International Journal of Innovative Studies in Sociology and Humanities (IJISSH)
Volume: 1 Issue: 1 | September 2016 www.ijissh.org

Multifunctionality of Hyphen in Bangla Text Corpus: Problems and Challenges in Text Normalization and POS Tagging

Niladri Sekhar Dash

Linguistic Research Unit, Indian Statistical Institute, Kolkata

Abstract: *Complete understanding of a text not only depends on the words and sentences used in the composition, but also on the proper decipherment of the role of punctuation marks used along with words in sentences. That means, there are some unique textual elements which require closer investigation to know how they exert their functions in the texts. A hyphen, as a member of this clan, requires extra attention as it reveals multifunctionality in usage in texts, the proper understanding of which may lead us to capture the actual sense encoded with the words used by the text composers. Beyond this crucial cognitive issue, the study of the role of the hyphen is also relevant to achieve control over the formal structure of texts, exploring the intricacies layered in textual architecture, analysing sentences, identifying proper names, capturing sense variation of words, and many other linguistic functions in texts. What we understand is that a hyphen can offer enormous potentiality in text understanding, text processing, and sense disambiguation of linguistic expressions. Therefore, an elaborate discussion on the usage patterns of hyphen becomes necessary in the context when we want texts to be rightly interpreted and properly processed for both man and machine learning. Keeping this observation in mind, in this paper, I have made an attempt to discuss the multifaceted usage of the hyphen as noted in the Bangla text corpus containing samples of written texts from more than eighty subject areas. The observations that I furnish in this paper are shaped up with examples and instances obtained from the corpus to address several linguistic and computational issues of the language. These are linked not only to Bangla language text but are also relevant to the texts of other Indian languages.*

Keywords: *corpus, character, frequency, function, occurrence, sense, words*

1. INTRODUCTION: DEFINING HYPHEN

Hyphen is an orthographic symbol that discharges various linguistic functions within a piece of text. It helps to identify syllabic pauses in utterance as well as to dissolve ambiguities in forms and meanings of words and sentences. It also prevents us from deducing wrong meanings from a word or a construction used in a piece of a text. Therefore, the functional importance of a hyphen is almost similar to that of other characters that are used in writing a text in a language.

Etymologically, the term *hyphen* is derived from the Latin word *huphen* which means "together". It is again derived from Greek by joining two Greek words, namely, *hupo* "under" + *hen* "one" (Crystal 1995). If we look at the usage varieties of punctuation marks used in a piece of a text, we can argue that perhaps the *hyphen* is one of the most troublesome punctuation marks, which display a wider range of variations in usage in a natural language text. Since no authority of any sort has ever tried to regulate its use in texts, there is no regularity in its use in many languages, and the Bangla language, like many Indian languages, is one of the members of this category.

The *Collins COBUILD English Dictionary* (1996) has identified hyphen as a noun and observed that it is a sign, which is used to join words together to make a compound word or to indicate that the first part of a word has been written at the end of one line and the remaining part of it is written at the beginning of the next line. The *Illustrated Oxford Dictionary* (1998) also observed that hyphen is a "sign used to join words semantically or syntactically, to indicate the division of a word at the end of a line, or to indicate a missing or implied element". The *Concise Oxford Dictionary of Current English* (1998) has described that a *hyphen* can be a considered as a noun or as a verb. It is a sign, which has been used "...either to join the words semantically and syntactically, or to indicate the division of a word at the end of a line, or to indicate a missing element in a sentence".

The definitions furnished above from the dictionaries are partly true since these dictionaries failed to encompass the wide varieties of use of the hyphen revealed in a natural language text. It is found that the use of the hyphen, in

the case of compound word formation, is mostly arbitrary, particularly when the formative members of the compound words consist of single characters or syllables. In such cases, the language users usually enjoy full liberty at the time of using this particular punctuation mark. As a result, except in some unavoidable situations, the hyphen is used quite arbitrarily in texts. In some occasions, it is not used although it is required; on the contrary, it is used on those occasions where it is not required at all. The examples presented in Section 6 can substantiate this observation.

My clear understanding is that the analysis of hyphen use will largely be based on written text samples. A spoken text may contain a hyphen mark, but there is no way to understand it until and unless the speech is transcribed into written form. Moreover, there are issues in the selection of strategic means for encoding prosodic elements of speech into written form within the existing limitations of standard keyboards that may not allow one to resort to special characters available in the keyboards.

Keeping all these issues in view, in this paper, I have made an attempt to understand the identity and role of a hyphen in Bangla written text. In Section 2, I have tried to shed some light on the history of study of punctuation marks including that of hyphen in texts; in Section 3, I have referred to the debate with regard to the actual identity of hyphen in a language script; in Section 4, I have focussed on the importance and relevance of studying hyphen in language processing and language technology; in Section 5, I have presented a short discussion on the punctuation marks used in the Bangla texts; in Section 6, I have classifies the multi-functionalities of hyphens in the Bangla texts; in Section 7, I have proposed some methods about how the hyphenated words should be treated at the time of PSO tagging, chunking, and other language processing works; and in Section 8, I have tried to find out the importance of the present study in language description, application, and computation.

2. SHORT GLIMPSE INTO HISTORY

Historians have made studies to record that hyphen has served very different functions at different points in time, depending on the nature of the audience for which the text was put into writing. In a fascinating narration, Parkes (1993) has presented a nice study on the history of use of punctuation marks in texts, which has gone all the way back to ancient Greek and Latin texts. He has clearly shown that punctuation marks have served different functions in different languages and cultures in the history of writing. In ancient Greece, one important goal of writing was to preserve spoken language in visual form and help the students to become better orators by following the usages of punctuation in texts. Since people did not read texts silently until much later, in those days, it was necessary for the learners to follow the usage of punctuations quite carefully. He has also showed that the number of punctuation marks in the Bible varied greatly from one age to the next based on target readership (Parkes 1993: 17). In the early years, when the readers were homogeneous (i.e., native speaking monks), there are fewer punctuation marks. In the later years, when the readers are heterogeneous in nature across far-flung countries, there was more use of punctuation marks per page in the Bible (Houston 2014 : 21).

The use of punctuation marks not only varied across ages, it also varied based on the scripts of languages. That means punctuation marks are not always uniform in shape and function in all languages, although their use are noted in almost all languages. For instance, in Turkish and Germanic scripts, an umlaut mark is placed over a vowel (e.g., ä); in French a stress mark is put on a vowel (e.g., é), a comma is used below a consonant grapheme (e.g., ç), an arrow mark is used over a vowel grapheme (e.g., â); in Polish script the consonant *L* is cut with a cross bar; in Czech an inverted arrow is used over the consonant *r*; in Spanish an interrogative sentence begins with an upside-down interrogative mark (¿) followed by another interrogative mark that is used at the end of the same sentence.

