Modelling the Organization and Processing of Bangla Polymorphemic Words in the Mental Lexicon: A Computational Approach

Tirthankar Dasgupta Manjira Sinha Anupam Basu Indian Institute of Technology Kharagpur, Kharagpur 721302 iamtirthankar@gmail.com, manjira87@gmail.com, anupambas@gmail.com

Abstract

In this paper we try to present psycholinguistically motivated computational model for the access and representation of Bangla polymorphemic words in the Mental Lexicon. We first conduct a series of masked priming experiment on a set of Bangla polymorphemic words. Our analysis indicates a significant number of words shows morphological decomposition during the processing stage. We further developed a computational model for the processing of Bangla polymorphemic words. The novelty of the new model over the existing ones are, the proposed model not only considers the frequency of the derived word but also considers the role of its constituent stem, suffix and the degree of affixation between the stem and the suffix. We have evaluated the new model with the results obtained from the priming experiment and then compare it with the state of the art. The proposed model has been found to perform better than the existing models.

KEYWORDS: Mental Lexicon, Morphology, Decomposition, Psycholinguistics, Masked Priming.

1 Introduction

The mental lexicon refers to the organization of words in the human mind and their interactions that facilitates fast retrieval and comprehension of a word in a context. One important goal of cognitive science is to understand the organization of mental lexicon as it will help to model how brain processes language. This knowledge will benefit the development various NLP applications that includes text comprehension, lexicon development, information retrieval, text summarization and question answering.

One of the key investigation areas in psycholinguistics is the representation and processing of morphologically complex words in the mental lexicon. That is, for a native speaker, whether a polymorphemic word like "unpreventable" will be processed as a whole (Bradley, 1980; Butterworth, 1983) or will it be decomposed into its individual morphemes "un", "prevent", and "able" and finally recognised by the representation of its stem (morphemic model)(Taft and Forster, 1975; MacKay, 1978). It has been argued that people do have the capability of such decomposition as they can understand novel words like "unsupportable". However, there has been a long standing debate whether such decomposition are obligatory or are they applicable to only those situations where the whole word access fails (Taft, 2004) (partial decomposition model) (Caramazza et al., 1988: Baaven et al., 1997: Baaven, 2000). An alternative to the morphemic and partial decomposition model is the full listing model that assumes decomposition is not at all an obligatory process and the 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). Several computational models have been developed to predict the processing of polymorphemic words. The obligatory decomposition model (Taft, 2004) accounts for the fact that decomposition of a polymorphemic word depends on the frequency of its constituent stem (or the base word). Therefore, higher the stem frequency, easier is the decomposition. On the other hand, the full listing model (Burani and Laudanna, 1992) states that the whole word frequency facilitates the recognition of a polymorphemic word. The dual route access model (Baaven et al., 1997) argues that the decomposition of a polymorphemic word into its constituent morphemes depends on the surface frequency of that word; if the frequency crosses a threshold then the word is accessed as a whole otherwise it is accessed via its parts.

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 existing studies have 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) have also been considered. Any such investigation for Indian languages has not been reported so far, though they are considered to be morphologically richer than many of their Indo-European cousins. On the other hand, several cross-linguistic experiments have indicated that mental representation and processing of polymorphemic words are not language independent (Taft, 2004). The conclusion drawn in one language cannot be generalized to the others without repeating the experiments on them. Bangla, in particular, supports stacking of inflectional suffixes and it has a rich derivational morphology inherited from Sanskrit and some borrowed from Persian, and Arabic, and shows abundance of compounding.

