Mental Representation and Access of Polymorphemic Words in Bangla: Evidence from Cross-modal Priming Experiments

Tirthankar Dasgupta, Monojit Choudhury[1], Kalika Bali[1], and Anupam Basu
Indian Institute of Technology Kharagpur, West Bengal 721302, India
[1]Microsoft Research Lab India, Bangalore 560080, India
E-mail: {iamtirthankar,anupambas}@gmail.com
{monojitc,kalikab}@microsoft.com

Abstract

We conduct a cross-modal priming experiment to determine the mental representation and access strategies for morphologically derived words in Bangla. Analysis of reaction time and error rates indicates that morphologically related words trigger priming effects, even when they are phonologically opaque, and phonologically related but morphologically unrelated words do not exhibit priming. Thus, there is a strong evidence for morphological decomposition of derivationally suffixed words in the minds of the Bangla speakers, though certain suffixes do not seem to evoke decomposition rules. We also observe that word recognition time is also strongly affected of certain orthographic features.

1 Introduction

Understanding the organization of the *mental lexicon* is one of the important goals of cognitive science. *Mental lexicon* refers to the representation of the words in the human mind and the various associations between them that help fast retrieval and comprehension of the words in a given context. 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 and very much a subject of research in psycholinguistics. A clear understanding of these phenomena will not only further our knowledge of how the human brain processes language, but also help in developing apt pedagogical strategies and find applications in natural language processing.

One of the key questions that psycholinguists have been investigating for a long time and debating a lot about is the mental representation and access mechanisms of polymorphemic words: whether they are represented as a whole in the brain or are understood by decomposing them into their constituent morphemes. That is to say, whether a word such as "*unimaginable*" is stored in the mental lexicon as a whole word or do we break it up "*un-*" , "*imagine*" and "-*able*", understand the meaning of each of these constituent and then recombine the units to comprehend the whole word. Such questions are typically answered by designing appropriate priming experiments or other lexical decision tasks. The reaction time of the subjects for recognizing various lexical items under appropriate conditioning reveals important facts about their organization in the brain. (See Sec. 2 for models of morphological organization and access and related experiments).

There is a rich literature on organization and lexical access of polymorphemic words where experiments have been conducted mainly for English, but also Hebrew, Italian, French, Dutch, and few other languages (Frost et al., 1997; Grainger, et al., 1991; Drews and Zwitserlood, 1995). However, we do not know of any such investigations for Indian languages, which are 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 not quite language independent (Taft, 2004). Therefore, the findings from experiments in one language cannot be generalized to all languages making it important to conduct similar experimentations in other languages. Bangla, in

Proceedings of ICON-2010: 8th International Conference on Natural Language Processing, Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2010

particular, features stacking of inflectional suffixes (e.g., *chhele* + *TA* + *ke* + *i* “to this boy only”), a rich derivational morphology inherited from Sanskrit and some borrowed from Persian and English, an abundance of compounding, and mild agglutination.

This work describes some initial experiments to understand the organization of the Bangla mental lexicon at the level of *morphology*. Our aim is to determine whether the mental lexicon decomposes morphologically complex words into its constituent morphemes or does it represent the unanalyzed surface form of a word. We apply the cross modal repetition priming technique, described in (Marslen-Wilson et al., 1994), to answer this question specifically for derivationally suffixed polymorphemic words of Bangla. We observe that morphological relatedness between lexical items triggers a significant priming effect, even when the forms are phonologically unrelated. On the other hand, phonologically related but morphologically unrelated word pairs hardly exhibit any priming effect. These observations are similar to those reported for English and indicate that derivationally suffixed words in Bangla are accessed through decomposition of the word into its constituent morphemes.

Further analysis of the reaction time and error rates per word and per subject reveal several interesting facts such as (a) apart from usage frequency, word length and presence of certain orthographical features also affect the recognition time of a word, and (b) certain derivational suffixes inherited from Sanskrit, which usually make the derived word phonologically or semantically opaque, do not trigger priming; this indicates that these morphological relations are no longer recognized or internalized by the modern Bangla speakers. These and similar other observations make us believe that understanding the precise nature of the mental representation of morphological processes in Bangla (as well as other Indian languages) is a challenging and potent research area that is very little explored.

