Compositionality in Bangla Compound Verbs and their Processing in the Mental Lexicon

Tirthankar Dasgupta[1]
TCS Innovation Lab, New Delhi
iamirthankar@gmail.com

Manjira Sinha[1]
Xerox Research Center India, Bangalore
Manjira87@gmail.com

Anupam Basu
Department of Computer Science and Engineering, IIT Kharagpur
anupambas@gmail.com

Abstract

We conduct a cross-modal priming experiment to determine the mental representation and access strategies for compound verbs (CV) in Bangla. Analysis of reaction time indicates that compositionality among CVs triggers priming effects for both the constituent verbs. On the other hand non-compositional CVs exhibit priming only for the polar verb. Thus, compositional CVs are decomposed into their constituent verbs during processing. On the other hand, non-compositional verb phrases are represented and accessed as a whole in the minds of a Bangla speaker. The reaction time data thus collected are used to evaluate our vector space model for compositionality judgment.

1 Introduction

The term compositionality refers to the idea that the meaning of a complex expression is derived from (i) its structure and (ii) the meanings of its constituents (refer to (Fodor and Pylyshyn, 1988) for details). Compound verbs (henceforth CV) in Bangla are known to exhibit continuum of compositionality. Thus, some of them are compositional (khete bashA "sit to eat"), some are non-compositional (jAliye mArA "to irritate") and some in between (ghure berAno "loitering").

There is a considerable linguistic debate on whether CVs are considered as two distinct words connected by some syntagmatic rule or whether they form a single lexical unit (Paul, 2010; Chakrabarti et al., 2008; Butt, 1995) and whether semantic compositionality plays any role deciding the above. These linguistically and computationally challenging issues play an important role in developing lexical resources like WordNet (Fellbaum, 2010).

Since, none of the linguistic arguments and computational approaches has so far led to any consensus, we here for the first time performed a series of psycholinguistic experiments on Bangla compound verbs to study and model how the *mental lexicon*(ML), organize and process such complex expressions. The term mental lexicon refers to the storage of words in the human mind (Aitchison, 2005). A clear understanding of the structure and processing mechanism of CVs in the ML may provide us insight about how to represent and process CVs in computational lexicons. Further, the reaction time of the subjects for recognizing various lexical items under appropriate conditioning may lead us to develop more robust computational models of automatic identification of CVs.

A plethora of works have been done to explore the representation and processing of words in the mental lexicon (refer to Seidenberg, 2012 to get a detail account of it). Attempts have also been made to study the processing of Bangla derivationally suffixed morphologically complex words (Dasgupta et al., 2010; Dasgupta et al., 2012, 2014) and Bangla compound words (Dasgupta et al., 2015). However, to the best of our knowledge

[1] Work done during the tenure at IIT Kharagpur.

D S Sharma, R Sangal and E Sherly. Proc. of the 12th Intl. Conference on Natural Language Processing, pages 65–70, Trivandrum, India. December 2015. ©2015 NLP Association of India (NLPAI)

no such prior attempts have been made to psycholinguistically analyze the representation and processing of the Bangla compound verbs.

Based on the above discussion, the objective of this paper is to explore the semantic compositionality in Bangla CVs and its role in the possible organization and processing of CVs in the ML. Accordingly, we first use empirical techniques to collect user judgment of compositionality for Bangla CVs. Next, we perform the cross modal priming experiment to understand the processing of CVs ML. We found that Bangla CV exhibit continuum of compositionality. Highly compositional verb phrases V_1V_2 trigger significant priming effect for both its constituent verbs V_1 and V_2. On the other hand, non-compositional verb sequences exhibit priming effect only for V_1. These observations together imply that the mental lexicon decomposes compositional verb sequences into their constituent verbs and recognize separately during processing. On the other hand non-compositional verb sequences are organized and accessed as a whole. Based on the dataset collected from the above experiments, we further develop a vector space model to compute semantic compositionality in a verb sequence. The predicted values are compared with the human judgment scores and the priming experiment results where a significant correlation is found.

2 Compositionality Judgments

We could not find any Bangla CV dataset publicly available and annotated in terms of compositionality. Thus, we choose 300 verb sequences and presented them to 36 native Bangla speakers. Similar to the work discussed in (Reddy et al., 2011), we request each subject to give compositionality score of the verb sequences by asking a) how literal the verb phrase is in a given context and b) how literal the constituent verbs are. Subjects were asked to rate the compositionality under 1-10 point scale, 10 being highly compositional and 1 for non-compositional. In order to validate the user annotated data, we compute the inter-annotator agreement using the Fleiss Kappa measure (Fleiss et al., 1981). We found an agreement of κ= 0.68.

3 The Priming Experiment

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 called a functional web) in the brain. When another stimulus (say word T) is then presented to the individual within a short duration, 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 (Tulving et al., 1982). 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, 2003; Spivey et al., 2012) for a detailed account of such phenomena and different types of priming techniques).

