J Psycholinguist Res
DOI 10.1007/s10936-014-9302-x

Computational Modeling of Morphological Effects in Bangla Visual Word Recognition

Tirthankar Dasgupta · Manjira Sinha · Anupam Basu

Abstract In this paper we aim to model the organization and processing of Bangla polymorphemic words in the mental lexicon. Our objective is to determine whether the mental lexicon accesses a polymorphemic word as a whole or decomposes the word into its constituent morphemes and then recognize them accordingly. To address this issue, we adopted two different strategies. First, we conduct a masked priming experiment over native speakers. Analysis of reaction time (RT) and error rates indicates that in general, morphologically derived words are accessed via decomposition process. Next, based on the collected RT data we have developed a computational model that can explain the processing phenomena of the access and representation of Bangla derivationally suffixed words. In order to do so, we first explored the individual roles of different linguistic features of a Bangla morphologically complex word and observed that processing of Bangla morphologically complex words depends upon several factors like, the base and surface word frequency, suffix type/token ratio, suffix family size and suffix productivity. Accordingly, we have proposed different feature models. Finally, we combine these feature models together and came up with a new model that takes the advantage of the individual feature models and successfully explain the processing phenomena of most of the Bangla morphologically derived words. Our proposed model shows an accuracy of around 80 % which outperforms the other related frequency models.

Keywords Mental lexicon · Morphological decomposition · Masked priming · Visual word recognition · Frequency effects · Suffix productivity

Introduction

The term *mental lexicon* refers to the access, representation and processing of the words in the human mind and the various associations between them that help fast retrieval and

T. Dasgupta (✉) · M. Sinha · A. Basu
Department of Computer Science and Engineering,
Indian Institute of Technology, Kharagpur 721302, West Bengal, India
e-mail: iamtirthankar@gmail.com

Published online: 02 July 2014

comprehension of the words in a given context (Aitchison 2005; Marslen-Wilson et al. 1994; Taft and Forster 1975). Words are known to be associated with each other at various levels of linguistic structures namely, orthography, phonology, morphology and semantics. However, the precise nature of these relations and their interactions are unknown. Understanding the organization of the mental lexicon is one of the important goals of cognitive science. A clear understanding of the structure and the processing mechanism of the mental lexicon will further our knowledge of how the human brain processes language. Further, these linguistically important and interesting questions are also highly significant for computational linguistics (CL) and natural language processing (NLP) applications. Their computational significance arises from the issue of their storage in lexical resources like WordNet (Fellbaum 2010) and raises important questions like, how to store morphologically complex words, in a lexical resource like WordNet keeping in mind the storage and access efficiency.

One of the key issues on which psycholinguists have been investigating for a long time is the representation and processing of morphologically complex words in the mental lexicon. That is to say, whether for a native speaker, a polymorphemic word like "unpreventable" will be processed as a whole or will it be decomposed into its individual morphemes "un-", "prevent", and "-able" and finally recognized by the representation of its stem (morphemic model). It has been argued that people certainly have the capability of such decomposition since they can understand novel words like "unsupportable". However, there has been a long standing debate whether such decompositions are obligatory (i.e morphemic) or are they applicable to only those situations where the whole word access fails (Taft 2004) (partial decomposition model). An alternative to the morphemic and partial decomposition model is the full listing model that assumes decomposition is not at all involved and initial processing of words are performed in terms of the whole word representation in the mental lexicon (Burani and Caramazza 1987; Burani and Laudanna 1992; Caramazza et al. 1988). Such issues are typically addressed by designing appropriate priming experiments (Frost et al. 1997; Aitchison 2005) or other lexical decision tasks.

Priming results in faster recognition of a stimulus (called the, target) based on the previous exposure of another stimulus (called the prime). Therefore, if the prime and the target words are morphologically related (say, MANLY and MAN), then going by the decomposition model, as soon as the prime (MANLY) is presented to a subject, it will be decomposed into its constituent stem (MAN) and the suffix (-LY) and be recognized individually. Thus, recognition of the target word starts well before it is presented to the subject. Naturally, this will result in a faster recognition of the target as compared to the case when the target is preceded by a morphologically unrelated word (say, MOTHER and MAN) where no such decomposition of the prime is possible. On the other hand, considering full-listing model, recognition of the target must be independent of the prime. Thus, time to recognize the target MAN preceded by the prime MANLY must be equal to the case when it is preceded by MOTHER. Hence, if priming by a morphologically related word results in faster recognition of the target, it may be assumed that decomposition has played its role.

The priming experiments can be classified according to the mode of representing the prime and target words: (a) when both are visually presented (Bentin and Feldman 1990; Ambati et al. 2009; Frost et al. 1997; Marslen-Wilson et al. 2008), (b) primes are auditorily presented but the targets are visually presented (Marslen-Wilson et al. 1994; Marslen-Wilson and Tyler 1997; Marslen-Wilson and Zhou 1999; Marslen-Wilson et al. 2008), (c) targets are auditorily presented but the primes are visually presented (Marslen-Wilson et al. 1994). These experiments demonstrate that across the languages, recognition of a target word (say happy) is facilitated by a prior exposure of a morphologically related prime word (e.g., happiness). Since morphological relatedness often implies orthographic, phonological and

semantic similarities between two words, several attempts have been made to factor out other priming effects from morphological priming (Bentin and Feldman 1990; Drews and Zwitserlood 1995).

The masked priming paradigm, where the prime word is placed in between a forward mask and a target word such that it cannot be consciously perceived (Bodner and Masson 1997; Davis and Rastle 2010) also shows some interesting ways of examining morphological effects in word recognition (Forster and Davis 1984). Through such experiments morphological priming effects are shown to exist in the absence of semantic priming for Hebrew (Frost et al. 1997), phonological priming (Crepaldi et al. 2010), and orthographic priming for French (Grainger et al. 1991) and Dutch (Drews and Zwitserlood 1995).

A cross modal priming experiment has been conducted for Bangla derivationally suffixes words by Dasgupta et al. (2010) where strong priming effects have been observed for morphologically and phonologically related prime-target pairs; weak priming is observed for morphologically related but phonologically opaque pairs and no priming is observed for morphologically unrelated pairs. Apart from this, we do not know of any other cognitive experiments on morphological priming in Bangla or other Indian languages.

Several attempts have been made to provide computational models that can predict the processing of a given polymorphemic words. The obligatory decomposition model (Taft 2004) accounts for the fact that, decomposition of a polymorphemic words depend upon the frequency of the constituent stem (or the base word). Therefore, higher the stem frequency, easier it is to decompose. On the other hand, the full listing model (Burani and Laudanna 1992) states that the access to a polymorphemic word depends upon the frequency of the whole word. Thus, higher is the surface frequency of a word is, the easier it is to be recognized. The dual route access model (Baayen et al. 1997) argues that whether or not a polymorphemic word will be decomposed into its constituent morpheme, depends upon the surface frequency of that word; that is, if the frequency of a polymorphemic word crosses a threshold then the word will be accessed as a whole otherwise it will be accessed via its parts.

Experiments on English inflected words (Taft and Forster 1975), argued that lexical decision responses of polymorphemic words depends upon the base word frequency. In other words, if recognition of a polymorphemic word always takes place through decomposition, then higher the frequency of the stem is (called, base frequency), the shorter is the time to recognize the word (called, Reaction Time or RT). Previous experiments have shown such base frequency effects in most of the cases but not for all (Baayen et al. 1997; Bertram et al. 2000; Bradley 1980; Burani and Caramazza 1987; Burani et al. 1984; Colé et al. 1989; Schreuder and Baayen 1997; Taft and Forster 1975; Taft 2004).

Later, the dual processing race model (Baayen 2000) was proposed where both the full-listing and morphemic path compete among each other and depending upon the frequency of base and the surface word any one of the paths are chosen. The model proposes a specific morphologically complex form is accessed via its parts if the frequency of that word is above a certain threshold of frequency, then the direct route will win, and the word will be accessed as a whole. If it is below that same threshold of frequency, the parsing route will win, and the word will be accessed via its parts. However, what the dual processing model fails to explain is whether the stem frequency of a derived word is also involved during the recognition process.

The obligatory decomposition (morphemic) model has been proposed by Taft (2004) for inflectional suffixed English words and showed that stem frequency of a word plays an important role during decomposition of a derived word. Further, it argued that access to a polymorphemic word always takes place via two phases, a) decomposition and b) recombination. Therefore, during recognition, any polymorphemic word will be first decomposed

into its constituent morphemes where the morphemes will be individually recognized and then in the combination phase they will be recombined together to recognize the whole word.