The rest of the paper is organized as follows: Sec. 2 presents related works; Sec. 3 describes experiment design and procedure; Sec. 4 summarizes the observations and discusses the findings; Sec. 5 concludes the paper by grounding this work in a more general context.

2 Related Works

Over the last few decades many studies have attempted to understand the representation and processing of morphologically complex words in the brain for various languages. Most of the studies are designed to support one of the two mutually exclusive paradigms: the *full-listing* and the *morphemic* model. The *full-listing model* claims that polymorphic words are represented as a whole in the human mental lexicon (Bradley, 1980; Butterworth, 1983). On the other hand, *morphemic model* argues that morphologically complex words are decomposed and represented in terms of the smaller morphemic units. The affixes are stripped away from the root form, which in turn are used to access the mental lexicon (Taft and Forster, 1975; Taft, 1981; MacKay, 1978). Intermediate to these two paradigms is the *partial decomposition model* that argues that different types of morphological forms are processed separately. For instance, the derived morphological forms are believed to be represented as a whole, whereas the representation of the inflected forms follows the morphemic model (Caramazza et.al., 1988).

Traditionally, *priming experiments* have been used to study the effects of morphology in language processing. *Priming* is a process that results in increase in speed or accuracy of response to a stimulus, called the *target*, based on the occurrence of a prior exposure of another stimulus, called the *prime* (Tulving et al., 1982). Here, subjects are exposed to a prime word for a short duration, and are subsequently shown a target word. The prime and target words may be morphologically, phonologically or semantically related. An analysis of the effect of the reaction time of subjects reveals the actual organization and representation of the lexicon at the relevant level.

Priming effects are observed because of the way our brain is supposed to function. Presentation of a stimulus (say a word P) activates a group of neurons (often termed as a *functional web*) in the brain. There are numerous functional webs in the brain; typically a neuron participates in multiple functional webs with varying degree association. When another stimulus (say word T) is then presented to the individual within a short duration (typically 300 ms), the recognition of T is faster if the functional web of T shares a lot of neurons with that of P. This fast reaction time to recognize a stimulus due to the prior exposure to a functionally related stimulus is known as priming. Thus, through priming experiments, we can identify word pairs whose functional webs have a stronger overlap, which in turn reveals how brain organizes the words in the mental lexicon. See Pulvermüller (2002) for a detailed account of such phenomena.

Here we briefly survey some priming experiments for studying morphological processes. These experiments can be classified according to the mode of representing the *prime* and *target* words: (a) when both are visually presented (Frost et.al., 1997), (b) primes are auditorily presented but the targets are visually presented (Marslen-Wilson and Tyler, 1997, 1998; Marslen-Wilson and Zhou, 1999), (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; Stolz and Feldman, 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), 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), and orthographic priming for French (Grainger, et al., 1991) and Dutch (Drews and Zwitserlood, 1995).

We do not know of any cognitive experiments on morphological priming in Bangla or other Indian languages, though there are works on other kinds of priming experiments (see for example, Vasishth et al. (2010) and Ambati et al. (2009)).

3 The Experiment

In order to study the organization of morphologically derived words in Bangla we conduct a *cross-modal repetition task* as described in (Marslen-Wilson et al., 1994). In this technique, a subject hears an audio stimulus and immediately at the offset of the spoken prime, gets a visual representation of the target word[1]. The prime and the target words are either morphologically related or phonologically transparent to each other. A pair of word is said to be *morphologically related* if they meet the following conditions[2]:

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

A pair of word is said to be phonologically transparent if whole or a significant part of one word is fully or partly contained in the other word.

Based on the auditory prime and the visual target probe, the subjects are asked to make some lexical decision, which in this case is to say whether the visually presented target is a valid word in the language. The above experiment is also repeated with the same target words but with different auditory inputs called the *control words*. The control words do not have any morphological or phonological 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).

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

[1] This study follows the experiment 1 of (Marslen-Wilson et al., 1994); however, for the sake of readability we describe the design process and other details.
[2] Thus, we do not consider other morphological processes of Bangla, such as inflections and prefixed forms.

linguistically apparent morphological processes need not have equal priming effects or any effect at all.