Experiment design: We conduct the cross-modal repetition priming experiments as described in (Marslen-Wilson et al., 1994). Here, a subject gets an audio stimulus (called prime (P)) of a verb sequence V1V2 and immediately at the offset of the spoken prime, gets a visual representation of the target (T) word V1. The same audio stimulus is presented to the subject after a random number of trials with the target V2. Based on the auditory prime and the visual target probe, the subjects are asked to decide whether the visually presented target V1 or V2 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 controls(C). The control verb sequences do not have any relatedness with the targets. For example, “kheye felo” (complete eating) and “khAOyA” (to eat) or “kheye felo” and “felo” (to drop) are P-T pairs, for which the corresponding C-T pair could be “niye jAO” (take away) and “khAOyA” (to eat). 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). For both the experiments, RTs of all the targets were recorded and analyzed to observe priming effects.

Materials: The 300 verb pairs as discussed in section 2, are grouped into two classes *highly compositional* (C1) and *non-compositional* (C2). We randomly choose 30 V_1V_2 pairs from each class as primes. For each prime, we have two targets V_1 and V_2, which makes 120 P-T pairs. Consequently, 120 C-T and 240 fillers (or non-words) have been chosen to restrict the subjects to apply any strategic guess during the experiment. Overall, we have collected RTs for 480 word pairs in one experiment. A set of 10 P-T and C-T pairs were also collected for a practice session before the true experiments. However, the RTs of these practice pairs are not included in any analysis.

The experiment was conducted using the DMDX software tool that plays 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 pressing either the “K”' key for a VALID word and “S”' for INVALID word. 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. The experiments were conducted on 36 native Bangla speakers; 29 of them have a graduate degree and 7 hold a post graduate degree. The age of the subjects varies between 25 to 35 years.

Results: The RTs with extreme values (>2000ms) and those for incorrect lexical decisions (about 3.2%) were excluded from the data. The raw RTs for all correct responses were inversely transformed (Ratcliff, 1993) and entered into a mixed-design analysis of variance with two factors: priming (primed and unprimed), and condition (C1, C2). In the subject’s analysis (F1), condition and priming were treated as repeated measures. In the items analysis (F2), priming is treated as repeated factor and condition as independent factor. Table 1 summarizes the mean RTs for the prime (P) and control (C) sets of the V1 and V2 for the two classes.

Class	Example	Avg. Comp	Avg RT	
			V_1(P,C)	V_2(P,C)
C1	*khete basA* (sit to eat)	8.8	666,688	661, 696
C2	*kheye felA* (to eat)	2.3	657,679	687, 659

Table 1: Average compositionality and RTs for the different word classes

There was a strong main effect of priming with faster RTs to primed (667ms for V_1 and 687ms for V_2) than unprimed (679ms and 659ms) targets, $F_1(1, 36)= 47.60$, $p<.01$; $F_2(1, 60)=26.00$, $p<0.01$. There was a main effect of condition $F_1(1, 30)=11.69$, $p<0.01$, $F_2(1,60) =3.51$, $p<0.01$, and a significant condition by priming interaction, indicating that priming effects varied across conditions. Thus, we have observed that when V_1 is presented as target, significant priming occurred for words belonging to both C_1 and C_2. These results are statistically significant according to the 2-way ANOVA test. On the other hand, when V_2 is presented as target with the same prime stimulus V_1V_2, priming is observed only in C2. This result indicates that, compositional verb sequences exhibits priming for both the constituent verbs V_1 and V_2 whereas, non-compositional verb sequences (C1) exhibit priming only with V_1. This may be accounted for by the fact that non-compositional verb sequences derive their meaning mainly from the meaning of its first constituent verb V_1, thus, priming occurs only with V_1. On the other hand, as semantic compositionality between verb sequences increases, both the constituent verbs (V_1 and V_2) plays role in determining the meaning of the whole expression. Thus, both the constituent verbs exhibit priming behavior when preceded by the prime V_1+V_2. The above observations together imply that:

Compositional verb sequences are in general represented in terms of their constituent verbs in the mental lexicon; lexical access and comprehension is facilitated through decomposition of the surface forms into the corresponding constituent verbs. On the other hand non compositional verb sequences are represented and accessed as a whole.

4 Computational Model

Based on the collected data, we now try to develop a vector space model of semantic compositionality to predict the organization and processing of verb phrases in the mental lexicon. We evaluate our model with the data collected from our human judgment compositionality score and the cross-modal priming experiments.

In our vector space model we represent a word's meaning in an n-dimensional space. Here, meaning of individual words is represented in terms of its co-occurrences observed in a large corpus. These co-occurrences are stored in a vector format acquired from a corpus following the different procedures as suggested in the literature (refer to (Mitchell and Lapata, 2008) for a detailed overview).