The effect of morphological family size was observed by Schreuder and Baayen (1997). It has been shown that the response latencies of morphologically complex words in English significantly depend on the morphological family size of the word in concerned. Similar observations have been made in the works of Baayen et al. (2006), Pylkkänen et al. (2004), Jong et al. (2000), Carlisle and Katz (2006), Bertram et al. (2000), Schreuder and Baayen (1997). Closer to the present scope, by Prado et al. (2005) models the paradigmatic structure of a morphologically complex word. The work describes a distributed connectionist model of visual word recognition that explores how the paradigmatic effect can describe the lexical decision tasks of complex words. Milin et al. (2009) also studied the paradigmatic effect of morphologically complex words through information theoretic approach. Here, thereaction time of a complex word has been modeled based on the entropy of that word. Ford et al. (2010) in their work analyzed the role of stem and suffix family size. They observed both the stem as well as the suffix family sizes plays important role in the recognition of morphologically complex words.

In spite of the plethora of work that has been done to understand the representation and processing of polymorphemic words in the mental lexicon, a coherent picture is yet to be emerged. Further, most of the studies reported so far conducted experiments mainly in English, Hebrew, Italian, French, Dutch, and few other languages (Frost et al. 1997; Forster and Davis 1984; Grainger et al. 1991; Drews and Zwitserlood 1995; Taft and Forster 1975; Taft 2004). However, we do not know of any such investigations for Indian languages, which are considered to be morphologically richer than many of their Indo-European cousins. On the other hand, several cross-linguistic experiments indicate that mental representation and processing of polymorphemic words are language dependent (Taft 2004). Therefore, the findings from experiments in one language cannot be generalized to all languages. Hence, it is important to conduct similar experimentations in other languages. Bangla, in particular, supports stacking of inflectional suffixes, a rich derivational morphology inherited from Sanskrit and some borrowed from Persian and English, and an abundance of compounding, as well as mild agglutination.

Accordingly, the objective of this paper is to present computational models that can be used to understand the organization and processing of Bangla derivationally suffixed polymorphemic words in the mental lexicon. Our aim is to determine whether the mental lexicon processes Bangla morphologically complex words in terms of full-listing, morphemic or partial decomposition model. For this, we first conducted the masked priming experiment over a set of 500 Bangla morphologically complex words and collected reaction time data from 28 subjects. The experimental result shows that priming behavior is observed only for those cases where the prime is the derived form of the target and having a recognizable suffix (like, *sonAli–sonA* (GOLDEN–GOLD), and *bayaska–bayasa* (AGED–AGE). Weak priming is observed for cases where the prime is a derived form of the target but do not have a recognizable suffix (like, *sabhAba* (HABIT)–*sbAbhAbika* (NATURAL))or when the prime and the target are not morphologically related at all but have a recognizable suffix (like, *AmadAni* (IMPORT)–*Ama* (MANGO)). These observations initially indicate the obligatory decomposition model proposed by Taft and Forster (1975), Taft (2004) that assumes polymorphemic words to be processed via decomposition. Further analysis of the RTs obtained in the experiments indicate that processing of Bangla polymorphemic words may be achieved by the dual route decomposition model as proposed by Baayen (2000). However, contrary to the idea of considering base and/or surface frequency as sole predictor of processing Bangla polymorphemic words in the mental lexicon, we have explored the individual roles of

different features of a morphologically complex words like, the relative frequency between the base and surface word, type-token ratios, and role of suffixes (their family size, type-token ratio, and productivity) in morphological decomposability. Accordingly we have proposed different models. We have evaluated the proposed models with the results obtained from the priming experiment. Finally, we combine the role of all these characteristics and develop a more robust computational model that can predict the organization and processing of Bangla morphologically complex words. We have evaluated our proposed model with derivationally suffixed Bangla words and found that the performance of our proposed model outperforms the performance of the existing ones.

The rest of the paper is organized as follows: section "Psycholinguistic Study of Bangla Polymorphemic Words through Masked Priming Experiments" presents related works; section "Applying Frequency Models to Bangla Polymorphemic Words" describes the masked priming experiment performed over a set of Bangla morphologically complex words. Section "Model 3: Relative Frequency between Base and the Derived Words" describes different frequency based models and their performance in predicting the processing mechanisms of Bangla polymorphemic words; section "Exploring the Role of Suffixes in Processing of Bangla Words" describes the newer models of word recognition; sect. "Model-6: Combining Model-3, Model-4, and Model-5" concludes the paper by summarizing the observations and discusses the findings.

Psycholinguistic Study of Bangla Polymorphemic Words through Masked Priming Experiments

In order to study the effect of priming on morphologically derived words in Bangla, we execute the masked priming experiment as discussed in Forster and Davis (1984), Rastle et al. (2000), Marslen-Wilson et al. (2008) for Bangla derivationally suffixed words. In this technique the prime is placed between a forward pattern mask and the target stimulus, which acts as a backward mask. This is illustrated below.

Mask (500 ms) ########
Prime (72 ms) *sonAli* (GOLDEN)
Target (500 ms) *sonA* (GOLD)

The prime and the target words are either morphologically and/or semantically related or orthographically transparent to each other. A pair of word is said to be morphologically related if they meet the following conditions:

a) One word is the derived form of the other
b) The derived form has a recognizable suffix

For example, the word pairs *bADiOyAlA* (House keeper) and *bADi* (House) are morphologically related since, *bADiOyAlA* is derived from *bADi* and has a recognizable suffix *-OyAlA*. A pair of word is said to be orthographically transparent if whole or a significant part of one word is fully or partly contained in the other word. Orthographically transparent words may or may not be morphologically or semantically related to each other. For example, *maShA* (mosquito) and *maShAla* (flame) are orthographically transparent but morphologically not whereas, our previous example, *bADIOyAlA* (House Master) and *bADi* (House) are both orthographically transparent and morphologically related.

After presenting the target probe, the subjects were asked to make a lexical decision whether the given target is a valid word in that language. The same target word is again

probed after a random amount of time, but with a different visual probe called the control word. The control words do not have any morphological, orthographic or semantic relatedness with the target. For example, *baYaska* (aged) and *baYasa* (age) is a prime-target pair, for which the corresponding control-target pair could be *naYana* (eye) and *baYasa* (age).[1]

The time taken by a subject to complete the lexical decision task after the visual presentation of the target is defined as the response time (RT). The RTs between a prime-target and the corresponding control-target pair are compared to identify whether there is enough evidence of morphologically structured lexical representation. Experiments in English and other languages show that in general the RT between the prime-target pair is significantly less than that of the control-target pair, implying the presence of morphological priming effect. Nevertheless, all linguistically apparent morphological processes need not have equal priming effects or any effect at all.

Materials and Methods

We selected 500 prime-target pairs, where the primes are related to the targets either in terms of morphology, semantics and/or orthography. In order to factor out the effects of semantics or orthography, we adopted the same technique discussed in Rastle et al. (2000), Marslen-Wilson et al. (2008) and classified the words into five different classes each consists of 100 word pairs. Words in these classes are classified according to their morphological, semantic and orthographical relationship. For example, class-I words or [M+S+O+] consists of word pairs that are morphologically (M+), semantically (S+) as well as orthographically (O+) related. Here, the "+" (as in M+) indicates relatedness and "−" indicates unrelatedness. Similarly, words that are morphologically unrelated but orthographically related will be represented as [M−S−O+] and so on. We also introduces a special class of words [M'+S−O+] which are similar to the word class [M−S−O+], however, this words consists of a valid and transparent Bangla suffixes. For example, words like, *AmadAni*(import) consists of a valid Bangla suffix (-dAni) and a valid stem *Ama*(Mango). However, *AmadAni* and *Ama* does not have any morphological connection among them. These classes of words have been introduced to observe the priming phenomena for pseudo suffixed words. Table 2 describes these five classes with examples.

It is interesting to note that while it is very easy to collect word pairs belonging to class I, it is hard to come up with morphologically derived word forms in Bangla which are orthographically unrelated. In fact, almost all the native Bangla suffixes (e.g., *-A, -I, -li, -oYA*) do not change the form of the root to which it attaches. However, there are some derivational processes inherited from Sanskrit, where the root forms are phonologically distinct from the derived ones, e.g., *hatyA* (to kill)–*hi.nsA* (violence, i.e., desire to kill).

For each of the 500 target words, we selected another set of 500 control words. These control words are similar to the prime words in terms of word length, and number of syllables. However, they are neither morphologically related nor orthographically transparent to the targets. Some statistics about the prime, target, and control words are presented in Table 1.

As discussed earlier, after hearing the auditory prime, a visual probe is presented to the subjects based on which some lexical decision have to be made. Thus, it is essential to restrict the subjects to make any strategic guess about the relationship between prime and the target word pairs. This can be achieved by introducing some filler in between the actual prime-target or control-target pairs. We constructed a set of 500 filler pairs which can be categorized into

[1] This study follows the experiment 1 of Rastle et al. (2000); however, for the sake of readability we briefly describe the design process and other details.