3.1 Materials

We selected 56 prime-target pairs, where the primes are derivationally suffixed forms of the targets. In half of these pairs, the words are phonologically related, whereas for the other half the derived forms phonologically differ from the targets. These classes are respectively referred to as [**M+P+]** (or class I) and **[M+P-]** (or class II). In order to further factor out the effect of phonological priming, another 28 prime-target pairs were selected, where the prime phonologically contains the target but they are not morphologically related. This set of prime-target pairs are denoted as **[M-P+] (**or class III**).** Table 1 describes these three classes with examples.

It is interesting to note that while it is very easy to collect word pairs belonging to class I and class III, it is hard to come up with morphologically derived word forms in Bangla which are phonologically opaque. 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 84 target words, we selected another set of 84 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 nor phonologically related to the targets. Some statistics about the prime, target, and control words are presented in Table 2.

Class	**Explanation**	**Example**
I [M+P+]	Words are related morphologically and phonologically	*nibAsa* (residence) - *nibAsi* (resident),
II [M+P-]	Words are morphologically related but phonologically opaque	*pAna* (drink) - *pipAsa* (thirst),
II [M-P+]	Words are phonologically related but not morphologically	*ghaDi* (watch) - *ghaDiYAla* (crocodile)

Table 1: Dataset for the Experiment

Word Type	Avg. Word Length	Avg. No. of complex characters	Average Corpus Frequency
Target	4.0	0.26	1463
Prime	6.4	0.464	1582
Control	6.2	0.472	1514

Table 2: Some statistics of the target, prime, and control words

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 fillers in between the actual prime-target or control-target pairs. We constructed a set of 84 filler pairs which can be categorized into the following three sets of 28 word pairs each: (a) where the prime is a valid word but the target is not, although it is phonologically contained in the prime and is obtained by deleting some word final character-string, e.g., *kapAla* (forehead) – *kapA* (non-word); (b) where the target is a valid word but the prime is not, although it phonologically contains the target and is derived by adding a suffix to the target, e.g *hAtAri* (non-word) – *hAta* (hand); and (c) where both the prime and target are valid words without any morphological and phonological relatedness.

Thus, in all, there are 252 word pairs including 84 prime-target pairs, 84 control-target pairs and 84 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.

Another set of 10 prime-target and prime-control words were collected for a practice session before the true experiments. However, the RTs of these practice pairs are not included in any analysis.

3.2 Procedure

The experiment was conducted using a custom-made graphical user interface (GUI) that randomized the word pairs, played the auditory stimulus and then showed the visual probe for 200ms. Corresponding to each visual probe, subjects had

3000ms to perform the lexical decision after which the system presents the next audio stimulus. The subject performs the decision task by clicking either the "Valid Word" or the "Invalid Word" button of the evaluation GUI. The system automatically records the RT, which in this case is the time between the onset of the visual probe and clicking of one of the buttons 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 10 trial word pairs. As discussed earlier, the experiment is divided into two 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 12 minutes. Between these two phases, there was a break of at least 2 hours. Since a continuous session of 12 minutes require a lot of attention and is tiring for the subjects, we further divided each phase of the experiment into four small sessions of three minutes each. There was a break for two minutes between the sessions.

3.3 Participants

The experiments were conducted on 14 highly educated native Bangla speakers; 9 of them have a graduate degree and 5 hold a post graduate degree. The age of the subjects varies between 22 to 35 years.

4 Results and Discussion

The RTs with extreme values and those for incorrect lexical decisions (about 3.2%) were excluded from the data. We define extreme RT values for each subject as the median RT value of that subject over all prime and control pairs plus/minus 500 ms. Table 3 summarizes the mean RTs for the prime and control sets for the three classes. The *p*-values for *two-sample t-test* and *paired t-test* are also indicated, where the prime and corresponding control RTs have been considered as the two samples or items within a pair. We observe that the average RTs for Bangla control-target pairs are more than the corresponding prime-target ones for [M+P+] and [M+P-] classes. These results are statistically significant according to the *t* statistics. However, we see no significant difference between the prime and control RTs for [M-P+] class. Thus, strong priming effects are observed when the

Class	Average RT (in ms)		*p* values	
	Prime	Control	sample	paired
[M+P+]	1005	1044	< 0.05	<0.01
[M+P-]	1045	1070	< 0.1	<0.07
[M-P+]	1077	1058	> 0.1	<0.09