In case of Indian language scripts the story is somewhat different. Most of the Indian language scripts did not have the use of punctuation marks in texts till the Roman script, through English, came into India and influenced many Indian language scripts to adopt the punctuation marks in their writing systems. Before that Indian language scripts had mostly a sentence terminal marker known as *pūrṇacched* (full stop) and a one of two punctuation marks like the *dash* and the *arrow sign.*

3. IDENTITY CRISIS OF HYPHEN

The functional identity of the hyphen is halfway between spelling and punctuation. It becomes an issue of spelling when we try to choose between the three alternative forms made with/without a hyphen mark: *field-work*, *field*

work, and *fieldwork*. Here we have to formulate a policy and implement it to determine if the word should carry hyphen or simply ignore it. On the other hand, hyphenation becomes an issue of punctuation when we are placed in the middle of a long word at the end of a line where we have to decide what are the acceptable segments before and after the hyphen that is supposed to be used in the word as the entire length of the word has to broken into two parts because the word cannot be accommodated in the single line. Here we have two alternatives, either we can select a new word with similar sense and replace the long word, or we can keep the long word unchanged with a conviction that replacement of it with a new word will not serve our purpose. In that case, we can carry out the hyphenation process by leaving a single letter stranded in the second line (e.g., *scenari-o*). This is unacceptable because such a letter is known as 'widow' and it does not go with the normal practice of hyphenation in texts. It may be acceptatble if we leave a single vowel letter stranded in the first line (e.g., *a-synchronic*), where the isolated vowel grapheme may stand as a morph with an implicit/explicit meaning. Here also, we have to select between two options: either divide the word at the syllable boundaries (e.g., *punc-tu-al*), or break the work at the morpheme boundaries (*be-wild-er-ment*). We normally prefer morpheme boundaries as the place of hyphenation since it provides at least one pronounceable syllable on each side of the hyphen. Thus, we do not hyphenate a word as *norma-lly*, but do as *normal-ly*. The basic idea is to make a segment, as far as possible, pronounceable with (often implicit) sense.

There are also cases where we find that words whose spelling may or may not ask for for a hyphen mark (e.g., *sociocultural* or *socio-cultural*). It may happen that the word happens to fall at such a point that it is possibe to find the line ending at the string '*socio-'* and we are then forced to use a hyphen to deal with the orthographic requirements of the string. In such a case, the use of the hyphen may be ambiguous because the target readers can have two different readings: *sociocultural* or *socio-cultural*. Here it is difficult to inform the readers whether we want the spelling of the word to be counted as a hyphenated or as a non-hyphenated word string. In such a case it is possible to formulate a solution to make it a principle that, if the spelling of a word is intrinsically hyphenated, then we shall show two hyphens (e.g., *socio--cultural*): the hyphen at the end of the first line belongs to the punctuation system; while the one at the start of the second line is the part of the spelling of the word. Such theoretical postulation, however, does not have any realistic representation in the real life situation, as far as the texts of natural languages concerned and as reflected in the text samples captured in corpora.

4. HYPHEN IN LANGUAGE TECHNOLOGY

The role of a hyphen in natural language processing activities (such as, *texts segmentation, prosodic annotation, syntactical analysis, information retrieval, POs tagging, chunking, morphological analysis, compound decomposition, local word grouping, lexical collocation, coding in multilingual systems,* etc.) has become an issue of serious concern due to its multifaceted role in texts. Even though hyphen is not seen as an area of teaching knowledge, we can hardly deny its role in reading and writing. This is also true for natural language processing, where hyphen plays an important role in text processing tasks. A hyphen, like other punctuation marks, is a 'natural tag' of information and indicator on which most of the text processing techniques rely. It is, therefore, necessary to explore and study all the issues of hyphen use in the context of processing words, compounds, phrases, and sentences within multilingual, multi-writing, and multi-coding activities of natural language texts.

In recent years language processing activities are confronted with new issues found in texts. Indeed, it is now understood that we have to work not only on isolated sentences but on the range of structured and unstructured texts available from various sources. For instance, the texts that are being collected from the internet, homepages, websites, blogs, and twitters are made of different forms and formats with many non-textual elements embedded into these. Since such texts are hardly normalized, all the text processing techniques require pre-processed texts in order to conduct syntactical, semantic, and pragmatic analysis on these. Each text has two structures: formal and discursive. The formal structure refers to the form and composition of a text resulted from the coding in a typographical system, from and layout. The discursive structure, on the other hand, is shaped up with the content and information encapsulated within the text. The latter one depends on the earlier.

At the time of pre-processing we have to understand the formal structure of a text (e.g., *title, subtitle, text fragmentation, sentences, paragraphs, propositions, words, quotations, item list, spatial gap, images, diagrams, captions, boxes, tables, punctuations, formulae,* etc.), before we exploit its discursive structure (e.g., *temporal, spatial,*

topic, event, concepts, relations between concepts, terms, anaphoric links, discourse, etc.). That means the success in the analysis of the discursive structure of a text largely depends on how the formal structure of the text is processed, accounted for, and normalized.

Without complete control of formal structure, text processing tasks will not be operational in the true sense of the term. This issue will hardly appear when we try to work with isolated sentences. However, for semantic analysis, the text used in a sentence must be segmented into manageable linguistic units that are superior or inferior to the normative sentences, by taking into account the semiotic marks clearly and formally known by a computer. Hyphens (as well as other punctuation marks and typographic signs) are still the most relevant semiotic marks as they can provide sharp indications for formal text segmentation and structuring. In fact, these should be treated as indicators — the rudiments of textual linguistics.

The traditional description of the use of punctuation marks in texts generally is normative in nature as they do not allow the formation of rules that could lead to automatic segmentation of forms. Furthermore, these descriptions do not consider the semantic analysis of polysemous punctuation marks like a hyphen, comma, semicolon, colon, dash, parenthesises, etc. in texts. In reality, however, these punctuation marks play a crucial role in the semantic structuring of texts, as their analysis improves the level of accuracy in text segmentation and discursive structuring. Based on this argument, we should try to understand the role of hyphen in the following issues:

(a) Formal segmentation of words,

(b) Defining textual architecture,

(c) Syntactic analysis of phrases and sentences,

(d) Identification of proper names, compounds, reduplicated words, affixed words, multi-word strings, cardinal-strings, etc.

(e) Understanding meaning variation of words,

(f) Understanding grammatical texture of words and larger units, and

(g) Word sense disambiguation,

When we keep these functions in view, we can clearly understand that hyphenation offers enormous potentialities to text processing tools, by disambiguating an expression. Although this may seek help from other contextual elements, in general, a hyphen can help in capturing the sense evoked in terms in usage (Blake and Bly 1993: 131). What is important here is that text processing tools must dissolve the issues involved in the use of the hyphen in words before the text is put to any work of language processing. Keeping this argument in mind, in the following sections, I would like to discuss the use of the hyphen in Bangla text corpus with examples and instances obtained from the corpus itself to resolve some linguistic and computational issues of text normalization, part-of-speech tagging, and words sense disambiguation (Section 6 and Section 7). Before that, I like to present a short review on the state and status of the study of punctuation marks (including hyphen) in Bangla as it will provide readers a larger canvas to pinpoint the position of a hyphen in the panorama of punctuation use in the written Bangla text corpus.