Table 1 Statistics of the target, prime and control words

Word type	Avg. word length	Avg. no. of complex characters	Avg. corpus frequency
Target	4.0 (2.0)	0.260 (0.11)	32.63 (10.10)
Prime	6.4 (1.9)	0.464 (0.29)	25.82 (7.04)
Control	6.2 (1.2)	0.472 (0.12)	25.14 (8.33)

Number in parenthesis signifies standard deviations

the following five sets of 100 word pairs each: (a) where the prime is a valid word but the target is not, although it is orthographically contained in the prime and is obtained by deleting some word final character-string, e.g., *kapAla* (fore-head)–*kapA* (non-word); (b) where the target is a valid word but the prime is not, although it orthographically contains the target and is derived by adding a suffix to the target, e.g., *hAtAri* (non-word)–*hAta* (hand); where the prime is valid but the target is not, and is obtained by swapping the individual alphabets of the target, e.g., *pAgalAmo* (madness)–*pAlaga* (non-word); and (c) where both the prime and target are valid words without any morphological and phonological relatedness.

Thus, all together, there are 1,500 word pairs including 500 prime-target pairs, 500 control-target pairs and 500 fillers. Before presenting the word pairs to each subject, they are randomized and divided into two set, such that the prime-target pair and the corresponding control-target pair are present in different sets. Moreover, each set contains exactly half of the prime-target and half of the control-target pairs.

Procedure

The experiment was conducted using the DMDX software tool.[2] Corresponding to each visual probe, subjects had 3,000 ms to perform the lexical decision after which the system presents the next masked prime followed by a visual stimulus. The subject performs the decision task by pressing either the "K" button (for valid word) or the "S" button (for invalid word) of a standard QWERTY keyboard. The system automatically records the reaction time (RT), which in this case is the time between the onset of the visual probe and pressing of one of the keys by the subject.

Before starting the real experiment all the subjects were given a short training about the task. A trial run was also performed using the separately collected 20 trial word pairs. As discussed earlier, the experiment is divided into five different phases. The experimental procedure for both the phases is same except that the prime and control words are different. The duration of each phase is about 25 min. Since a continuous session of 25 min require a lot of attention and is tiring for the subjects, we further divided each phase of the experiment into five small sessions of five minutes each. There was a break for ten minutes between the sessions.

Participants

The experiments were conducted on 32 highly educated native Bangla speakers; 27 of them have a graduate degree and 5 hold a post graduate degree. The age of the subjects varies between 22 and 35 years (Table 2).

[2] http://www.u.arizona.edu/~kforster/dmdx/dmdx.htm.

Table 2 Dataset for the experiment

Class	Explanation	Examples
M+S+O+	Morphologically derived, stem and suffix are transparent and decomposable, semantically and orthographically related	nibAsa (residence)-nibAsi (resident)
M+S+O−	Morphologically derived, stem and suffix are opaque, semantically related but orthographically not	mitra (friend)-maitri (friendship)
M'+S−O+	Morphologically unrelated, transparent stem and suffix, semantically unrelated but orthographically related	Ama (Mango)-AmadAni (import)
M−S+O−	Semantically related but Morphologically and Orthographically unrelated	jantu (Animal)-bAgha (Tiger)
M−S−O+	Morphologically and semantically unrelated but orthographically related	ghaDi (watch)-ghaDiYAla (crocodile)

Results

The RTs with extreme values and those for incorrect lexical decisions (about 1.8 %) were excluded from the data.[3] We have also discarded one prime-target pair from our dataset due to its incorrect spelling. Further, four subjects have to be excluded from the experiment due to their inconsistent and extremely high error rates. Overall, we have analyzed the RTs of 490 prime-target and 490 control-target pairs for a total of 28 subjects. Table 3 summarizes the average RTs for the prime and control sets for the five classes. The RT and error rate data were submitted to by-subject and by-item analyses of variance with the following main factors: priming relation (prime vs. control) and relation classes (M+S+O+, M+S+O−, M'+S−O+, M-S+O−, and M−S−O+).

We observed that, overall, the average RTs for Bangla control-target pairs are more than the corresponding prime-target ones. In other words, priming relation had a significant effect over the control relations. We have computed the by subject and by item F-scores as F_1 (1, 23) = 32.42, $p < .002$; F_2 (1, 485) = 48.93, $p < .005$. "Correct" responses to targets were faster when they appeared after the primes than unrelated controls. The priming effects of individual classes along with their significance values are depicted in Table 3. To summarize, strong priming effects are observed when the target word is morphologically derived and has a recognizable suffix, semantically and orthographically related with respect to the prime[M+S+O+] ($F_1(1, 23) = 18.21$, p = 0.001, $F_2(1, 96) = 21.13$, $p < 0.03$); although statistically significant, but week priming is observed for word pairs belonging to [M+S+O−], and [M'+S−O+]; no priming effects are observed when the prime and target words are orthographically related but share no morphological or semantic relationship [M−S−O+] ($F_1(1, 23) = 17.34$, $p = 0.006$, $F_2(1, 96) = 13.47$, $p < 0.004$) or only semantically related but without any morphological or orthographic relation[M−S+O−]. These results thus rules out the possibility that priming in [M+S+O+] could be due to individual effects of orthographical or semantic relatedness.

[3] Any RT value that falls outside the range of Average RT 500 ms is considered as extreme.

Table 3 Average RT for the word classes, the F-Score and p values

Class	Avg. RT (in ms) and error rates (in %)					ANOVA	
	Prime	Error	Control	Error	Diff	F-score	p value
[M+S+O+]	523	2.40	589	1.20	66	$F_1(1, 23) = 18.21$	$p < 0.001$
						$F_2(1, 96) = 21.13$	$p < 0.030$
[M+S+O−]	653	2.00	660	1.60	7	$F_1(1, 23) = 10.04$	$p = 0.07$
						$F_2(1, 94) = 13.13$	$p < 0.06$
[M'+S−O+]	554	2.49	542	1.86	12	$F_1(1, 23) = 12.42$	$p < 0.009$
						$F_2(1, 94) = 11.93$	$p < 0.040$
[M−S+O−]	606	3.12	597	2.11	−19	$F_1(1, 23) = 17.56$	$p < 0.02$
						$F_2(1, 96) = 18.39$	$p < 0.005$
[M−S−O+]	690	3.69	657	3.64	−43	$F_1(1, 23) = 19.67$	$p = 0.001$
						$F_2(1, 95) = 15.53$	$p < 0.008$

Analysis of RTs for Lexical Items

It is interesting to look at the individual lexical items whose priming behavior deviates from that of their class. For instance, *akarma* (useless work)–*akarmaNyA* (worthless girl), *pAkA* (smart)–*pAkAmo* (street smartness), *srama* (labour)–*sramika* (worker) and *kShamA* (forgiveness)–*kShamaNIYa* (forgivable) exhibit the least priming effect in [M+S+O+].

In [M+S+O−] class, prime-target pairs like, *pAna* (to drink)–*pipAsA* (thirsty), *dharA* (hold)–*dhairya* (patience) and *chalA* (move)–*chAlita* (controlled) show no priming effect despite there is a strong morphological association between the prime-target pairs. In general, we observe that participants are unable to recognize the morphological connection between most of the derivationally suffixed word pairs in the [M+S+O−] class. Examples include *suhRRida* (friend)–*souhArdya* (friendship), *uchit* (appropriate)–*auchitya* (appropriateness) and *hatyA* (murder)–*hi.nsA* (violence). One explanation for this is, Bangla inherits these morphological forms from Sanskrit and the derivational process is unknown.

Another important observation from the experiment is that a significant number (around 38 %) of prime-target pairs belonging to the [M+S+O+] class shows weak or no priming despite their high morphological association. For example, pairs like, *ghana* (dense)–*ghanatba* (density) and $ga \sim Nga$ (river Ganges) – $ga \sim Ngajala$ (water from Ganges), *jiba* (living being)–*jibanta* (alive), *chora* (thief)–*chorAI* (smuggled) etc. shows very weak priming effect. In order to eliminate possible experimental errors, we repeated the same priming experiment with these words to the same set of subjects and obtained the same result for all the pairs (although the average RT for some of the target deviates from that of the original result but this did not change the overall results).