Table 3**:** Average RT for the word classes and the *p*-values (see text for details)

target word is morphologically related and phonologically contained in the prime word; weak priming effect is observed when the target is morphologically related to but not phonologically contained in the prime; moreover, no priming effects are observed when the prime and target words are phonologically related but share no morphological relationship, which rules out the possibility that priming in [M+P+] could be due to phonological relatedness. These observations together imply that

> *Bangla derivationally suffixed words are in general represented in terms of their constituent morphemes; lexical access and comprehension is facilitated through decomposition of the surface forms into the corresponding constituents.*

Marslen-Wilson et al. (1994) observed similar results for English and Arabic where significant priming is found for morphologically transparent prime-target pairs but priming effect is not observed for morphologically unrelated pairs.[3] The RTs were further analyzed for each participant and each word. A two sample *t*-test for the par-

[3] While the priming trends observed in (Marslen-Wilson et al., 1994) are exactly similar to those reported here, there is a striking difference between the absolute values of the RTs (typically 400-500 ms higher than the values that are usually reported in literature). In order to rule out the possibility of any methodological error in our setup, we repeated the experiments with some of the English word items used in (Marslen-Wilson et al., 1994). The participants were native Bangla speakers, who had 10 to 12 years of formal English-as-second-language education. Surprisingly, their RTs for English words were 100 - 200 ms lower than those for Bangla. This might be due to various factors, and calls for further research. The residual difference of 200-300 ms, perhaps, is explained by the fact that the participants had to make a mouse-click for selecting the appropriate option on the screen, instead of pressing a real button or key which is a more common technique.

ticipants indicates that 8 out of 14 participants exhibited statistically significant priming effects ($p < 0.05$) for [M+P+] class.

Class	Sign Score Range		
	-10 to -4	**-3 to +3**	**+4 to +11**
[M+P+]	15	13	0
[M+P-]	8	18	2
[M-P+]	5	15	8

Table 4: Classification of the Sign Test Values According to Classes and Sign Score Ranges

The corresponding numbers for [M+P-] and [M-P+] are 5 ($p < 0.05$) and 3 ($p < 0.08$) respectively. As we shall discuss later, participants with poor priming effects also had larger error rates, which indicates that perhaps they were not paying enough attention during the experiments. Nevertheless, the overall pattern of priming at the level of individuals is similar to that of the global trend. This further confirms our general conclusion.

4.1 Analysis of RTs for Lexical Items

We also looked at the RTs for each of the 84 target words. Since we had only 14 of observations, one from each participant, we decided to conduct a *sign test* instead of the usual parametric tests of significance (e.g., *t-test*). The sign test is performed by taking the signs of the differences of the RTs for a corresponding prime-target and control-target pair, and then summing (or averaging) over the signs. The null hypothesis here is that the average or sum is 0 (i.e., there are equal number of cases where control RT is greater than prime RT and vice versa). The results are summarized in Table 4. Since we subtracted the control RT from the prime RT, a negative sign indicates priming. Therefore, the smaller the value of the sum for a target word, the more significant is the priming effect. We consider a value less than or equal to -4 as significant. In other words, a target is considered to be significantly primed by the prime word if out of 14 responses, RT for the prime-target was smaller than the RT for the corresponding control-target in at least 9 cases.

The results of Table 4 are in agreement with those in Table 3: maximum priming is observed for words in [M+P+], some priming is evident in [M+P-], but for most of the lexical in [M-P+] no priming effect was observed.

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) and *kShamA* (forgiveness) – *kShamaNIYa* (forgivable) exhibit the least priming effect (sign test sum = +2) in [M+P+]. However, this may not be surprising because, *akarma* and *kShamaNIYa* are very infrequent words, which makes them difficult to recognize.

In [M+P-] class, *pAna* (to drink) – *pipAsA* (thirsty), *dharA* (hold) – *dhairya* (patience) and *chalA* (*move*) – *chAlita* (controlled) show no priming effect (sign score = +3). Apart from phonological and semantic non-transparency, here the target words *pAna*, *chalA* and *dharA* are polysemous[4]. Word sense ambiguity is known to lead to high RTs.