We have considered the lemmatized context words around the target word in a window of size 200 as the co-occurrences. For the purpose of lemmatization, we have used a Bangla morphological analyzer having an accuracy of around

95%. A Bangla corpus of 33 million words is already available[2]. The corpus contains document from different domain like literature, tourism, news and personal blogs. The top 10000 frequent content words from the corpus are used for the feature co-occurrences. To measure similarity between two vectors, cosine similarity (*sim*) is used. We have used the raw co-occurrence frequency as the values of the constructed vector.

Given a CV W_3 with the constituents W_1 and W_2, the compositionality score S_3 of W_3 is computed as $S_3 = f(S_1, S_2)$. Where, S_1 and S_2 are the literality scores of W_1 and W_2 respectively and f is the semantic compositionality function defined in Column 1 of Table 1. S_1 and S_2 are computed as

$$S_1 = sim(v_1, v_3) \text{ and}$$

$$S_2 = sim(v_2, v_3).$$

Where, v_1, v_2 and v_3 be the co-occurrence vectors of W_1, W_2 and W_3 respectively and sim is the cosine similarity between two vectors computed as:

$$sim(v_1, v_2) = \frac{v_1 . v_2}{|v_1||v_2|}$$

The primary idea behind the constituent based compositionality model is the fact that if a constituent word is used literally in a given verb sequence it is possible that the verb sequence and its constituent share common co-occurrences.

We compare the compositionality score S_3 of all the five composition functions namely, *f1, f2, f3, f4, and f5* with the human annotated compositionality score. The constant parameters (α, β, γ) for all the five models have been computed using list square linear regression technique. We perform a 6 folded cross validation over the test data. The performance of the individual models is reported in Table 2 below. We can observe that the composition functions *f1*and *f5* are better predictor of phrase based compositionality than the other models.

#	F	ρ, R^2	(α, β, γ)
1	$\alpha\, v_1 + \beta v_2 = v_3$	0.73, 0.80	0.02, 0.40
2	$\gamma v_1 v_2 = v_3$	0.40, 0.71	0.32
3	$\alpha v_1 + \beta v_2 + \gamma(v_1 . v_2) = v_3$	0.43, 0.77	0.01, 0.41, 0.33
4	$\alpha v_1 = v_3$	0.30, 0.55	0.12
5	$\beta v_2 = v_3$	0.75, 0.89	0.73

Table 2: Correlation between the human judgment compositionality and the predicted S3 by each of the composition function along with the goodness of fit.

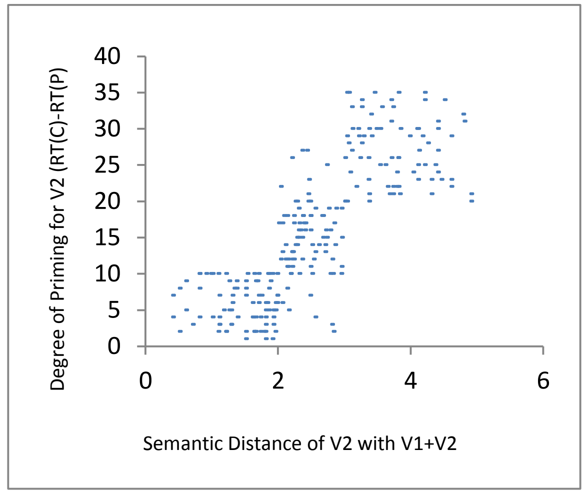

Figure 1: Change in degree of priming of V_1 with respect to the semantic distance of V_1 and V_1+V_2

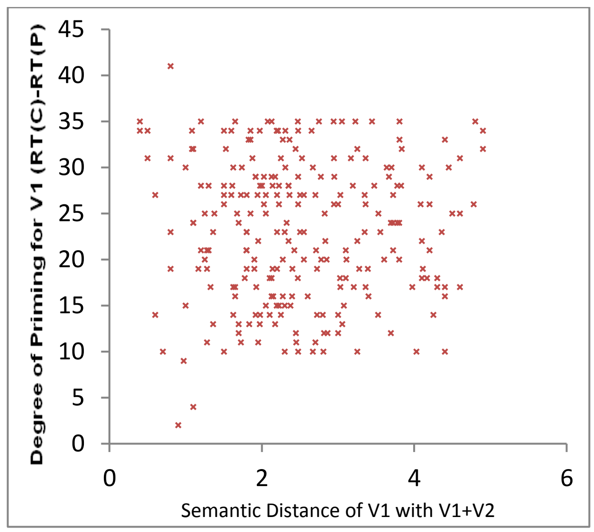

Figure 2: Change in degree of priming of V_2 with respect to the semantic distance of V_2 and V_1+V_2

In the final phase of our work, we have compared the phrase level compositionality score S_3 and the literality scores S_1 and S_2 with the priming behavior of targets V_1 and V_2. The correlation results are reported in Table 3. We observe S_3 has got a high correlation with the priming behavior of the second constituent verb V_2. On the other hand, weak correlation exists between S_3 and V_1. The observations are quite expected, as priming behavior in V_1 is observed irrespective of the semantic compositionality of the phrase V_1V_2,

[2] www.cel.iitkgp.ernet.in

whereas, priming behavior in V_2 changes as compositionality in V_1V_2 changes. This is also reflected in Figure 1 and Figure 2 respectively.