Analysis of High RT Lexical Items

We also observed that the RTs for certain pair of words were significantly higher than what one would expect and consistently so across all the participants. Manual inspection of these words indicates that the target or the corresponding prime/control words in such cases have one of the following properties:

Table 4 List of Bangla words having conjugate characters and their average RT across 28 subjects

Word	Corpus Frequency	Length	Avg. RT
jIbanta(alive)	26	6	641 (67)
bayaska(old)	34	6	773 (73)
hindustAna(India)	1	11	754 (98)
ghanatba(density)	6	5	774 (76)
(sUryAsta)(sunset)	20	9	846 (53)
kerAnigiri(clarkship)	8	10	1,078 (102)
lambAi(length)	1	6	1,132 (111)
rAShTrIYa(national)	113	9	1,227 (94)

Number inside parenthesis signifies standard deviations

a) Very infrequent
b) Long in terms of the number of characters present (>7)
c) Presence of certain conjugates such as (*Sh*+*T*), ($l + p$) and ($\sim N + g$), and other irregular or non-transparent glyphs (g + u) and (h+RRi) in the target
d) Incorrect spelling of the target (e.g., *sharira* instead of *sharIra*)

Frequency effects on recognition time are well studied (Forster and Davis 1984; Taft 2004) and explain observation (a). It is quite well known that visual word recognition time and accuracy depends on several factors such as, font size, font type, eccentricity, i.e., the angle of the visually represented word from the focus of the eye, and the crowding effect, i.e., the physical length of a word [see, e.g., Jo (2000)]. Therefore, observation (b) is also not surprising. However, the last two observations are specific to Bangla orthography and throw up some interesting research questions.

The Bangla script uses a large number of non-transparent glyphs for conjugates and also some consonant-vowel pairs. These glyphs have been a point of discussion amongst the scholars of Bangla language, especially for pedagogical reasons: non-transparency in character representation leads to poor recognition and recall of the glyphs as well as the words containing them; this negatively affects the learning process in young children. Therefore, there have been proposals for using the less common but easy to recognize transparent forms of these glyphs. We do not know of any systematic study that explores and quantifies the cognitive load associated with the learning and processing of the glyphs with varying degree of transparency. Since such a study is beyond the scope of the current work, the experimental items were not prepared to specifically identify glyph recognition complexities. Nevertheless, we do observe an effect of glyph transparency and glyph usage frequency on word recognition time. Uncommon and non-transparent glyphs (e.g., (Sh+T)) have highest recognition time, whereas very frequent glyphs (e.g., (k+Sh)), even if non-transparent do not seem to have a negative effect on the recognition time of the words. Table 4 depicts a list of Bangla words containing different conjugate characters and their average RT over the 28 subjects.

High recognition time and error for incorrect spellings, or non-words, is a well-known fact. However, it is interesting in the context of Bangla because Bangla does not distinguish between short and long vowels in pronunciation, even though the distinctions are traditionally maintained in the written forms. Recently, there have been several controversial proposals for spelling reforms where all long vowels are to be replaced by their shorter counterparts. The unintentional error in our dataset, *sharira* (body) instead of the more commonly found and popularly acceptable form *sharIra*, was accidentally discovered when we observed very high RT for the pairs involving this item as the target. Thus, it might be argued that speakers who have learnt the traditional spellings will find it hard to recognize their new spellings.

Table 5 Comparison of RTs between Bangla words with their conventional and un-conventional spelling forms. Number inside parenthesis signifies standard deviations

Bangla Words with their conventional forms	Average RTs		Unconventional Representations	Average RTs	
	RT	STDEV		RT	STDEV
অক্ষি(akShi)	957	107	অক্ষী (akShI)	991	87
অক্ষুণ্ণ (akShuNNa)	1386	31	অক্ষুণ্ন (akShuNNa)	1432	79
পরীক্ষা (parIkShA)	1341	89	পরিক্ষা (parikShA)	1373	111
চিকিত্সা (chikitasA)	933	51	চীকীত্সা (chIkItasA)	954	40
অজীর্ণ (ajIrNa)	1403	116	অজীর্ন (ajIrna)	1457	81
রোগিণী (rogiNI)	775	120	রোগিনী (roginI)	874	115
অজ্ঞানী (aj~nAnI)	1179	111	অজ্ঞানি (aj~nAni)	1246	53
অজ্ঞানবাদী (aj~nAbAdI)	1224	63	অজ্ঞানবাদি (aj~nAbAdi)	1287	100
অঞ্জলি (a~njali)	1027	75	অঞ্জলী (a~njalI)	1117	93
মনসিজ(manasija)	1221	112	মনসীজ (manasIja)	1232	47

In order to extend this argument we have conducted a separate lexical decision experiment. Here, we chose 80 Bangla words that have different accepted spelling conventions. The words were shown to 21 subjects using the procedure as discussed in Baayen et al. (1997), Taft (2004). We asked subjects to recognize whether a given word is valid or not. Similar to the priming experiment discussed above, we have recorded the reaction time of individual words per subject. An illustration of some typical Bangla words and their average RT is depicted in Table 5. We found that, for most of the cases the RT of those words that exhibits a more common form of representation is significantly lower than the words having an uncommon representation $F_1(1, 20) = 11.4$, $p < 0.05$; $F_2(1, 80) = 23.11$, $p < 0.02$. This is not a surprising conclusion, though the exact nature and extent of difficulty in perceiving the new forms is a topic of further research.

Analysis of Error Rates

During priming experiments, participants can make an incorrect lexical decision on whether a word is valid or invalid. The errors could be due a participant's incorrect judgment about validity of a word or a wrong selection made despite of a correct judgment. In general, it has been observed that error rates and RT for non-words are higher than valid words. Table 6 reports the error rates and RT for the prime-target, control-target and the fillers. As expected, we observe high error rates and high RT for fillers, which mostly consist of non-words as target or prime. In fact 81 % of the total errors for the fillers are for the non-words. The overall error rate, however, is quite low.

Recall that test of significance for individual subjects revealed 28 out of 32 participants showed statistically significant priming effects ($p < 0.03$), which led us to hypothesize that the remaining four participants were not paying good attention during the experiments or are not well exposed to Bangla due to their educational medium.

Table 6 Comparison of the RT and error rates between prime, control and fillers

Class	Average RT (m.sec)	Error (%)
Prime	579	1.2
Control	654	1.9
Fillers	1,011	6.2

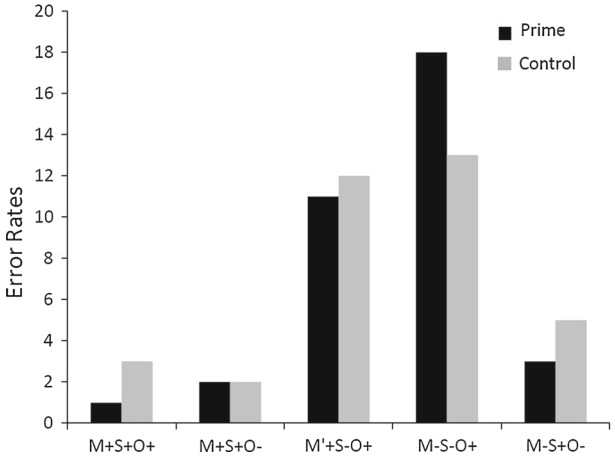
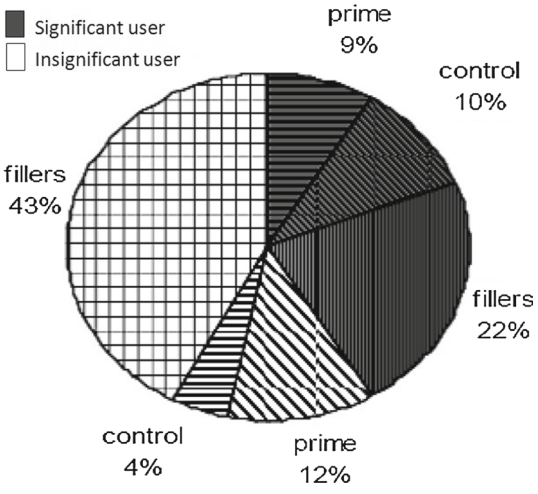


Fig. 1 Comparison of error rates across word classes

Fig. 2 Comparison of error rates for different categories of lexical items. The *gray* and the *white cells* are respectively for participants who displayed significant and insignificant priming effects



Therefore, we would expect their error rates to be higher than that of the other 28 participants. Figure 1 plots the histogram of error rates for the significantly primed (left bars) and non-significantly primed participants. Overall error rate of the former class of participants (41 %) is much less than that of the latter (59 %), which matches our speculation. Again, as one would expect, the maximum errors are made for fillers. Among the valid words, the highest error rates are observed for the class [M−S+O−] and [M−S−O+] (see Fig. 2). Recall that these are the classes for which we do not observe any priming effect.