In general, we observe that participants are unable to recognize the morphological connection between certain derivationally suffixed word pairs in the [M+P-] class. Sign test scores for them are close to 0. Examples include *suhRRida* (friend) – *souhArdya* (friendship), *uchit* (appropriate) – *auchitya* (appropriateness) and *hatyA* (murder) – *hi.nsA* (violence). Bangla inherits these morphological forms from Sanskrit.

It is worth mentioning here that sign test has also revealed weak or no priming for a few strongly associated prime-target pairs such as *ghana* (dense) – *ghanatba* (density) and *ga~Nga* (river Ganges) – *ga~Ngajala* (water from Ganges). We could not identify any particular reason for this anomalous behavior, though it could be simply due to the noisiness of RT data.

4.2 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:

(a) very infrequent
(b) long in terms of the number of characters present (≥ 5)

[4] *pAna* also means "get" or "achieve" (2nd person, honorific, simple present) and "beetle leaf"; *dharA* also means "earth"; *chalA* also means "continue".

(c) presence of certain conjugates such as ষ্ট (*Sh+T*), ল্প (*l+p*) and ঙ্গ (*~N+g*), and other irregular or non-transparent glyphs গু (*g + u*) and হৃ (*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 (Foster 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., ষ্ট) have highest recognition time, whereas very frequent glyphs (e.g., ক্ষ = *k+Sh*), even if non-transparent do not seem to have a negative effect on the recognition time of the words.

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.

Class	Avg. RT (ms)	Error (%)
Prime	1000	1.9
Control	1047	1.3
Fillers	1406	6.2

Table 5. Comparison of the RT and error rates between prime, control and fillers.

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. 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.

4.3 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 5 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 for [M+P+] categories, 8 out of 14 participants showed statistically significant priming effects, which led us to hypothesize that the remaining 6 participants were not paying good attention during the experiments or are not well exposed to Bangla (quite unlikely though). Therefore, we would expect their error rates to be higher than that of the other 8 participants. Figure 1 plots the pie chart of error rates for the significantly primed (grey cells) and non-significantly primed (white cells) participants.

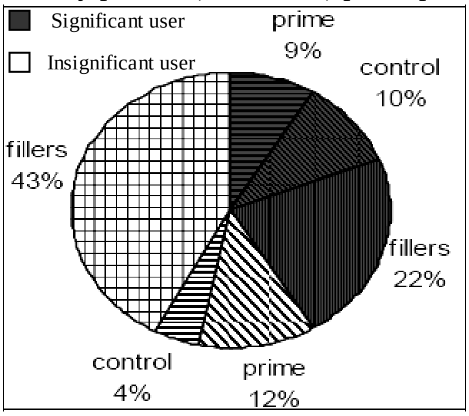

Figure 1. Comparison of error rates for different categories of lexical items. The gray and the white cells are respectively for participants who displayed significant and insignificant priming effects for [M+P+]

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-P+] (see Figure 2). Recall that this is the class, for which we do not observe any priming effect.

5 General Discussions and Conclusions

In this paper we presented some preliminary psycholinguistic experiments to identify the basic representation and processing of Bangla polymorphemic words. Our initial results show that morphologically related prime-target pairs do prime each irrespective of their phonological relatedness. On the other hand, prime-target pairs that are morphologically opaque do not exhibit any priming effects even if they are phonologically related. These observations lead us to believe that mental representation and access of polymorphemic words in Bangla are typically achieved by decomposition of the word into its constituent morphemes, which is the basic premise of the obligatory decomposition model (Taft, 1988; 2004). However, it would be premature to conclude anything concrete based only on the current experimental results.

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.

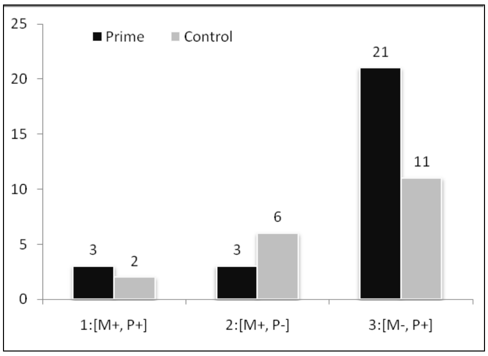

Figure 2: Comparison of error rates across word classes.