Similar observations are found for S1 and S2. Overall, the results in Table 3 are in agreement with that of Table 1: compositional verb phrases exhibit priming for both its constituents whereas non-compositional verb phrases shows priming only to its first constituent verb (V1).

	V_1	V_2
S_3	0.04	0.78
S_1	0.73	0.01
S_2	0.12	0.74

Table 3: Correlation between priming in V1 and V2 and the computed cosine similarity scores (S1, S2).

5 General Discussion and Conclusion

In this paper we have presented some preliminary psycholinguistic experiments to identify the role of compositionality in the representation and processing of Bangla compound verbs in the mental lexicon. In order to do so, we first computed the user judgment of compositionality score. We found that there is a continuum of compositionality between Bangla compound verbs. Next, we employ the cross-modal repetition priming experiment over a set of Bangla compound verbs. Our initial results show that non-compositional verb sequences exhibit priming only with V_2. On the other hand compositional verb sequences exhibit priming with both the constituent verbs V_1 and V_2. These observations together lead us to believe that mental representation and access of compound verbs in Bangla typically depends on the compositionality of the constituent verbs. Compositional verb phrases are accessed via decomposition of the phrase into its constituent verbs whereas non-compositional phrases are accessed as a whole.

In the next phase of our work, we try to develop a vector space model of semantic compositionality to predict the organization and processing of verb phrases in the mental lexicon. We evaluate our model with the data collected from our human judgment compositionality score and the cross-modal priming experiments. Comparing the models output with that of the empirically collected data, we claim that the proposed vector space model correctly predicts the semantic compositionality of Bangla verb sequences and their possible organization and processing in the mental lexicon. However, it would be premature to conclude anything concrete based only on the current experimental results. We also observe that several other factors including usage frequency and verb ordering affect the overall word recognition time and access mechanisms. Each of these factor calls for rigorous experimentation for understanding the exact nature of their interdependencies that we intend to perform in future.

Acknowledgement

We are thankful to all the participants who actively helped us to conduct the psycholinguistic experiments. We are also thankful to TCS Innovation Lab, Delhi for funding us to travel and attend the conference.

Reference

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

Butt, M. (1995). *The structure of complex predicates in Urdu*. Center for the Study of Language and Inf.

Chakrabarti, D., Mandalia, H., Priya, R., Sarma, V., and Bhattacharyya, P. (2008). Hindi compound verbs and their automatic extraction. In *International Conference on Computational Linguistics*, pages 27–30.

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

Fleiss, J. L., Levin, B., and Paik, M. C. (1981). The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2:212–236.

Fodor, J. A. and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1):3–71.

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

Mitchell, J. and Lapata, M. (2008). Vector-based models of semantic composition. *proceedings of ACL-08: HLT*, pages 236–244.

Paul, S. (2010). Representing compound verbs in indo wordnet. In *Golbal Wordnet Conference*, pages 84–91.

Pulvermüller, F. (2003). *The neuroscience of language: on brain circuits of words and serial order*. Cambridge University Press.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological bulletin*, 114(3):510.

Reddy, S., McCarthy, D., and Manandhar, S. (2011). An empirical study on compositionality in compound nouns. In *Proceedings of IJCNLP*.

Spivey, M. J., Joanisse, M., and McRae, K. (2012). *The cambridge handbook of psycholinguistics*. Cambridge University Press.

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

Seidenberg, M. S. (2012). Computational models of reading. *The Cambridge handbook of psycholinguistics*, page 186.

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*. International conference on natural language processing (ICON). pp. 58-67.

Dasgupta T., Sinha M., Basu A. (2012). "*Computational Models to understand the Access and Representation of Bangla Polymorphemic Words in the Mental Lexicon*". International Conference on Computational Linguistics (COLING), 2012, pp-235-244.

Dasgupta T., Sinha M., Basu A. (2014), "*Computational Modeling of Morphological Effects in Bangla Visual Word Recognition*", Journal of Psycholinguistic Research (JOPR), Springer, 2015.

Dasgupta T., Sinha M., Basu A. (2015). *Computational Models of the Lexical Representation of Bangla Compound Words in the Mental Lexicon*", Journal of Psycholinguistic Research (JOPR), Springer, 2015