Discussion

As explained earlier, the effect of priming with a morphologically derived word vindicates decomposition, leading to reduced RT of the target. However, it is apparent from the above results that all polymorphemic words do not decompose during processing. This contradicts the obligatory decomposition model of Taft and Forster (1975), Taft (2004). Naturally, the question that arises is what are the other factors that are responsible for the decomposition of Bangla polymorphemic words? In order to answer this we need to further investigate the processing phenomena of Bangla derived words. One notable means is to identify whether the stem or suffix frequency of a polymorphemic word is involved in the processing stage of that word. For this, we apply the existing frequency based models to the Bangla polymorphemic words and try to evaluate their performance by comparing their predicted results with the result obtained through the priming experiment.

Applying Frequency Models to Bangla Polymorphemic Words

Model-1: Base Word Frequency Effects

The base word frequency model states that the probability of decomposition of a Bangla polymorphemic word depends upon the frequency of its constituent stem. Thus, a polymorphemic word that constitutes a high frequency stem will be decomposed faster than a word having low stem frequency. In order to compare the results with respect to that of the masked priming experiment discussed in the previous section, we made a slight change to the original model. We propose that if the stem frequency of a polymorphemic word crosses a given threshold value τ, then the word will decomposed into its constituent morpheme. The model is formally represented as:

$$Decomposability\,(w) = \begin{cases} TRUE, & if\ \log_{10}\left(frequency\left(W_{stem}\right)\right) \geq \tau \\ FALSE, & if\ \log_{10}(frequency(W_{stem})) \leq \tau \end{cases}$$

The value of τ is computed as the log of average base word frequency of Bangla words from a corpus[4]. This returns the value of τ as 0.09. We apply model-1 to a set of 500 morphologically derived words. According to model-1, words like, *pathika* (318),[5] *jalA* (15), *bADiwAlA* (19), and *baYaska* (34) will be decomposed into their constituent stem and suffixes during the processing stage. The reason behind this is that, all these words are derived from very high frequency stems like, *patha* (2241), *jala* (1736), and *bADi* (1118). Thus, priming phenomena will be observed if these stems (considered as targets) are preceded by the derived words (i.e the primes). Since, prior exposure of the prime will result in decomposition of the derived prime word into its morphemes and thus the recognition of the target will start well before the actual target is probed. Similarly, according to model-1, derived words like *ginnipanA*, *rAjakIYa*, and *nibAsi* will not be primed and thus not be decomposed during the processing stage of the Bangla polymorphemic words. The predicted values of the model are evaluated with respect to the results obtained from the priming experiment discussed in section. The performance of the model is computed in terms of Precision, Recall, F-Measure and Accuracy. The confusion matrix along with the computed results is depicted in Table 7. We observed

[4] Corpus frequency is computed by combining the CIIL, and Anandabazar corpus and literary works of Rabindranath Tagore, and Bankim Chandra available from (www.ciil.org, iitkgp.ernet.in and nltr.org).

[5] Number in the parenthesis represents the frequency of a word in the corpus.

Table 7 Summarizing the results of base word frequency model

	Model-1: Base word frequency (BF) (values out of 500 words)	Performance	
False positive	135	Precision (%)	60
True negative	111	Recall (%)	78
True positive	199	F-Measure (%)	68
False negative	56	Accuracy (%)	62

that the model possess an accuracy of 62 %. However, from the Table 4 we observe the false positive and false negative values to be around 26 and 11 % respectively. This indicates for these 26 % of the words, the base word frequency model predicts no morphological decomposition due to extremely low base word frequency (ranges between 1 and 7 out of 4 million) but the priming experiment shows high degree of morphological decomposition. Similarly, for the On the other hand, for about 11 % of the word the model fails to explain why around 26 % (like, *ekShatama*, *juYADi* and *rAjakiYa*) words having extremely low base word frequency (ranges between 1 and 7)shows high degree of priming. Moreover, the model also fails to explain the negative decomposability of 11 % words (like, *laThiYAla*, *dAktArakhAnA*, and *Alokita*) despite having high root word frequencies (ranges between 100 and 1,100). Hence, in the next section we proceed to experiment with the derived word frequency model to get a better model that can be used to explain the above exceptions.

Model-2: Derived Word Frequency Effect

In this model we try to validate the priming phenomena with respect to the whole word frequency. The hypothesis is that, if a specific morphologically complex form is above a certain threshold of frequency, then the whole word access will be preferred instead of decomposition model, and thus no priming effect will be visible in this case. On the other hand if the derived word frequency is below that same threshold of frequency, the parsing route will be preferred, and the word will be accessed via its parts. The derived word frequency model can be formally represented as:

$$Decomposability\,(w) = \begin{cases} TRUE, & if\ \log_{10}(frequency\,(w)) \leq \tau \\ FALSE, & if\ \log_{10}(frequency(w)) \geq \tau \end{cases}$$

In order to apply this model to Bangla polymorphemic words, we have computed the threshold value to be the average corpus frequency of words which comes out to be 1.33. Therefore, a Bangla morphologically complex word whose surface frequency exceeds the threshold limit of τ will be accessed as a whole otherwise; it will be decomposed into its parts. For example, words like *sonAli*(179), *galAbAji* (334), and *suryAsta* (407) must be processed as a whole and words like, *ginnipanA*, *juYA.Di*, and *ekaShatama* will be parsed into their constituent morphemes namely *ginni*, *juYA*, and *ekaSha*. Similar to the approach discussed in model-1, the same 500 polymorphemic words were given as an input to the model. The predicted values of the model are then compared with the actual data collected from the priming experiment (see Table 8 for the confusion matrix along with the computed results). From the results depicted at Table 8, we observe that the model can be used to explain the possible decomposition of low frequency derived words (like, *juYA.Di*, *nishThAbAna*, and *ekaShatama*) which model-1 fails to explain. Thus, the false positive value for the present model is lower than that of model-1 (21 %). However, model-2 performs poorly due to the

Table 8 Summarizing the results of surface word frequency model

	Model-2: surface word frequency model (SF) (values out of 500 words)	Performance	
False positive	111	Precision (%)	58
True negative	88	Recall (%)	51
True positive	155	F-Measure (%)	54
False negative	143	Accuracy (%)	49

high false negative value (28 %). This implies the model fails to recognize the potentially decomposable words (like, *meghalA*, *pAkAmo* and *AkAShamandala*) properly.

Discussion

From the above results we observe that, Model-1 predicts that the priming/decomposition will take place if the base word frequency is high, irrespective of the frequency of the prime. However, the prediction of the model was not validated when the prime as well as the target words are both having high frequency. On the other hand, Model-2 predicts that priming/decomposition will take place if the prime is of low frequency. However, the model was not validated from the experimental results for low frequency prime and low frequent target pairs. Hence, the two extremes of paring call for a newer model.

Model 3: Relative Frequency between Base and the Derived Words

In a pursuit towards an extended model, we combine the model 1 and 2 together to observe if and how their combination can predict the parsing phenomena. One way to combine the base and derived word frequency is through regression analysis. In accordance to the technique discussed in Hay and Baayen (2001), we took the log of frequency of both the base and the derived words and plotted their values in a log-log scale. In order to get the best-fit curve over the given dataset we have used the least square fit regression method, the equation of the straight line being:

$$\log_{10}(BaseFrequency) = 0.346 \times \log_{10}(SurfaceFrequency) + 1.611$$

We propose that any point that falls above the regression line will be parsed into its constituent morphemes during processing. On the other hand, points situated below the regression line will be accessed as a whole. In other words, given the surface frequency of a derived word W, the equation above can predict the frequency of the corresponding base word. If the predicted frequency of the base word is greater than the actual frequency of the Base word then the point lies above the regression line and thus, during processing these words will be accessed via the decomposition model. This is depicted in Fig. 3 which illustrates the surface and base word frequency distribution of 2,000 Bangla polymorphemic words. The model predicts that those points that lie on or above the regression line will be parsed during processing whereas points lying below the regression line will be accessed as a whole. The results are depicted in Table 9. We observe that the model performs much better (with false negative and false positive values below 17 %) than the previous two models.