5. PUNCTUATIONS IN BANGLA TEXT CORPUS

The Bangla written text corpus under investigation contains a whole range of punctuation marks derived from the *Devnāgarī* (Sanskrit) and the *Roman* (English) script. The list of punctuation marks used in the Bangla text composition includes the followings:

(1) pūrṇacched “full stop”, (2) kamā “comma”,

(3) kolon “colon”, (4) semikolon “semicolon”,

(5) praśnabodhak cihna “question mark”, (6) bismaysūcak cihna “exclamatory mark”,

(7) ekokti cihna “single quotation mark”, (8) dvirukti cihna “double quotation mark”,

(9) pratham bandhanī "parentheses", (10) dwitīya bandhanī "brace", `

(11) tṛtīya bandhanī "square bracket", (12) ḍyās "dash",

(13) bindu "dot", (14) anukti cihna "ellipsis",

(15) tārakā cihna "asterisk", (16) bikalpa cihna "stroke",

(17) ūrdhakamā "apostrophe", and (18) hāiphen "hyphen".

It is noted by many scholars that similar to some other Aryan languages, Bangla language normally uses the *pūrṇacched* "full stop" mark at the sentence terminal position to indicate the end of a declarative construction. On the other hand, all the Dravidian language scripts, like that of Roman, use the *full stop* at sentence terminal position. Other punctuation marks used in the Bangla texts are almost similar to those used in the Roman (English) script (Roy 1989). Besides these primary punctuation marks, some other orthographic symbols and signs are also used in the Bangla prose text composition. These include *bakrākṣar* "italics", *tir cihna* "arrow mark", *śatkarā cihna* "percentage mark", *samān cihna* "equal sign", *ataeb china* "therefore sign", *candrabindu* "moon dot", etc. (In Bangla writing convention, *candrabindu* "moon dot" is a unique symbol, which is used immediately before the name of a person to imply that the person is no longer alive). Furthermore, various mathematical and geometric symbols are also used in the Bangla corpus text to execute specific linguistic-informative functions.

There have been many studies with regard to the use of characters in the text, particularly vowel and consonant characters as well as clusters in a language (Miller, Newman and Friedman 1958, Miller 1951). In case of Bangla it is also noted that the information with regard to frequency of occurrence of characters in the text provides many important insights into understanding the language and its properties (Das, Bhattacharya and Mitra 1984, Mallik and Nara 1994, Mallik and Nara 1996, Mallick 2000). In support of this observation, it becomes important to know the frequency of occurrence of punctuation marks in a written text, which can help us understand how text composers deploy punctuation symbols as useful devices for information embedding in text composition (Dash and Chaudhuri 1998). Keeping this idea at background, a simple frequency study is carried out on the Bangla text corpus of nearly five million words to count the frequency of occurrence of the punctuation marks in the language, It is found that (Table 1), with respect to other punctuation marks used in the corpus, hyphen is used quite moderately, whereas comma, similar to English texts (Bayraktar *et al.* 1998), is the highest in usage followed by *pūrṇacched, semicolon, question mark,* and *colon* (Dash and Chaudhuri 2000).

Table 1: *Frequency of use of punctuation marks in the Bangla text corpus*

Punctuation Marks in Bangla	%-age	Punctuation Marks in Bangla	%-age
Pūrṇacched (full stop)	16.26	hāiphen (hyphen)	8.89
kamā (comma)	21.32	ḍyās (dash)	2.95
kolon (colon)	6.16	ekokti cihna (single quotation mark)	2.32
semikolon (semicolon)	15.27	dvirukti cihna (double quotation mark)	3.75
praśnasūcak cihna (question mark)	7.38	bindu (dot)	1.65
bismaysūcak cihna (exclamatory mark)	4.36	anukti cihna (ellipsis)	1.13
pratham bandhanī (parentheses)	2.13	tārakā cihna (asterisk)	0.50
dwitīya bandhanī (brace)	1.74	bikalpa cihna (stroke)	0.50
tṛtīya bandhanī (square bracket)	1.34	Others	1.00
ūrdhakamā (apostrophe)	1.35		

In the list above (Table 1), the hyphen is in the fourth position (8.89%), after the *comma* (21.32%), *pūrṇacched* (16.26%), and *semicolon* (15.27%). All these four punctuation marks constitute more than 60% of the total occurrence of punctuation marks in the corpus (vis-a-vis, Bangla language). However, I assume that the percentage of hyphen may slightly change towards lower or greater percentage once the confusion of use of hyphen with that of the *dash* is solved. It is often noted that language users are not much careful in the use of hyphen and *dash*, as they are not clearly informed about the distinct nature and function of these two punctuation marks. As a result, both the symbols (due to their close structural affinity) are often interchanged in use in texts. Quite often it is noted that hyphen is used in the places where a *dash* is a legitimate candidate.

6. MULTIFUNCTIONALITY OF HYPHEN IN BANGLA TEXT CORPUS

Not much study has been done empirically on the patterns of usage of a hyphen in Bangla texts, although there have been some intuitive attempts for understanding its nature and function in the language. For instance, Roy (1989: 137) has presented a short discussion on its use in 19th century Bangla prose texts, while Bhattacharya (1965), Bhattacharya (1992), Chatterji (1926: 211), Chatterji (1974), Chakrabarti (1994: 78), Bhattacharya (1999: 57-58), Ray (1997) and Bhattacharya (2000) have primarily focussed on its use in the present Bangla texts. All these studies have primarily been dependent either on the intuitive evidence hatched by the investigators or on small text samples designed specifically to gather instances and formulate observation. Since there was no large multidisciplinary Bangla text corpus available to the scholars, there has never been any attempt to understand the multi-functionality of a hyphen in the language with reference to its actual usages varieties as reflected in the empirical text databases of the language. In this respect, the present study reported here is probably the first effort of its kind where I have made an attempt to understand the role of a hyphen in the language with examples collected directly from a written Bangla text corpus containing a lot of empirical text samples representing more than eighty-five subject domains. I argue that the observations presented in this study will be highly useful for Bangla language teaching, verifying previously made observations, as well as for designing tools and techniques for Bangla language processing.

The use of the hyphen, with regard to the use of other punctuation marks in the language, is found to be the most complex phenomenon in the Bangla texts corpus. The guidelines proposed in widely acclaimed dictionaries are not sufficient enough to define the roles of a hyphen in the Bangla texts since there is no regularity in its use and its function in the language. In general, traditional dictionaries inform us that hyphen is used in the text for the following four functions:

(a) Joining words together to make compound words,

(b) Implying division of words at the end of a line,

(c) Indicating a missing or implied element within words, and

(d) Getting some idea of the meaning of words.