The objective of this paper is to understand the organization and processing of Bangla derivationally suffixed words in the mental lexicon. Our aim is to determine whether the mental lexicon decomposes morphologically complex words into its constituent morphemes or it represents the intact surface form of a word and subsequently develop a robust computational model. To achieve this, first we have conducted the masked priming experiment and gathered reaction time data for next level analysis. The experimental results show that priming occurs only for those cases where the prime is the derived form of the target and have a recognizable suffix (like, sonA-sonAli (GOLD-GOLDEN), and bayasa-bayaska(AGE-AGED). Weak or no priming is observed for cases where the prime is a derived form of the target but do not have a recognizable suffix or when the prime and the target is not morphologically related at all. These observations instigate the basic assumptions of the obligatory decomposition model (Taft and Forster, 1975; Taft, 2004) that polymorphemic words are always processed via decomposition. Deeper analysis of the experimental data reveals that processing of Bangla polymorphemic words may be explained by the dual route decomposition model proposed by (Baayen, 2000). However, unlike the dual route model, our proposed model not only considers the frequency of the derived word but also the role of its constituent stem, suffix and the degree of affixation between them. Our proposed model is the first ever attempt to computationally predict the processing mechanism of a polymorphemic word in any Indian language. We have evaluated our proposed model against the priming experiment results and also compared our performance with that of the existing models in other languages. We have found that our proposed model provides good accuracy for Bangla polymorphemic words which reinforces the language dependent nature of word processing phenomena.

The rest of the paper is organized as follows: section 2 presents related works; section 3 describes the masked priming experiment performed over a set of Bangla morphologically complex words; section 4 compares the performance of different frequency based models in predicting the processing mechanisms of Bangla polymorphemic words; section 5 describes the proposed models of word recognition in Bangla; the last concluding section contains the summary of the observations and discusses the findings.

2 Related Works

Their is a rich literature on representation, organization and accessing of polymorphemic words in the mental lexicon. Typically, priming experiments, and frequency models are used to address such issues. 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. For details please refer to the literature (Caramazza et al., 1988; Bodner and Masson, 1997; Tulving et al., 1982). 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)(Bodner and Masson, 1997)(Davis and Rastle, 2010) (Forster and Davis, 1984) (Frost et al., 1997) (Crepaldi et al., 2010) (Grainger et al., 1991)(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.

Class	Examples
M+S+O+	nibAsa (residence)-nibAsi (resident)
M+S+O-	mitra (friend) - maitri (friendship)
M'+S-O+	Ama (Mango)- AmadAni (import)
M-S+O-	jantu (Animal)- bAgha (Tiger)
M-S-O+	ghaDi (watch)- ghaDiYAla (crocodile)

Table 1: Dataset for the Experiment. M=Morphology, S=Semantics, O=Orthography. + implies related, - implies unrelated.

In the frequency model analysis, (Taft and Forster, 1975) with his experiment on English inflected words, argued that lexical decision responses of polymorphemic words depends upon the base word frequency. In other words, 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). (Baayen, 2000) proposed the dual processing race model 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.

3 Psycholinguistic Study of Bangla Polymorphemic Words through Masked Priming Experiments

We apply the masked priming experiment discussed in (Forster and Davis, 1984; Rastle et al., 2000) (Marslen-Wilson et al., 2008) for Bangla morphologically complex words. Here, the prime is placed between a forward pattern mask and the target stimulus, which acts as a backward mask. This is illustrated below.

 $mask(500ms) \#\#\#\#\# \rightarrow prime(50ms) sonA(GOLD) \rightarrow target(500ms) sonAli(GOLDEN)$ 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 but with a different visual probe called the control word. The control shows no relationship 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).

There were 171 prime-target and control-target pairs classified into five different classes. The prime is related to the target either in terms of morphology, semantics and orthography depending upon the class in which they belong. For example, class-I primes are morphologically, semantically as well as orthographically related where as class-V primes are related only in terms of semantics. The five different class along with their examples are discussed in Table 1.

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

Results:

The RTs with extreme values and those for incorrect lexical decisions (about 3.2%) were excluded from the data¹. Table 2 summarizes the average RTs for the prime and control sets for the five classes. The p-values for two-sample t-test and paired t-test are also indicated, where

¹Any RT value that falls outside the range of Average $RT \pm 500ms$ is considered as extreme

the prime and corresponding control RTs have been considered as the two samples or items within a pair. We observe that, 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; no priming effects are observed when the prime and target words are orthographically related but share no morphological or semantic relationship; although not statistically significant, but weak priming is observed for prime target pairs that are only semantically related. The results for [M+S+O+] and [M-S-O+] classes are statistically significant according to the t-statistics. However, we see no significant difference between the prime and control RTs for other classes.