Each of these factors call for rigorous experimentation for understanding the exact nature of their inter-dependencies. However, morphological complexity of Bangla throws up many more research questions that we plan to investigate in the future.

First, Bangla features several morphological phenomena such as, inflection, derivation, affix stacking, compounding, echo words and reduplication, as well as complex combinations of these processes. Second, Bangla not only uses native affixes, but also those inherited/borrowed from Sanskrit, Perso-Arabic and English languages. Third, certain morphological processes are rare, while some are quite common. Fourth, there is also a complex interrelation between morphology and semantics[5], phonology and orthography. Fifth, certain morphological processes might be easier to acquire and are perhaps acquired much early during child language acquisition, while some other processes might be internalized much later. It would be interesting to systematically explore all these dimensions and identify the morphological processes and factors that trigger morpheme decomposition rules in the brain versus those forms which are understood as whole word through the *full-listing* model.

[5] In our experiments, we observed that the target *sbabhAba* (nature) is not primed by the derived form *sbAbhAbika* (natural/normal), even though both the words are very frequently used. One possible reason could be semantic opaqueness of the derived form, because the most commonly used sense of *sbabhAba* is "character" while *sbAbhAbika* usually means "normal".

Finally, it is worthwhile to contemplate a little on the broader goals of this research. Understanding the representation and processing of words in the brain is a small, but fundamental, step towards the broader objective of understanding how humans acquire and process languages. This knowledge does not only resolve one of the holy grails of cognitive science, but also is useful for designing better NLP systems based on the understood organic principles. For instance, the lexicon of Bangla is quite large due to the productive morphological processes. Thus, a full listing of words seems an unreasonable approach. On the other hand, many NLP applications including *Machine Translation* and *Information Retrieval* require word and phrasal dictionaries. Understanding the organization of the words in the Bangla mental lexicon can provide very useful pointers to which words should be listed in the lexicon and which one are to be processed on the fly using a morphological analyzer. A representation scheme for the computational lexicon based on the principles of mental lexicon organization is expected to perform better because its success and failure in processing words are expected to meet the expectations of the end user.

References

Ambati B., Dulam G.R., Husain S. and Indurkhya B. (2009). Effect of jumbling the order of letters in a word on reading ability for Indian languages: An eye-tracking study. In N.A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

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

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

Bradley, D. (1980). Lexical representation of derivational relation, *Juncture*, pp. 37-55, Saratoga, CA: Anma Libri.

Butterworth, B. (1983). Lexical Representation, *Language Production*, Vol. 2, pp. 257-294, San Diego, CA: Academic Press.

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

Drews, E., and Zwitserlood, P. (1995).Morphological and orthographic similarity in visual word recognition. *Journal of Experimental Psychology:HumanPerception andPerformance*, 21, 1098–1116.

Forster, K.I., and Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 680–698.

Frost, R., Forster, K.I., & 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*, 829–856.

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

Jo, E. (2000). Crowding affects reading in peripheral vision. *Intel Science Talent Search*. Stuyvesant High School.

MacKay,D.G.(1978), Derivational rules and the internal lexicon. *Journal of Verbal Learning and Verbal Behavior*, 17, pp.61-71.

Marslen-Wilson, W.D., & Tyler, L.K. (1997). Dissociating types of mental computation. *Nature*, *387*, pp. 592–594.

Marslen-Wilson, W.D., & Tyler, L.K. (1998). Rules, representations, and the English past tense. *Trends in Cognitive Sciences*, *2*, pp. 428–435.

Marslen-Wilson, W.D., Tyler, L.K., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review*, *101*, pp. 3–33.

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

Masson, M.E.J., and Isaak, M.I. (1999). Masked priming of words and nonwords in a naming task: Further evidence for a nonlexical basis for priming. *Memory & Cognition*, *27*, 399–412.

Pulvermüller, F. (2002). *The Neuroscience of Language*. Cambridge University Press.

Stolz, J.A., and Feldman, L.B. (1995). The role of orthographic and semantic transparency of the base morpheme in morphological processing. In L.B. Feldman (Ed.) *Morphological aspects of lan-*

guage processing. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Taft, M., and Forster, K.I.(1975). Lexical storage and retrieval of prefix words. *Journal of Verbal Learning and Verbal Behavior*, Vol.14, pp. 638-647.