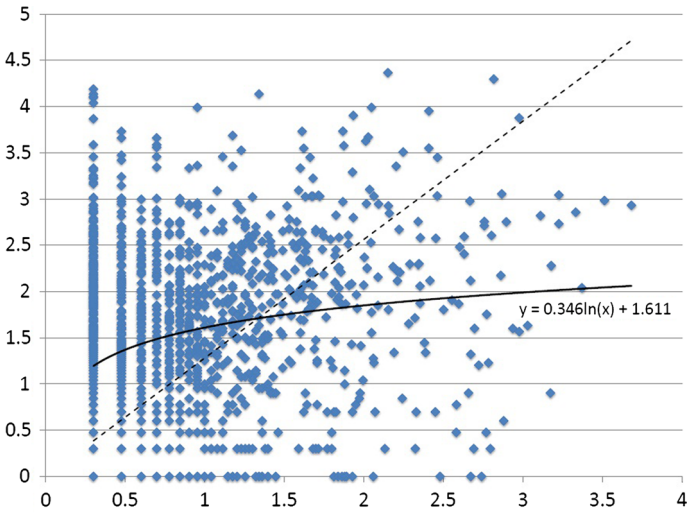


Fig. 3 The relation between log derived frequency and log base frequency for 2,000 different Bangla polymorphemic words. *Solid line* represents *least squares fit regression line*

Table 9 Summarizing the results of relative frequency model

	Model-3: base and surface word frequency ratio (values out of 500 words)	Performance	
False positive	88	Precision (%)	70
True negative	143	Recall (%)	75
True positive	199	F-Measure (%)	72
False negative	67	Accuracy (%)	69

We validate our model by comparing its predicted results with the results obtained from the masked priming experiment on 500 Bangla polymorphemic words. The results of the predicted values of the model along with accuracy are depicted in Table 9. The present model shows an accuracy of 69 %. Consequently a significantly high number of words (31 %) are wrongly classified by the present model. This may be accounted for by the fact that most of the derived words that could not be correctly classified by the present model are composed of low frequency stem and suffixes. This led us to further modify the existing model to study the role of individual suffixes during the morphological decomposition of Bangla polymorphemic words.

Exploring the Role of Suffixes in Processing of Bangla Words

One of the key issues that have not been addressed in Model-3 is the fact that whether the regression analysis between base frequencies on derived frequency across suffixes will generate any variation in the slope and intercept of the resulting line. It has been observed that, for English, regressing between base and derived word frequency generates different

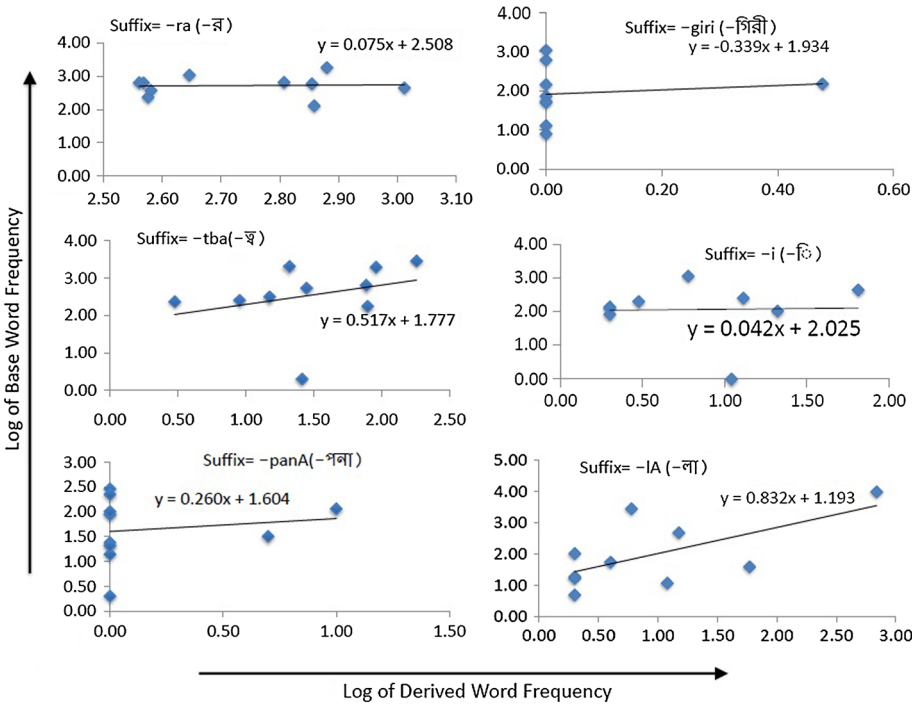


Fig. 4 The relation between log derived frequency and log base frequency for four affixes. The *lines represent least squares regression lines*

slope and intercept values. Hay and Baayen (2001) showed that suffixes belonging to high intercept values shows higher tendency to decompose than suffixes with low intercept values.

In this section, we would like to examine the same for Bangla. Therefore, we will try to examine whether the regression analysis between base and derived frequency of Bangla words varies between suffixes and how these variations affect word decomposition. For this, we choose six different Bangla native suffixes with varying degree of token frequencies. For each suffix, we choose 10 different derived words. Figure 4 illustrates the chosen suffixes corresponding to different suffix classes and their base word and derived word frequencies. Finally, we plot the regression line between words under each suffix and found that the intercept of the regression line for Bangla suffix shows considerable variation.[6] Figure 4 illustrates the regression analysis of the six different Bangla suffixes. We observe that those suffixes having high value of intercept are forming derived words whose base frequencies are substantially high as compared to their derived forms. Moreover, we also observe that high intercept value for a given suffix indicates higher inclination towards decomposition rather than whole word access.

From the above analysis we observe that decomposition of a Bangla polymorphemic word not only depends upon the base and derived word frequencies, but also depends upon the characteristics of the given suffix. That is, whether or not a polymorphemic word will be accessed via decomposition or by whole word access depends on several factors like the frequency distribution between the base word and the derived word, type and token frequency

[6] Similar results were reported for the English suffixes in Hay and Baayen (2001).

of the suffix, and the degree of affixation between the base word or the stem and the suffix. Thus, in spite of having both derived and stem frequency ratio and suffix type/token ratio falls below the threshold frequency τ, a Bangla polymorphemic word may not show the decomposition phenomena due to the fact the degree of affixation between the stem and the suffix may be weak. Therefore, in the following sections we will explore the degree of affixation between the stem and the affix. Accordingly, we will first identify the role of suffix frequencies (type and token) in determining the decomposition of Bangla polymorphemic words.

Model-4: Suffix Type/Token Ratio Model

The type frequency is defined as the total number of distinct words associated with an affix. On the other hand, token frequency of a suffix is the total number of times a suffix is attached with a word. In this model the type token frequency ratio of individual suffixes was taken into account to study the decomposition of Bangla polymorphemic words. As suggested earlier, lesser is the token frequency of a suffix greater is its chances in getting parsed in a word attached with it. Type frequency of a suffix exhibits the potentiality of a suffix in forming an entirely new word. In other words, it is a count of how many different types of words a suffix can derive from the base word. Taking the ratio between the type and the token frequency of every suffix that can attach with a given stem, we determine the degree of affixation of a given stem and a suffix. Through this information we try to predict the access mechanisms of Bangla polymorphemic words. We believe as the degree of affixation between a stem and a suffix decreases the higher is the probability of decomposition of the target derived word. Therefore, hypothesis for this model can be given as, for a given Bangla polymorphemic word if the type/token frequency ratio (in logarithmic scale) of a given suffix, attached to a word, exceeds a predefined threshold τ, then the word will be accessed as a whole otherwise the derived word will be decomposed into the corresponding stem and suffix. The threshold value for the surface and stem frequency ratio is computed by taking the average of the ratio between surface word and base word frequency of around 2,000 polymorphemic words. We estimated the average and hence the threshold to be around 0.08. Therefore, the proposed model can be represented as:

$$Decomposability\,(w) = \begin{cases} TRUE, & if\ \frac{frequency(Type(W_{suffix}))}{frequency(Token(W_{suffix}))} \leq \tau \\ FALSE, & otherwise \end{cases}$$

Similar to the previous models, our new model is evaluated over a set of 500 Bangla polymorphemic words where the stem and the suffixes are transparent (i.e the suffix is fully or partly recognizable). The performance of the model as presented in Table 11 shows 69 % accuracy.

Although, model-4 does not throw any improvement over model-3 in terms of accuracy, we observed that model-4 performs best in determining the true negative values (see Table 11) and thus, can better predict those words which does not shows the decomposition phenomena. On the other hand, model-3 possesses a high precision of 70 % and can better detect the true positive values (199) as compared to model-4. Therefore, despite of having same accuracy, both the model shows equal strength in classifying different types of word. This observation is further illustrated in Table 10 which depicts a list of words that were given as an input to model-3.