However, all these functions constitute only a part of the varieties it exhibits in the Bangla texts. The use of the hyphen in Bangla compound words is arbitrary, especially when elements of compounds are made of a single syllable. Except for some unavoidable situations, it is used randomly: in some occasions, it is not used although it was required; on the contrary, it is used in certain contexts where it is not required. I have noted more than thirty (30) variations of use of the hyphen in Bangla texts for carrying out different linguistic functions, namely, orthographic, lexical, grammatical, syntactic, semantic, and discourse (Fig. 1).

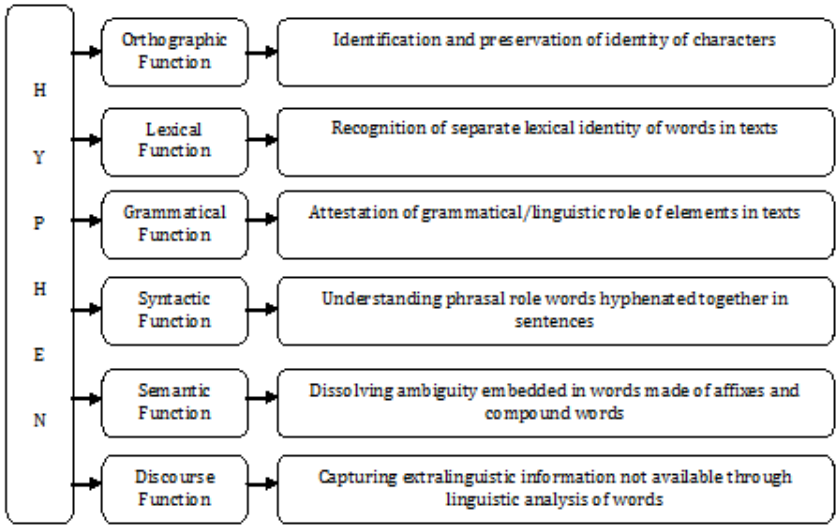

Fig 1: *Classification of function of hyphen used in the written Bangla text corpus*

I shall try to address each type of use of the hyphen in the language in the following sub-sections with examples taken from the Bangla text corpus.

6.1 Orthographic Function

The use of the hyphen in between a proper name and a suffix or a case ending is a regular practice in Bangla where the goal is to keep the proper names unchanged in their orthographic representation. This shows that a hyphen mark is used in between a proper name and an inflection mark, such as, *kālidās-er* "of Kalidas", *sṭeṭsmyān-e* "In the Statesman", *sombār-e* "on Monday", *ṭeligrāph-er* "of Telegraph", *deś-er* "of Desh", *ājkāl-er* "of Ajkal", *bartamān-er* "of Bartaman", etc., where hyphen serves as a device to keep the original orthographic forms of the proper names unaffected when there is a chance for adding case markers with these words. This kind of use of hyphen has been a recent trend in Bangla writing practice, which is clearly manifested in the Bangla text corpus. However, there is no legitimate reason behind such practice as it does not supply any extra information except giving some ideas about the actual orthographic forms of the proper names involved in suffixation or inflection. The normal practice of writing in Bangla is that all the proper names and words are joined with a suffix or a case ending without the use of a hyphen, such as *kālidāser, sṭeṭsmyāne, sombāre, ṭeligrāpher*, *deśer*, *ājkāler*, *bartamāner*, etc.

A hyphen mark is consciously deployed to retain the original shape of some Bangla nouns (not proper names) unaffected when these nouns are tagged with a suffix or an inflection, because without this practice the lexico-semantic identity of the nouns will be greatly ambiguous, as shown below:

- pad-er "of lexeme",
- pā-ṭi "the leg"
- ka-ṭā "how many"
- sai-ṭā "the signature"
- bau-er "of wife"
- mā-rā "mother and others", etc.

This is also not a regular practice of writing nouns in Bangla text. Generally, a case marker or an inflection is tagged at the end of these words without inserting a hyphen mark, such as *pader, pāṭi, kaṭā, saiṭā, bau-er, mārā,* etc.

When some English words are transliterated into Bangla, a hyphen mark is normally used between the English word and the Bangla case marker or suffix that is supposed to be tagged to it, as shown below.

- roḍ-e "on road"
- mesin-ke "to machine",
- phon-er "of phone",
- ḍren-e "in drain",
- rum-er "of room"
- mal-e "in mall", etc.

This is again not a regular practice for writing English words tagged with Bangla case endings, since in most cases, English words are usually tagged with Bangla case endings without a hyphen in between, as in, *skuler* "of school", *klāse* "in class", *sṭeśane* "at station", *meśiner* "of machine", *male* "in the mall", etc.

Most of the English group verbs, in their transliterated forms, are written in Bangla with a hyphen mark which is put between the two formative forms, e.g., *pik-āp* "pick up", *bāi-pās* "by pass", *mek-āp* "make up", *phalo-an* "follow on", etc. The use of the hyphen in transliterated English group verbs is highly useful as it helps to identify them as single units having special sense or meaning.

When a case marker is attached to a word which ends with the consonant grapheme *khaṇḍa-ta* (part-ta), a hyphen is inserted in between the case marker and the word end just to retain the word-final consonant grapheme

unchanged in form, e.g., *kaiphiyaṯ-er* “of answer”, *bhabiṣyaṯ-er* “of future”, *ijjaṯ-er* “of prestige”, *bajjāṯ-der* “to the obstinate people”, etc. There is a justification for adding a hyphen in these words, i.e., the *part-ta* cannot take any vowel allograph. Therefore, we have two options: either we have to use a hyphen to retain *part-ta* or we have to change *part-ta* into *full-ta* if we want to remove hyphen. The addition or removal of a hyphen does not cause any change in pronunciation and meaning of the words. When the hyphen is removed, the word-final consonant grapheme *khaṇḍa-ta* (part-ta) is changed into the normal consonant grapheme *full-ta* when the case marker or the suffix is tagged to the word, as in, *kaiphiyaṯer, bhabiṣyaṯer, ijjaṯer, bajjāṯder,* etc. Although it is known to us that replacement of *khaṇḍa-ta* (part-ta) with *full-ta* does not affect the meaning of the words in Bangla, in one instance at least, I have found that both *full-ta* and *khaṇḍa-ta* are not mutually inclusive, since replacement of one by the other can cause change in meaning of the words, e.g., *parabhṛt* (with *khaṇḍa-ta*) "crow" : *parabhṛta* (with *full-ta*) "cuckoo".