Taft, M.(1988). A morphological decomposition model of lexical access. *Linguistics*, 26, pp. 657-667.

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

Tulving, E., Schacter D. L., and Heather A.(1982). Priming Effects in Word Fragment Completion are independent of Recognition Memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, vol.8 (4).

Appendix-I: Dataset for the Experiment

Words are arranged in a sequence of triplets (target, prime and control), for class-I, II and III separated by semicolon. The fillers contain only prime-target pairs.

Class-I triplets: Morphologically and phonologically related prime-target pairs (M+P+)
অকর্ম, অকর্মণ্যা, অকম্পমান; অষ্ট,অষ্টধাতু, অংশভাগী; আইন, আইনসংগত,অক্ষরজ্ঞান; সোনা, সোনালি, রুপোলি; কন্ঠ, কন্ঠস্থ, সংগতি; কনক,কনকচম্পা, সংগঠিত; কবচ, কবচধারী, সংগীতকলা; কবর, কবরস্থান, সমবয়স্ক; কবল, কবলগ্রস্ত, সময়রক্ষা; কম, কমজোর, সমবায়ী; ক্ষমা, ক্ষমণীয়, সাদামাঠা; গুরু, গুরুকুল, সাহুকার; ঘন, ঘনত্ব, অগ্নিভ; ঘর, ঘরোয়া, সুফল; খরচ, খরচা, মরচে; গঙ্গা, গঙ্গাজল, আলিঙ্গন; গতি, গতিবিধি, প্রতিনিধি; দানব, দানবিক, পৌরানিক; পথ, পথিক, কথিত; অক্ষর, অক্ষরজ্ঞান, অনির্দিষ্ট; নিষ্ঠা, নিষ্ঠাবান, সাক্ষরতা; নিবাস, নিবাসী, পরবাস; পাগল, পাগলামো , দাড়ানো; বয়স, বয়স্ক, নয়ন; শ্রম, শ্রমিক, মরচে; বাড়ি, বাড়িওয়ালি, কারখানা; কেরানি, কেরানিগিরি, পরলোকগমন; ফুল, ফুলদানি, পাদানি

Class-II triplets: Morphologically related but phonologically opaque pairs (M+P-)
স্বভাব,স্বাভাবিক, আলোচনা; প্রকৃতি, প্রাকৃতিক, আশাবাদী; উচিত, ঔচিত্য, আশঙ্কা; অন্তর, আন্তরিক, কথাবার্তা; দিন, দৈনিক, কদাচিৎ; ভূত, ভৌতিক, কনিষ্ঠ; হত্যা, হিংসা, গুমটি; পান, পিপাসা, ঘড়িয়াল; ধরা, ধৈর্য, খাতা; করা,কর্তব্য,খাটনি; করা, কার্য, ঘুসি; অঞ্চল, আঞ্চলিক, গঙ্গাজল; মিত্র, মৈত্রি, ঘনত্ব; পশু, পাশবিক, কমজোর; দেব, দৈবিক, পথিক; অনু, আণবিক, খালাসি; চলা, চালিত, ঘটক; কল্পনা , কাল্পনিক , দানবিক; চারু, চারিমা, খরচা; কূল, কৌলিন্য , ক্ষমণীয়; মুখ, মৌখিক , চকিত; সুহৃদ, সৌহার্দ্য, কন্ঠস্থ; যন্ত্র, যান্ত্রিক, শ্রমিক; মরা, মৃত, ছুঁচো; সুন্দর, সৌন্দর্য, চতুর্দিক; জঙ্গল, জংলি, অংশী; শরির, শারিরীক, ধারাবাহিক; অর্থ, আর্থিক, কচুরি