From Table 10, we observe that words like, *meghlA* (CLOUDY), *nibAsI* (RESIDENT) and *Alokita* (SHINE) despite of having very week priming effects, are wrongly classified as

Table 10 List of sample prime target pairs given as an input to model-3 and model-4 and their performance

Prime-targets	Base/surface frequency ratio	Priming type	Model-3 result	Model-4 result
jIbanta- jIba (lively–living)	0.47	0	1	0
bA.DioYAlA- bA.Di (Housekeeper–House)	0.01	1	1	1
bayaska- bayasa (Old–Age)	0.05	1	1	1
nibAsI- nibAsa (Residence–Resident)	0.04	0	0	1
meghalA- megha (Cloudy–Cloud)	0.02	0	0	1
Alokita- Alo (Lightning–Light)	0.02	0	0	1
rAShTrIYa- rAShTra (National–Nation)	2.05	1	0	0
nAchunI- nAcha (Dancer–Dance)	0.05	0	0	0

Priming type = 1 implies significant degree of priming is observed for the word pairs, and priming type = 0 implies no priming or less priming is observed. For Model-3 and Model-4 Result, 1 implies the model correctly classifies the decomposition of the derived word and 0 implies failure to classify the word correctly

Table 11 Summarizing the results of Type/Token ratio model

	Model-4: type/token ratio (values out of 500 words)	Performance	
False positive	100	Precision (%)	50
True negative	158	Recall (%)	85
True positive	100	F-Measure (%)	63
False negative	15	Accuracy (%)	69

decomposable words because of their low base and surface words frequency ratios (0.04, 0.02, and 0.02 respectively). On the other hand, when these words are provided as an input to model-4, they have been correctly classified as non-decomposable. This may be accounted due to the fact that suffixes attached to these words have got low type/token ratios (0.01, 0.03, and 0.018 respectively) and thus difficult to decompose. However, both the proposed models fails to explain the decomposition of word like, *rAShtriYa* (NATIONAL) and non-decomposition of word like *nAchuni* (DANCER) which needed a more deeper analysis. Nevertheless, the above experimental data and our observation further strengthen our claim that only base and surface word frequencies are not the only factors responsible for the decomposition factor and suffix properties plays equally important role in determining the decomposition of Bangla polymorphemic words in the mental lexicon. Hence, we argue that combining the above two models can better predict the decomposability of Bangla polymorphemic words. But, before that we would further like to analyze whether along with the type/token ratio, the productivity of a suffix plays any role in morphological decomposition (Table 11).

Model-5: Suffix Productivity in Morphological Decomposition

In this section our objective is to identify the degree of affixation of a given suffix and a word. In other word, we try to compute how well a given suffix can be attached with a given stem. This is done by computing the productivity of a suffix. Although it has been proposed that suffix type frequency can be a determiner of its productivity, yet it has been argued that productivity is multifaceted and can be assessed in different ways (Hay and Plag 2004). We, in this paper apply the same technique as proposed by Hay and Plag (2004) to compute the productivity of Bangla suffixes. There are mainly three components of productivity, P, P*,

Table 12 Correlation between the suffix type frequency, token frequency, happex count and conditioned degree of productivity

	Type frequency	Type frequency	Type frequency	Type frequency
Type frequency	–	0.97	0.91	−0.726
Token frequency	–	–	0.909	−0.694
Happex	–	–	–	−0.701
Productivity	–	–	–	–

and V. V is the "type frequency" of a suffix. That is, the number of different type of words with which the suffix is attached. P is the "conditioned degree of productivity" and is the probability that we are encountering a word with a suffix(S) and it is representing a new type. The productivity of a suffix S (denoted as P(S)) is therefore computed as:

$$Productivity\,(S_i) = P\,(W|S_i \cap frequency\,(w) = 1) = \frac{H_{count}(S)}{N_S}$$
$$= \frac{Number\ of\ happex\ with\ that\ suffix}{Number\ of\ token\ containing\ the\ suffix(N)}$$

Where, H_{count} is the number of hapaxes with the given affix S and N_S is the number of tokens containing the suffix (N). Hapaxes are those words which occur exactly once in the corpus. Hapaxes and their counts are important in linguistics because they reveal how potential a suffix is in forming an entirely new word, what is its strength in producing new and rare words.

P* is the "hapaxed-conditioned degree of productivity". It expresses the probability that when an entirely new word is encountered it will contain the suffix. It is measured by calculating all hapaxes in the corpus with that affix / total number of hapaxes in the corpus. Thus, P* is computed as:

$$P^* = P\,(Happex|S_i) = \frac{Number\ of\ happex\ in\ the\ corpus\ with\ the\ suffix\ S_i}{Total\ Number\ of\ happex\ in\ the\ corpus}$$

Finally, we add P and P* to get the productivity value of every suffix. We have chosen 27 suffixes, 9 of them are very frequent (type frequency ranges from 1,000 to 1,700 words and token frequency 3,000–7,000), 9 are moderately frequent and the rest are least frequent (type frequency below 100 and token frequency below 500). For every suffix, we have computed the type and token frequencies, the number of hapex count and their productivity. We also computed the correlation between the above factors (see Table 12).

We found that, for Bangla, both type and token frequencies significantly correlates among themselves as well as with the happex count which again is inversely correlates to the productivity of the suffix. This implies as the type/token frequency of a suffix increases the higher are the chances of the suffix to form happexes. Although a negative correlation is observed between type/token frequencies, happex count with the suffix productivity, however, no significant correlation could be drawn between them. Therefore, we aim to identify the role of suffix productivity in the processing of words in the mental lexicon. Accordingly, we computed both the conditioned degree of productivity (P) and hapaxed-conditioned degree of productivity (P*) and finally plotted a regression curve between them. The equation of the regression line is depicted in the equation below:

$$P = 0.040 \times P^* - 0.124$$

Table 13 Summarizing the results of suffix productivity model

	Model-5: suffix productivity (values out of 500 words)	Performance	
False positive	44	Precision (%)	84
True negative	129	Recall (%)	73
True positive	240	F-Measure (%)	73
False negative	87	Accuracy (%)	74

We hypothesized that, any point lying above the regression line will be processed via decomposition otherwise they will be processed as a whole. We have evaluated our model with the same set of 500 Bangla polymorphemic words that has been used for the priming experiments. Table 13 depicts the overall result of the evaluation. We observed that as the productivity of the suffix increases, the probability of decomposition of a word also get increases. For example, we observe that the suffixes "-wAlA", "-giri", "-tba", and "-panA" are highly productive (ranges between 0.6 and 0.9) as compared to the suffixes "-A", "-Ani", "-tama", and "-I". Therefore, words having productive suffixes will be more prone to decomposition than the less productive ones. We validate our model with the same 500 words that has been used to validate the previous models. We found an accuracy of around "74 %".

One important observation that can be made from Tables 11 and 13 is that, both the model 4 and model 5 performs best in determining the true negative values. It is also observed that Model-4 possess a high recall value of (85 %) but having a low precision of (50 %) on the other hand results of Model-3 and model-5 possess a high precision of 70 and 84 % respectively. This implies, model-4 can accurately predict those words for which decomposition will not take place. On the other hand model-3 and model-5 accurately identifies those words for which decomposition will occur. Thus, we argue that combining the above three models together can enhance the performance. Hence, in the next section we will present a new model that combines the power of the above three models in determining the decomposability of Bangla polymorphemic words.

Model-6: Combining Model-3, Model-4, and Model-5

From the discussions of the last section we combined model 3, 4 and 5 together to get a new enhanced model. The combination of the models were done by performing both logical AND an logical OR operation on the outputs of Model-3, Model-4 and Model-5. We observe that performance of the OR operation results in a slightly improved accuracy, but both of them are comparable. Thus we have considered performing the logical OR operation over the feature models. This is represented as:

$$Decomposability\,(w) = \begin{cases} TRUE, & if \quad (M3\,(w) \bigcup M4\,(w) \bigcup M5\,(w) = 1 \\ FALSE, & otherwise \end{cases}$$

Similar to the earlier models, we evaluate Model-6 with the same 500 words used in earlier models. The results are depicted in Table 14 (column 7). A comparison of results of our final proposed model with that of the existing ones is depicted in TableResultTable1. The performance of our final model shows an accuracy of 80 % with a precision of 87 % and a recall of 78 %. This outperforms the performance of the other models discussed earlier sections. However, around 22 % of the test words that include words like, *rAShTrIya, nAchuni,*

Table 14 Summarizing the comparative results of the existing frequency based models and our proposed models

	M1 BF	M2 SF	M3 LOG (SF) versus LOG (BF)	M4 TYP/TKN vsSF/BF	M5 P, P*,V	M6 COMBINED
False positive	135	111	88	133	44	32
True negative	111	88	143	212	129	175
True positive	199	155	199	133	240	228
False negative	56	143	67	20	87	64
Precision (%)	60	58	70	50	84	87
Recall (%)	78	51	75	75	73	78
F-Measure (%)	68	54	72	60	74	82
Accuracy (%)	62	49	68	69	74	80

M-1 to M-6 corresponds to Model-1 to Model-6. BF = Base frequency model, SF = Surface frequency model, TYP = Suffix type frequency, TKN = Suffix token frequency, Combine = Combining models 3, and 4 together

nishThAbAna, and juyADi, were wrongly classified by Model-5 which the model fails to justify. Thus, a more rigorous set of experiments and data analyses are required to predict access mechanisms of such Bangla polymorphemic words.