In the case of acrostic words, the use of the hyphen is a common practice in Bangla text writing. This gives an explicit advantage to the readers to understand that the acrostic words are actually made with several abbreviated forms, as the following examples show:

- bi-bi-si (British Broadcasting Corporation)
- āi-es-āi (Indian Statistical Institute)
- āi-āi-ṭi (Indian Institute of Technology)
- āi-āi-em (Indian Institute of Management)
- ṭi-āi-eph-ār (Tata Institute of Fundamental Research)
- āi-em-eph (International Monetary Fund) etc.

Hyphen is also used in between an awkward sequence of identical consonant graphemes across words, as noted in the following examples:

- tāp-prabar "heat-tempo",
- bābā-bāchā “utmost request”
- kāch-chāṛa “removing from side”,
- pāt-tāṛi “belongings”, etc.

The hyphen used to break a word at the end of a line is another crucial orthographic function, although it is not a regular feature of the spelling of words of a language. It is more common in the case of printed texts where words are broken carefully and consistently taking into account their appearance and structure. It is normally done to space texts accurately and to justify margins for proper text alignment. Since this is not a mandatory feature of a written text, with some care, it may be avoided totally in hand-written, typed, and word-processed texts.

6.2 Lexical Function

The use of hyphen for the formation of compound words constituting two or three words having different sense has been an age-old practice in Bangla writing system. Gradually, its use is extended to routine and occasional couplings of words especially when a reference to the sense of separate elements is considered important and unavoidable. Thus, hyphen is used between the following strings:

- Compounds constituting two nouns, e.g., *cor-ḍākāt* "thief and robber",
- Two adjectives, e.g., *rogā-moṭā* "thin and thick" ,
- A noun and an adjective, e.g., *man-gaṛā* "fabricated",
- A pronoun and a noun, e.g., *sei-din* "that day",
- Two proper names, e.g., *śelī-kīṭs* "Shelly and Keats",

- Words of similar meaning, e.g., *ṭākā-paysā* "notes and coins",
- Department and post, e.g., *kṛṣi-mantrī* "agriculture minister",
- Institution and designation, e.g., *skul-māsṭār* "school teacher",
- Place and occasion, e.g., *bārlin-alimpik* "Berlin Olympic",
- Two directions, e.g., *uttar-paścim* "North-West",
- Single-letter nouns and multi-letter nouns, e.g., *bhū-prakṛti* "geo-topography"
- Cardinal adjective and a noun, e.g., *du-belā* "two times",
- Frozen form (idiomatic expression), e.g., *mā-māṭi-mānuṣ* "mother-earth-people", etc.

The use of a hyphen between a prefix and a noun has a specific lexical function in the language. It clearly indicates that since the prefix is not an integral part of the word, it is used with a hyphen in between the two forms to recognize its separate lexical identity, as the following examples show:

- ku-najar "bad sight"
- su-samay "good time"
- a-baśya "indomitable"
- be-ijjat "non-prestige"
- bad-svabhāb "bad habit"
- gar-hājir "not present", etc.

Again, hyphen is found to be used between reduplicated words, onomatopoeic words, and echo words in Bangla with clear indication that at the time of text or word analysis, all these hyphenated strings should be considered as single lexical entry and be treated accordingly, as the following examples show:

- mājhe-mājhe "sometimes"
- jane-jane "to every person"
- biṣ-ṭis "poison and others"
- jhan-jhan "tinkling sounds"
- ban-ban "in high speed"
- bai-ṭai "books and others things", etc.

At the lexical level hyphen is also used between numeric combination as well as between alpha-numeric combinations, such as the followings:

- Two numbers, e.g., 2-3, 12-14,
- Two separate years, e.g., 1986-1990,
- A number and a word, e.g., 6-gaj "6-yard", 4-hāt "4-hands, 2-mānuṣ "2-men",
- A number and a particle, e.g., 1-ṭi "one", 2-ṭo "two", 3-ṭe "three", 7-ṭā "seven", etc.
- A number and a suffix, e.g., 4-ṭhā "fourth", 19-śe "on 19th", 3-rā "third", etc.

6.3 Grammatical Function

The use of hyphen has been instrumental for grammatical analysis of various linguistic items (e.g., letters, graphemes, allographs, stems, affix, diacritics, etc.) used in Bangla text. The presence of hyphen in these contexts

serves in two ways: it works as a mark of their identity and it confirms their contextual relevance in the texts. A language investigator can, with the help of the hyphen, understand that the linguistic items are used in the text for specific linguistic analysis and interpretation.

For linguistic analysis when some inflected words are used as nouns after the addition of further suffix markers with them, a hyphen is used in between the inflected form and the final suffix marker to highlight its distinct lexico-grammatical function, e.g., *ādhunik Bāṅglāy 'moder-er' byabahār prāy nei ballei cale* "The usage of 'moder-er' is almost redundant in modern Bangla", etc.

Generally, all affixes are tagged with a hyphen for recognizing their specific linguistic identity. While prefixes are tagged with hyphen immediately after them; suffixes are tagged with it immediately before them, and infixes are normally tagged with a hyphen immediately before and after them as the following examples show. A hyphen works as a mark of identification of affixes or dependent morphemes, and it performs a very clear linguistic function. This function, however, not a special feature of Bangla but true to English and other languages where affixes are used as word-formation properties.

Prefix : pra-, bi-, su-, ati-, prati-, etc.,

Suffix : -guli, -der, -ke, -te, -er, -re, -e, etc.

Infix : -a-, -i-, -u-, -o-, etc.

When Bangla graphemes and other orthographic symbols are used for grammatical analysis, a hyphen mark is normally used between the symbol and the suffix tagged with it, e.g., *a-ṭā bāṅglā barṇamālār pratham barṇa* "The vowel grapheme 'a' is the first letter in the Bangla alphabet", *h-ṭā antim barṇa* "'Ha' is the last letter", etc.

At the time of grammatical and linguistic analysis all the Bangla vowel graphemes and their allographs are always written with a hyphen in between the member constituents, as in, *u-kār* "u-allograph", *e-kār* "e-allograph", *o-kā*r "o-allograph", etc. Here hyphen performs a lexico-grammatical function to indicate that the allograph is linked with the vowel grapheme with which it is tagged with the hyphen.

Some Bangla consonant graphemes are almost similar in pronunciation. Therefore, such homophonous consonant graphemes, as well as their places of articulation, are normally written with a hyphen in between, such as the followings:

- mūrdhanya-ṇa “retroflex-ṇ”
- dantya-na “dental-n”
- tālabya-śa “palatal-ś”
- mūrdhanya-ṣ “retroflex- ṣ”
- dantya-sa “dental-s”, etc.

In these cases, there is a clear need for using a hyphen, because if we delete it from the sequence, there will be a problem in identification and understanding of the characters and their places of articulation.

