Lingual Information* Access, India.

Richard Evans. 2001. Applying machine learning toward an automatic classification of it. *Literary and Linguistic Computing, **16**(1):* 45-57.

S. Gupta, M. Purver, and D. Jurafsky. 2007. Disambiguating between generic and referential you in dialog. In *ACL Demo-Poster Sessions*, pp. 105-108

Yoko Iyeiri. 2004. The Use of the Auxiliary *Do* in Negation in *Tom Jones* and some other literary works of the contemporary Period. In *New Trends in Historical Linguistics: An Atlantic View*, ed. Moskowich-Spiegel F., Isabel & Crespo García, B.

A. Kawtrakul et al. 2006. A conceptual analysis of the Notion of Instrumentality via a Multilingual Analysis. *3rd ACL-SIGSEM Workshop on Prepositions, EACL 2006 workshop*, Italy.

B. F. Kelly. 2002. “Well you can’t put your swimsuit on top of your pants!”: Child-mother uses of in and on in spontaneous conversation. In *31st Stanford Child Language Research Forum* pp 69-78.

C. Müller. 2007. Resolving *It*, *This*, and *That* in unrestricted multi-party dialog. *ACL-07*, pp 816-823

Chris D. Paice and Gareth D. Husk. 1987. Towards the automatic recognition of anaphoric features in English text: the impersonal pronoun .it. In *Computer Speech and Language*, **2**:109-132.

B. Santorini. 1990. Part-of-Speech tagging guidelines for the Penn Treebank project. *Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania*. http://www.cis.upenn.edu/~treebank/home.html