Class-III triplets: Phonologically related but morphologically opaque pairs (M-P+)
কচু, কচুরি, দৈনিক; খর, খরগোশ, গুরুকুল; খাট, খাটনি, ঘরোয়া; খাড়া, খাড়াই, নিবাসী; খাত, খাতা, পাতা; গণ্ড, গণ্ডক, মন্দির; গুম, গুমটি, রুপোলি; গুরু, গুরুজন, অষ্টধাতু; ঘট, ঘটক, রজত; ঘড়ি, ঘড়িয়াল, রবিবার; মশা, মশাল, হতাশ; ঘুস, ঘুসি, হাঁসি; আইন, আইনস্টাইন, রবীন্দ্রনাথ; চক, চকিত, পালকি; চট, চটকানো, পালটানো; চতুর, চতুর্দিক, তদারকি; ছুঁচ, ছুঁচো, কাঁচ; টিপ, টিপন্নি, উৎপন্ন; চপ, চপলা, উতলা; খালাস, খালাসি, তামাশা; বর, বরবর, ধরাধরি; কাজ, কাজল, কোকিল; কাঠ, কাঠাল, দুর্বল; কাপ, কাপড়, দাপট; হরি, হরিণ, পলাশি; দেওয়া, দেওয়ালি, বাড়িওয়ালি; আম, আমদানি, কেরানিগিরি; ছিট, ছিটকিনি, ফুলদানি

Fillers-I: Valid target with invalid prime words
খানসামা, খানসা; পালোয়ান, পালোয়া; প্রতাপ, প্রতা; অকৃতদার, অকৃতদ; অগ্নিবাণ, অগ্নিব; কপাল, কপা; কবরস্থান, কব; কমজোরি, কমজ; কর্তব্য, কর্ত; কলাবতী, কলাব; খেজুর, খেজু; কাঁটালি, কাঁট; কাগজপত্র, কাগ; কোঠাঘর, কোঠাঘ; খণ্ডযুদ্ধ, খণ্ডযু; গড়াগড়ি, গড়াগ; গণমত্ততা, গণম; গবেষণা, গবেষ; গর্দান, গর্দ; গাফিলতি, গাফিল; কারসাজি, কারস; গন্তব্য, গন্ত; চিকিৎসা, চিকিৎ; চিত্রশিল্পী, চিত্রশি; অধিনায়ক, অধিনায়; প্রতিনিধি, প্রতিন; সেবক, সেব; জন্মভূমি, জন্মভ

Fillers-II: Valid prime with invalid target words
হীরে, হীরেবার; হিন্দু, হিন্দুশেষ; হাস্য, হাস্যচর; হাতা, হাতানি; হত, হতরি; স্মৃতি, স্মৃতিমর; স্বাস্থ্য, স্বাস্থ্যি; লাফ, লাফার; স্বাগত, স্বাগতির; স্বর, স্বরি; স্বপ্ন, স্বপ্নকারি; স্বদেশ, স্বদেশপ্রায; স্থূল, স্থূলেকির; স্ত্রী, স্ত্রীকের; স্কন্ধ, স্কন্ধকারি; চিত্র, চিত্রকান্ডি; সংযম, সংযমার; সন্তান, সন্তানন; শ্বেত, শ্বেতকেরা; শৌর্য, শৌর্যকার; শোভা, শোভানি; শিখা, শিখারি; লিচু, লিচুকার; লাথি, লাথিকেরা; লাঠি, লাঠিকের; জন্তু, জন্তুসুল; জন্ম, জন্মকা; ঠিক, ঠিকাপান

Fillers-III: Valid prime with valid target words
লঘুকরণ, ভূমিবন্টন; লক্ষ্মী, ভূষণ; রৌদ্র, ভূগোল; রোমাঞ্চ, ভুবন; তারা, ভবিষ্যৎ; রোজ, ব্রহ্মাণ্ড; রোগী, ব্রহ্মপুত্র; রাজহাঁস, বজ্র; যোগী, ব্যায়াম; যৌতুক, ব্যাবহার; যোদ্ধা, ব্যবধান; যোগদান, বোলতা; যুবক, বোকা; যুবরাজ, ফন্দি; যন্ত্রণা, ফলক; কদাচার, ফল; মোহিত, প্লাবন; মৌমাছি, প্রেরণা; মৌরলা, প্রাধান্য; মোম, প্রাণীবিদ; মন্দির, প্রহর; ভ্রমণকারী, নির্মিত; ভ্রাতা, গন্ধ; কম্বল, নির্বিঘ্ন; ভোটদান, নয়ন; ভেলকি, ধ্বজ; তুলনা, দখল; দক্ষিণ, ধাক্কা