Another unique usage of hyphen in the Bangla text is noted where it works in a manner similar to that of a postposition, as the following examples show:

- se 5-10 julāi bāṛite thākbe “He will stay at home from 5 to 10 July”,
- iskul 15-31 jānuāri bandha thākbe “School will remain close from 15 to 31 January”
- *e bachar pujor chuṭi 5-25 akṭobar paryanta paṛeche* "Puja leave this year is from 4 to 25 October".

Here hyphen acts just as a postposition because we have to use the Bangla postposition *theke* "from" at the same place if we want to remove the hyphen. In such cases, the question is whether we should tag the hyphen as a postposition or a punctuation.

6.4 Syntactic Function

In the case of a syntactic function, hyphen works as a syntactic link to provide a sense of continuation of meaning expressed within a sentence. It also saves a word from its repetitive use within a sentence. Thus, a hyphen serves to connect those words, which when combined with a hyphen between them, exhibit a kind of unbroken syntactic linkage similar to a sentence, as the following examples show:

- dakṣiṇer-dolā-lāgā-pākhi-jāgā-basanta prabhāte "In the spring morning shaken by swinging breeze of the south and awaken by birds call",
- kathāy-kathāy-rāg-karā mejāj "to-be-angry-with-every-word temper"
- nā-bāla-kathār-gopan-gabhīr-bedanā theke jege oṭhā "arising from the profound and concealed pain of un-spoken words", etc.

Sometimes a hyphen is used to link up the constituents of phrases with subsequent words, which are combined together to be used attributively in a sentence, as noted in the following examples.

- hājār-hāt-kālī "Goddess Kali with thousand cut-off hands",
- paśu-maṭsya-śikār "hunting of animals and fish",
- paṛe-pāoyā-caudda-ānā "easily got unexpected money",
- brek-iven-payenṭ "break-even-point",
- nagad-bhittik-lenden "cash-based transaction",
- kām-bajra-yān "lust-strong-vehicle", etc.

When the second member of a series of compound words is common, a hyphen is used in the place of the second word to avoid its regular repetition in the sentence. The second word is actually used with the last member of the series of compounds in the sentence. In this case, a hyphen performs the role of an elided word, such as the followings:

- sārā rājya juṛe śramik -, bekār -, bidyut -, khādya -, paribahaṇ -, o jal samasyā dekhā diyeche "The problems of labour, unemployment, electricity, food, transport, and water have cropped up in the state".

In this case, a hyphen has a real syntactic relevance, as it directly refers to the elision of the second element of a compound in a syntactic environment. The second word being common to all compounds is removed and the empty place is filled up with a hyphen.

6.5 Semantic Function

The pattern of use of hyphen can have a large impact on the 'meaning' of any particular form as variation in use of the hyphen in words can generate several different readings (Quirk *et al.* 1985). There has never been any attempt to understand the semantic function of a hyphen in Bangla text. It is noted that its use at some specific juncture within some words can indicate a syllabic pause which is instrumental in proper comprehension of the structure of words as well as in understanding their actual meanings. This is the first attempt to explore how its presence or absence within words can cause their meaning variations.

Hyphen is most often used between words and emphatic particles for extra emphasis, as the following examples show. The most notable thing is that the removal of hyphen will produce a form where tow identical vowel graphemes are used -- a phenomenon quite unnatural in the norm of writing Bangla words. Therefore, the use of the hyphen is inevitable in these places. Although the extra emphasis is indicated through the use of the emphatic particle, it has become intensified and visually appealing due to the presence of the hyphen. Thus the use of hyphen here transcends across semantic-cum-graphemic levels of the words.

- tumi-i "you yourself",
- ami-i "I myself",

- sei-i "he himself",
- mā-i "mother herself",
- se-o "he also",
- āro-o "too much", etc.

The use of the hyphen, in some words, not only helps us to find actual meaning of words but also stops us from deducting wrong meaning from the words. Following are some instances (Table 2) where presence or absence of hyphen can change the meanings of the words in Bangla

Table 2: *Change of meaning of words with/without hyphen*

Without hyphen	Meaning	With hyphen	Meaning
atithi	guest	a-tithi	non-occasion
abas	collapsed	a-bas	not in control
amaik	amiable	a-maik	without microphone
asukh	Illness	a-sukh	non-happiness
kaṭā	brownish yellow	ka-ṭā	how many
pāṭā	plank	pā-ṭā	the leg
amṛita	nectar	a-mṛrita	not dead
ākār	shape	ā-kār	ā-allograph
cāṭā	lick	cā-ṭā	the tea
ekār	alone	e-kār	e-allograph
mār	Kill/hit	mā-r	of mother
kuśāsan	the seat of Kush grass	ku-śāsan	bad ruling

Even mere displacement of hyphen within words can cause variation in meaning for some words. The role of a hyphen in these contexts is measured in its usage within the words for dissolving ambiguities in meaning. For instance, some Bangla words, such as *surataraṅga, akhyātanāmā, narakapāl, gaṇakabar*, etc. are ambiguous in a sense as they can be hyphenated at two different places to generate two different senses. In the following examples, for elucidation, a hyphen mark is displaced from one juncture to another within the words to show how two different meanings are generated (Table 3). Hyphen serves here as a marker for dissolving ambiguity embedded within surface forms since its placement helps to decipher the actual meaning of the words.

Table 3: *Variation in meaning due to displacement of hyphen within words*

Hyphenated words	Utterance	Glossary
sura-taraṅga	[ʃur-tɔrɔŋgo]	a wave of music
surata-raṅga	[ʃurɔt-rɔŋgo]	coital fun
a-khyātanāmā	[ɔ-khætonama]	not famous
akhyāta-nāmā	[ɔᵏkhæto-nama]	notorious
nara-kapāl	[nɔro-kɔpal]	human forehead
naraka-pāl	[nɔrɔk-pal]	the king of hell
gaṇa-kabar	[gɔno-kɔbɔr]	mass graveyard
gaṇaka-bar	[gɔnɔk-bɔr]	the best astrologer

The above examples (Table 2 and Table 3) show that the analysis of the semantic function of the hyphen can serve as a tool for dissolving lexical ambiguity embedded within words. In the case of morphological processing, words sense disambiguation, automatic text-to-speech conversion, recognition of functional behaviour of hyphen within words can help us in proper identification of syllabic length and pause which might eventually lead in designing systems that can extract as well as recognise the actual meaning of words without human intervention.

6.6 Discourse Function

Hyphen is sometimes used in between two consecutive words, which are not actually compounds but which, due to their peculiar co-occurrence, denote a sense of hesitation, appeasing, mode of action, pun, etc. as the following examples show:

- Hesitation : e.g., dicchi-debo "dilly-dally",
- Request : e.g., bābā-bāchā "appeasing",
- Callousness : e.g., dicchi-debo "dilly-dally",
- Hurriedness : e.g., sāt-tāṛātāṛi "in a haste",
- Pun : e.g., be-heḍ "without head"/"shameless", etc.