Class	Avg RT (in ms)		p values		Sign Score Range			
	Р	С	S	Pair	-14 to -4	-3 to +3	+4 to +14	
[M+S+O+]	623	689	< 0.00	< 0.01	24	4	18	
[M + S + O -]	658	660	< 0.09	< 0.06	6	14	19	
[M' + S - O +]	545	549	< 0.10	>0.20	5	7	19	
[M - S + O -]	602	597	>0.20	< 0.10	3	6	22	
[M-S-O+]	590	569	< 0.05	< 0.08	2	5	21	

Table 2: Average RT for the word classes, the p-values and the sign score ranges.

Analysis of RTs for Lexical Items:

We also looked at the RTs for each of the 171 target words. Since we had only 14 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 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 2. 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.

As explained earlier, the effect of priming with a morphologically derived word instigates decomposition, leading to reduced RT of the target. However, it is apparent from the above results that not all polymorphemic words tend to 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.

4 Applying Base Word and Derived Word Frequency Models

The *base word frequency model* (or, Model-1) states that a polymorphemic word that constitute 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:

$$Decomposability(W_i) = \begin{cases} TRUE, & \text{if } \log_{10}(frequency(W_{stem})) \ge \tau \\ FALSE, & \text{if } \log_{10}(frequency(W_{stem})) \le \tau \end{cases}$$
(1)

The derived word frequency model (or, Model-2) claims that, if a specific morphologically complex form is above a certain threshold of frequency, then the whole word access will be preferred and thus no priming effect will be observed 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. Here, the threshold value is computed as the log of average corpus frequency of words² which comes out to be 3 in our case. We apply model-1 and Model-2 to a set of 171 morphologically derived words. The predicted values of both the models are evaluated with respect to the results obtained from the priming experiment discussed in section. performances of the models are computed in terms of Precision, Recall, F-Measure and Accuracy. matrix along with the computed results is depicted in Table 4. We observed that Model-1 posses an accuracy of 62% where as Model-2 has an accuracy of 49%. Table 4 also shows that the false positive and false negative values to be around 11% and 26% respectively. This indicates for these 11% of the words, Model-1 predicts no morphological decomposition due to extremely low base word frequency (ranges between 1 to 7 out of 4 million) but the priming experiment shows high degree of morphological decomposition. On the other hand, model fails to explain why around 27% words (like, ekShatama, juYADi and rAjakiYa) having extremely low base word frequency (ranges between 1 to 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 to 1100). We observe that Model-2 can be used to explain the possible decomposition of low frequency derived words which the base word frequency model fails to explain. Thus, the false positive value for the present model is lower than that of the earlier one (21%). However, the present model performs poorly due to the high false negative value (28%). This implies the model fails to recognise the potentially decomposable words (like, meghalA, pAkAmo and AkAShamandala) properly.

From the above results we observed 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.

5 Combining the Base and the Derived Word Frequencies with Suffix Frequencies

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. We further tried to analyse

²Computed 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)

the role of suffixes in determining the decomposability of Bangla derivationally suffixed words. Accordingly, we followed the same regression based technique discussed in (Hay and Baayen, 2001) to derive relationship between the base and surface word frequencies. 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:

$$Log_{10}(Base\ Frequency) = 0.264 \times Log_{10}(Surface\ Frequency) + 1.822$$
 (2)

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 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 Figure 1 which illustrates the surface and base word frequency distribution of 171 Bangla polymorphemic words. The model predicts that those points that lie on or above the regression line will be parsed during processing where as points lying below the regression line will be accessed as a whole. Next, we compute the type and token





frequencies of the individual suffixes. 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. The hypothesis can be given as, for a given Bangla polymorphemic word if the type/token ratio 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. In order to compute the threshold ratio, we follow the same approach as discussed above. Therefore, we draw a parsing line which is the linear regression line passing from the origin. The slope of the line thus computed is the value of the threshold frequency τ . Thus, the proposed model can be viewed as:

$$Type \ Frequency(S_i) = 0.09 * Token \ Frequency(S_i)$$
(3)