General Discussion and Conclusion

In this paper we attempted to model the representation and processing of Bangla morphologically complex words. Our aim is to determine whether a Bangla polymorphemic word is accessed as a whole or is it decomposed into its constituent morphemes and is recognized accordingly. We tried to answer this question through two different angles. First, we have conducted a series of psycholinguistic experiments based on masked priming paradigm. The reaction time of the subjects for recognizing various lexical items under appropriate conditioning reveals important facts about their organization in the brain which are discussed in the paper.

Our initial results show that morphologically related prime-target pairs do prime each irrespective of their orthographic or semantic relatedness. On the other hand, prime-target pairs that are morphologically opaque do not exhibit any priming effects even if they are orthographically or semantically related. Further, RT analysis of individual words showed that a significant number of Bangla polymorphemic words do not decompose during processing. These observations lead us to believe that mental representation and access of polymorphemic word in Bangla shows the partial decomposition model. We also observe that several other factors including word usage frequency, orthographic complexities, word length and spelling affect the overall word recognition time and accuracy. Each of these factors call for rigorous experimentation for understanding the exact nature of their inter dependencies.

In the second approach, we tried developed a computational model that can predict the recognition process of Bangla polymorphemic words. In order to do so, we have explored the individual roles of different linguistic features of a Bangla morphologically complex word and accordingly proposed different feature models. We finally combine the individual feature models together and propose a new model that can accurately predict the processing of a Bangla morphologically complex word. The combination has been done by

performing both logical OR and logical AND operation over the outputs of the individual feature models. Performance of the logical OR operation is slightly better than that of the AND operation. Finally, we observed that, decomposition of Bangla morphologically complex words depends upon several factors like, the base and surface word frequency, suffix type/token ratio, suffix family size and suffix productivity. The performance of the combined model shows an accuracy of 80 % and this outperform the performance of the individual feature models described in the paper. However, our proposed combined model (MODEL-6) fails to explain the processing phenomena of rest of the 20 % words for which further experiments and RT analysis are required. To the best of the knowledge of the authors there is no other work on computational modeling of Bangla polymorphemic words against which we could benchmark our results.

References

Aitchison, J. (2005). *Words in the mind: An introduction to the mental lexicon*. London: Taylor & Francis.

Ambati, B., Dulam, G., Husain, S., & Indurkhya, B. (2009). *Effect of jumbling the order of letters in a word on reading ability for indian languages: An eye-tracking study*: Proceedings of the 31st Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.

Baayen, H. (2000). On frequency, transparency and productivity. In Booij, G., van Marle, J. (eds.) *Yearbook of morphology*, pp. 181–208.

Baayen, R., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language*, *37*(1), 94–117.

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *55*(2), 290–313.

Bentin, S., & Feldman, L. (1990). The contribution of morphological and semantic relatedness to repetition priming at short and long lags: Evidence from hebrew. *The Quarterly Journal of Experimental Psychology*, *42*(4), 693–711.

Bertram, R., Baayen, R. H., & Schreuder, R. (2000a). Effects of family size for complex words. *Journal of Memory and Language*, *42*(3), 390–405.

Bertram, R., Schreuder, R., & Baayen, R. (2000b). The balance of storage and computation in morphological processing: The role of word formation type, affixal homonymy, and productivity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(2), 489.

Bodner, G., & Masson, M. (1997). Masked repetition priming of words and nonwords: Evidence for a nonlexical basis for priming. *Journal of Memory and Language*, *37*, 268–293.

Bradley, D. (1980). Lexical representation of derivational relation. *Juncture*, pp. 37–55.

Burani, C., & Caramazza, A. (1987). Representation and processing of derived words. *Language and Cognitive Processes*, *2*(3–4), 217–227.

Burani, C., & Laudanna, A. (1992). Units of representation for derived words in the lexicon. *Advances in Psychology*, *94*, 361–376.

Burani, C., Salmaso, D., & Caramazza, A. (1984). Morphological structure and lexical access. *Visible Language*, *18*(4), 342–352.

Caramazza, A., Laudanna, A., & Romani, C. (1988). Lexical access and inflectional morphology. *Cognition*, *28*(3), 297–332.

Carlisle, J. F., & Katz, L. A. (2006). Effects of word and morpheme familiarity on reading of derived words. *Reading and Writing*, *19*(7), 669–693.

Colé, P., Beauvillain, C., & Segui, J. (1989). On the representation and processing of prefixed and suffixed derived words: A differential frequency effect. *Journal of Memory and Language*, *28*(1), 1–13.

Crepaldi, D., Rastle, K., Coltheart, M., & Nickels, L. (2010). 'Fell' primes 'fall', but does 'bell' prime 'ball'? masked priming with irregularly-inflected primes. *Journal of Memory and Language*, *63*(1), 83–99.

Dasgupta, T., Choudhury, M., Bali, K., & Basu, A. (2010). Mental representation and access of polymorphemic words in bangla: Evidence from cross-modal priming experiments. In *International conference on natural language processing*.

Davis, M., & Rastle, K. (2010). Form and meaning in early morphological processing: Comment on feldman, o'connor, and moscoso del prado martn (2009). *Psychonomic Bulletin & Review*, *17*(5), 749–755.

De Jong, N. H., Schreuder, R., & Harald Baayen, R. (2000). The morphological family size effect and morphology. *Language and Cognitive Processes*, *15*(4–5), 329–365.

Drews, E., & Zwitserlood, P. (1995). Morphological and orthographic similarity in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(5), 1098.

Fellbaum, C. (2010). Wordnet. *Theory and Applications of Ontology: Computer Applications*, pp. 231–243.

Ford, M., Davis, M., & Marslen-Wilson, W. (2010). Derivational morphology and base morpheme frequency. *Journal of Memory and Language*, *63*(1), 117–130.

Forster, K., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 680.

Frost, R., Forster, K., & Deutsch, A. (1997). What can we learn from the morphology of hebrew? A masked-priming investigation of morphological representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 829.

Grainger, J., Colé, P., & Segui, J. (1991). Masked morphological priming in visual word recognition. *Journal of Memory and Language*, *30*(3), 370–384.

Hay, J. & Baayen, H.(2001). Parsing and productivity. In *Yearbook of morphology*, p. 35.

Hay, J., & Plag, I. (2004). What constrains possible suffix combinations? On the interaction of grammatical and processing restrictions in derivational morphology. *Natural Language & Linguistic Theory*, *22*(3), 565–596.

Jo, E. (2000). Crowding affects reading in peripheral vision. *Intel Science Talent Search*, 1–15.

Marslen-Wilson, W., Bozic, M., & Randall, B. (2008). Early decomposition in visual word recognition: Dissociating morphology, form, and meaning. *Language and Cognitive Processes*, *23*(3), 394–421.

Marslen-Wilson, W., Tyler, L., et al. (1997). Dissociating types of mental computation. *Nature*, *387*(6633), 592–593.

Marslen-Wilson, W., Tyler, L., Waksler, R., & Older, L. (1994). Morphology and meaning in the english mental lexicon. *Psychological Review*, *101*(1), 3.

Marslen-Wilson, W., & Zhou, X. (1999). Abstractness, allomorphy, and lexical architecture. *Language and Cognitive Processes*, *14*(4), 321–352.

Milin, P., Kuperman, V., Kostic, A., & Baayen, R. (2009). Paradigms bit by bit: An information-theoretic approach to the processing of paradigmatic structure in inflection and derivation. *Analogy in Grammar: Form and Acquisition*, pp. 214–252.

Moscoso del Prado Martn, F., Deutsch, A., Frost, R., Schreuder, R., De Jong, N. H., et al. (2005). Changing places: A cross-language perspective on frequency and family size in dutch and hebrew. *Journal of Memory and Language*, *53*(4), 496–512.

Pylkkänen, L., Feintuch, S., Hopkins, E., & Marantz, A. (2004). Neural correlates of the effects of morphological family frequency and family size: An meg study. *Cognition*, *91*(3), B35–B45.

Rastle, K., Davis, M., Marslen-Wilson, W., & Tyler, L. (2000). Morphological and semantic effects in visual word recognition: A time-course study. *Language and Cognitive Processes*, *15*(4–5), 507–537.

Schreuder, R., & Baayen, R. (1997). How complex simplex words can be. *Journal of Memory and Language*, *37*, 118–139.

Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. *Quarterly Journal of Experimental Psychology Section A*, *57*(4), 745–765.

Taft, M., & Forster, K. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, *14*(6), 638–647.