The hyphen is used to denote a sense of continuation of sounds produced in excitement or exultation, as in, *co-o-o-o-r, co-o-o-o-r* “the thief”, *āmi khāba nā-ā-ā-ā-ā-ā* "I shall not eat", etc. In such cases, the emotional outbursts due to anger or fear are best represented through the use of a hyphen between the segments for indicating the continuation of the stream of emotion in the expression. In fact, the lack or absence of a hyphen in such cases will not be able to create the expected impact on the readers.

A hyphen is also used to indicate sounds produced by the musical instrument, as in, *tā-dhin-dhin-tā*, *sā-ā-ā-ā, re-e-e-e, gā-ā-ā-ā*, *do-re-mi-fa-so-la-ti-do*, etc. Here again, a hyphen performs an extralinguistic function due to which it is possible to understand the significance of the sounds in specific discourse context.

The hyphen is used to write proper names that have multiple word strings (mostly for Arabic, Persian, and Chinese proper names) in Bangla script. This is a new practice of using a hyphen in proper names in Bangla which is not noted in the case of their English transliteration, such as, *nāzim-ud-daulā, māo-se-tuṃ, imdād-ul-haq, rābāi-ben-ejrā, ciyān-kāi-shek, mahammad-bin-tughlak,* etc.

Another interesting use of hyphen is noted in a piece of text when the actual name of an individual is deliberately represented by a consonant grapheme to hide his/her identity, as in, *ka-bābu balte lāglen, kha-bābu śunte lāglen, ār ga-bābu ghumote lāglen* "Mr. X started speaking, Mr. Y started listening, and Mr. Z started sleeping", etc.

Hyphen is often used in case of idiophones (onomatopoeic words) where the emotion of the person concerned is expressed with elegant use of hyphen, such as the followings:

- hu-hu kānnā "bursting cry"
- cham-chame bhay "shivering fear"
- cin-cine byathā "piercing pain"
- ri-ri gā "irritating body"
- hi-hi hāsi "flowing laughter"
- tham-thame mukh "grimacing face", etc.

Perhaps, removal of the hyphen from these words would not have much dampened the mood of the persons expressed by these hyphenated words.

Finally, a hyphen is used within an exclamatory word to express the high rate of emotional load tagged with it, as in, *ho-yāṭ* "what" *sā-bbās* "bravo", *mā-go* "Oh my God", *du-cchāi* "shit!", *o-yā-o* "wow", *śā-lā* "stupid", etc.

7. ANNOTATING HYPHENATED WORDS

In the case of hyphenated words, we have to be careful at the time of POS tagging as several morpho-semantic issues are involved in them. I argue for adopting three different approaches to deal with such word strings.

7.1 First Approach

In case of those words where a formative element (e.g., inflection, particle, case marker, etc.) is separated from the word with a hyphen, it is better to tag the entire hyphenated word as a single lexical unit, since these formative elements are actually the part of the base form, and therefore, do not need to be separated with a hyphen. Moreover, even if the hyphen mark is removed, the original sense or meaning of the word is hardly affected, as the

examples given below (Table 4). In most cases, the ambiguity generated therein may be solved with reference to the local context where the word is actually used and put into a chunk.

Table 4: *Removal of hyphen at the time of POS annotation without hampering meaning*

Original Form	Revised Form	POS	Annotation	Gloss
ho-yāṭ	hoyāṭ	Indeclinable	hoyāṭ/NN/	"what"
mā-i	māi	Noun	māi/NN/	“mother herself”
Kālidās-er	Kālidāser	Noun	Kālidāser/NN/	“of Kalidas”,
Sṭeṭsmyān-e	Sṭeṭsmyāne	Noun	Sṭeṭsmyāne/NN/	“in Statesman”,
Sombār-e	Sombāre	Noun	Sombāre/NN/	“on Monday”,
pad-er	pader	Noun	pader/NN/	“of lexeme”,
Deś-er	Deśer	Noun	Deśer/NN/	“of Desh”,
mā-r	mār	Noun	mār/NN/	“of mother”,
cā-ṭā	cāṭā	Noun	cāṭā/NN/	“the tea”,
pā-ṭi	pāṭi	Noun	pāṭi/NN/	the leg”,

7.2 Second Approach

On the other hand, in the case of those word forms, where the hyphen is used between two potentially individual lexical items, which are capable of independent use, it is sensible to tag the words as well as the hyphen as separate linguistic entities, because here hyphen is a legitimate functional connector between the words. For instance, consider the following types (Table 5). In such cases, hyphen itself needs to be tagged separately with a separate tag meant for punctuation.

Table 5: *Preserving hyphen at the time of POS annotation without hampering meaning*

Original Form	Revised Form	POS	Annotation	Gloss
bhū-prakṛti	bhū-prakṛti	Noun	bhū/NN - prakṛti/NN/	“geo-nature”
ku-svabhāb	ku-svabhāb	Noun	ku/JJ/- svabhāb/NN/	“bad habit”
chu-mantar	chu-mantar	Noun	chu/JJ/ - mantar/NN/	“touch magic”
pik-āp	pik-āp	Finite verb	pik/FV/ - āp/PREP	“pick up”
u-kār	u-kār	Noun	u/NN/ - kār/NN/	“u-allograph”
cor-ḍākāt	cor-ḍākāt	Noun	cor/NN/- ḍākāt/NN/	“thief and robber”
śelī-kīṭs	śelī-kīṭs	Noun	śelī/NN/- kīṭs/NN/	“Shelley and Keats”
uttar-paścim	uttar-paścim	Noun	uttar/NN/- paścimNN/	“north-west”
rogā-moṭā	rogā-moṭā	Adjective	rogā/JJ/- moṭā/JJ/	“thin and thick
man-gaṛā	man-gaṛā	Adjective	man/NN/- gaṛā/JJ/	“fancy-made”

7.3 Third Approach

In case of those multiword strings, which are actually used as hyphenated phrases, it is sensible to tag each word in its relevant part-of-speech and combine them together as a chunk with special meaning and mark them accordingly without disturbing the position and use of the hyphens between the words, as the following examples show (Table 6). In such cases, hyphen itself needs to be tagged separately with a separate tag meant for punctuation.