Finally, we combine equation 2 (E2) and equation 3 (E43) together to get a new enhanced model. The combination of the models were done by performing a logical OR operation on the

	False	True	True	False	Р	R	F	A
	Positive	Negative	Positive	Negative	(%)	(%)	(%)	(%)
BF	46	38	68	19	40	54	46	62
SF	38	30	53	49	41	55	47	49
Combined	20	39	51	17	72	75	73	71

Table 3: Summarising the comparative results of the frequency based models. BF= Base frequency model, SF= Surface frequency model, Combine= Combining all the models together. P= Precision, R=Recall, F=F-Measure, A=Accuracy

outputs of E2 and E3. This is represented as:

$$Decomposability(W) = \begin{cases} TRUE, & \text{if } (E3 \lor E4) = 1\\ FALSE, & Otherwise \end{cases}$$
(4)

The enhanced model is evaluated over a set of 136 Bangla polymorphemic words where the stem and the suffixes are transparent (i.e the suffix is fully or partly recognizable). This is because, as automatic identification of opaque Bangla suffixes and computing the frequency is difficult. Thus, for the present model we have not considered the 39 Bangla derived words (belonging to the class [M+S+O-]) for which the stem and suffix is opaque. The results are depicted in Table 4. The performance of our final model shows an accuracy of 71% with a precision of 72% and a recall of 75%. This suppresses the performance of the other models discussed earlier. However, around 29% of the test words that includes words like, *rAShTrIya*, *nAchuni*, *nishThAbAna*, *and juyADi*, were wrongly classified which the model fails to justify.

Conclusion

In this paper we try to model the processing of Bangla words in the mental lexicon. Our aim is to determine whether such words are accessed as a whole or does it is decomposed into its constituent morphemes during recognition. We tried to answer this question through two different angles. First, we conduct a series of masked priming experiments. The reaction time of the subjects for recognizing various lexical items under appropriate conditioning reveals important facts about their organization and processing of words in the brain which are discussed in the paper. Next, we try to develop computational models that can predict the recognition process of Bangla words and validated the prediction through the results of priming experiment. We observed that appart from the surface and base word frequency, decomposition of a Bangla polymorphemic word depends upon the suffix with which the base is attached. The performance of our proposed model shows an improvement of 9% compared to the existing ones. However, further study is needed in order to concretize our claim. To the best of our knowledge there is no other work on computational modelling of Bangla polymorphemic words against which we could benchmark our results.

Acknowledgements

We thank Dr. Monojit Choudhury from Microsoft Research India for his valuable suggestions and SNLTR, Kolkata to support and sponsor our work. We also thank members of SNLTR, Kolkata and Communication Empowerment Lab, IIT Kharagpur for participating in our experiments.

References

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

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

Bentin, S. and 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., Schreuder, R., and Baayen, R. (2000). 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. and 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, pages 37-55.

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

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

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

Butterworth, B. (1983). Lexical representation. Language production, 2:257-294.

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

Colé, P., Beauvillain, C., and 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., and 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., and 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. and Rastle, K. (2010). Form and meaning in early morphological processing: Comment on feldman, o'connor, and moscoso del prado martín (2009). *Psychonomic bulletin & review*, 17(5):749–755.

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

Forster, K. and 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., and 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., and Segui, J. (1991). Masked morphological priming in visual word recognition. *Journal of memory and language*, 30(3):370–384.

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

MacKay, D. (1978). Derivational rules and the internal lexicon. *Journal of Verbal Learning and Verbal Behavior*, 17(1):61–71.

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

Rastle, K., Davis, M., Marslen-Wilson, W., and 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. and 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. and Forster, K. (1975). Lexical storage and retrieval of prefixed words. *Journal of verbal learning and verbal behavior*, 14(6):638–647.

Tulving, E., Schacter, D., and Stark, H. (1982). Priming effects in word-fragment completion are independent of recognition memory. *Journal of experimental psychology: learning, memory, and cognition*, 8(4):336–342.