Table 6: *Preserving hyphen at the time of POS annotation in multiword phrasal strings*

Original Form	Annotation	Gloss
dakṣiṇer-dolā-lāgā-pākhi-jāgā-basanta prabhāte	dakṣiṇer/NN/ - dolā/NN/ - lāgā/jJJ/ -pākhi/NN/ - jāgā/JJ/ - basanta/NN/ prabhāte/NN/	In spring morning shaken by swinging breeze of south and awaken by birds call
kathāy-kathāy-rāg-karā mejāj	kathāy/NN/ - kathāy/NN/ - rāg/NN/ - karā/JJ/ mejāj/NN/	To-be-angry-with-every-word-temper
nā-bāla-kathār-gopan-gabhīr-bedanā theke jege oṭhā	nā/IND/- bāla/FV/ - kathār/NN/- gopan/JJ/ - gabhīr/JJ/ - bedanā/NN/ theke/PP/ jege/FV/ oṭhā/JJ/	Arising from profound & concealed pain of un-spoken words”
hājār-hāt-kālī	hājār/NN/ - hāt/NN/ - kālī/NN/	Goddess Kali with thousand cut-off hands
paśu-maṭsya-śikār	paśu/NN/ - maṭsya/NN/ - śikār/NN/	Hunting animals and fish
nagad-bhittik-lenden	nagad/NN/ - bhittik/JJ/ - lenden/NN/	Cash-based transaction

8. CONCLUDING REMARKS

The blank space (or the empty space), which provides a spatial gap between the words in a sentence is not usually considered as a punctuation mark. However, after analysis of its function in texts, I argue that we should treat is as a punctuation mark because what is a measured pause in spontaneous speech is a calculated empty space in a piece of written text. While a pause differentiates a speech from an unbroken trail of sound, an empty space differentiates a sentence from a continuous string of characters stuck together without a gap in between. The management of empty space in writing is a crucial issue because its presence or absence between two or more consecutive words can affect the meaning of the words as well as our linguistic understanding of the expressions.

The development of a language script has a strong impact on the evolution of thought processes as well as on the enhancement of cognitive powers of a human society. A script is a form of knowledge representation that uses alphabets, diacritics, and other characters to encode and decode knowledge, and to convert auditory sounds into visual symbols so that the members of a speech community are able to capture the wisdom of the generations in visual form. The study of script of any language is, therefore, not only useful for understanding the linguistic behaviour of the people of a particular language community but also important for exploring their linguistic-cognitive interface that empower them to express events and concepts, knowledge and information, ideas and imagination through the accepted set of linguistic symbols and characters intelligible to them across generations.

In this context, the study of the patterns of usage of hyphen, similar to that of other characters used in the script, attracts our attention because it reflects into the cognitive process the members of a speech community deploy to construct ideas, share information, and embed knowledge for cross-fertilization of the thinking process of the members of the society. The information gathered from the diversity of use of hyphen in texts not only enriches the language users about its diverse usage varieties, but also supplies important insights to understand how, at the time of composing a piece of text, one has to be careful in appropriate use of the symbol, so that rather than being a simple decorative character, it can carry additional linguistic and discourse information with the support of which proper understanding of a text becomes simple. therefore the use of the hyphen is an important strategy of text composition, which should be properly studied, explained, and understood by the language users. The descriptive value of this study will transcend into the application when the language teachers use data and information from this study to teach learners the right uses of a hyphen in texts and when grammar books are written with instructions for its proper use by the language users.

REFERENCES

Bayraktar, M., B. Say and V. Akman (1998) An Analysis of English Punctuation: The Special Case of Comma. *International Journal of Corpus Linguistics.* 3(1): 33-58.

Bhattacharya, N. (1965) *Some Statistical Studies of the Bangla Language.* Doctoral Dissertation. Indian Statistical Institute. Kolkata. (MS).

Bhattacharya, S. (1992) *Bangla Uccharan Abhidhan (Bangla pronunciation dictionary).* Kolkata: Sahitya Sansad.

Bhattacharya, S. (1999) *Tistha ksanakal: biram cihna o anyanya prasanga (Pause for a while: punctuation marks and other issues).* Kolkata: Ananda Publishers.

Bhattacharya, S. (2000) *Bangalir Bhasa (The language of the Bangali).* Kolkata: Ananda Publishers.

Blake, G. and R. W. Bly (1993) *The Elements of Technical Writing*, New York: Macmillan Publishers (pg. 48).

Chakrabarti, N.N. (ed.) (1994) *Bangla: ki likhben, kena likhben (Bangla: what to write and why to write)*. Kolkata: Ananda Publishers.

Chatterji, S.K. (1926) *The Origin and Development of the Bangla Language.* Kolkata: Calcutta University Press. Reprinted by Rupa Publications, Calcutta in 1993.

Chatterji, S.K. (1974) *Bangala Bhasatattver Bhumika (An introduction to Bangla linguistics)*. Kolkata: Calcutta University Press.

Chatterji, S.K. (1993) *Bhasaprakash Bangala byakaran (Grammar of the Bangla language)*. Kolkata: Rupa Publications.

Crystal, D. (1995) *The Cambridge Encyclopaedia of the English Language.* Cambridge: Cambridge University Press.

Das, G., S. Bhattacharya and S. Mitra (1984) Representing Asamia, Bangla, and Manipuri Text in Line Printer and Daisy-Wheel Printer. *Journal of the Institution of Electronics and Telecommunication Engineers.* 30: 251-256.

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

Dash, N.S. and B.B. Chaudhuri (1998) Bangla Script: A structural Study. *Linguistics Today*. (2)1:1-28.

Houston, K. (2014) *Shady Characters: The Secret Life of Punctuation, Symbols, and Other Typographical Marks.* New York: W. W. Norton & Company.

Mallick, B.P. (2000) (ed) *Shes lekha: Linguistic Statistical Analysis.* Kolkata: Bangla Academi.

Mallik, B.P. and T. Nara (eds.) (1994) *Gitanjali: Linguistic Statistical Analysis.* ILCAA: Tokyo University.

Mallik, B.P. and T. Nara (eds.) (1996) *Sabhyatar Sankat: Linguistic Statistical Analysis.* Kolkata: Rabindra Bharati University Press.

Miller, G.A, E.B. Newman and E.A. Friedman (1958) Length-Frequency Statistics for Written English. *Information and Control*. 1: 370-389.

Miller, G.A. (1951) *Language & Communication.* New York: McGraw-Hills.

Parkes, M.B. (1993) *Pause and effect: an introduction to the history of punctuation in the West.* Berkeley: University of California Press.

Quirk, R., S. Greenbaum, G. Leech, & J. Svartvik (1985) *A Comprehensive grammar of the English language.* London ; New York : Longman.

Ray, P.S. (1997) *Bengali Language Handbook*. Kolkata: Bangla Akademi.

Roy, A.K. (1989) *Unish shataker Bangla Gadya : Ingreji prabhab (Bangla prose in the 19th century: the impact of English)*. Kolkata: Jignasa Publications.

AUTHOR'S BIOGRAPHY

Dr. Niladri Sekhar Dash is an Associate Professor in the Linguistic Research Unit, Indian Statistical Institute, Kolkata, India. For last 23 years, he is working in corpus linguistics, natural language processing, language technology, computational lexicography, and language digitization and documentation. To his credit, he has published 15 research monographs and more than 150 research papers in the areas of his expertise. As a visiting faculty, he has delivered lectures in more than 25 universities and institutes in India and abroad. Details: https://sites.google.com/site/nsdashisi/home/