Effect of Syntactic Features in Bangla Sentence Comprehension

Manjira Sinha
Xerox Research Center
India[1]
manjira87@gmail.com

Tirthankar Dasgupta
TCS Innovation Lab,
New Delhi[1]
gupta.tirthankar@tcs.com

Anupam Basu
IIT Kharagpur
anupambas@gmail.com

Abstract

Sentence comprehension is an integral and important part of whole text comprehension. It involves complex cognitive actions, as a reader has to work through lexical, syntactic and semantic aspects in order to understand a sentence. One of the vital features of a sentence is word order or surface forms. Different languages have evolved different systems of word orders, which reflect the cognitive structure of the native users of that language. Therefore, word order affects the cognitive load exerted by a sentence as experienced by the reader. Computational modeling approach to quantify the effect of word order on difficulty of sentence understanding can provide a great advantage in study of text readability and its applications. Plethora of works have been done in English and other languages to address the issue. However, Bangla, which is the fifth mostly spoken languages in the world and a relatively free word order language, still does not have any computational model to quantify the reading difficulty of a sentence. In this paper, we have developed models to predict the comprehending difficulty of a simple sentence according to its different surface forms in Bangla. In the course of action, we have also established that difficulty measures for English do not hold in Bangla. Our model has been validated against a number of user survey.

1 Introduction

Complexity of a sentence is the amount of effort a user needs to put in order to understand or comprehend the sentence. Sentence complexity is an important factor in accessing text readability, language acquisition and language impairment. When a reader scans (generally left to right) a sentence, she first processes the syntax (structure and word organization) and semantics (meaning represented by the words) and then reduces them to a semantic whole to store in the memory (Levy, 2013). The short-term memory of the reader engages in real time comprehension of a sentence. While processing a sentence, the short-term memory encounters two types of costs (Oya, 2011): storage cost of the structure built in memory so far and the integration cost due to the proper insertion of the current word into that structure. Therefore, the integration complexity depends upon the relative positions of the entities to be connected, i.e., word order of the sentence.
One of the important grammatical information for sentence interpretation is the word order as it determines the organizations of different grammatical features. It has great impact on the sentence complexity (Meltzer et al., 2010) as it influences both the storage and integration cost and expectation load. Different languages follow different construction rules to build sentences and thus different word orders. Research has been performed to the study effect of word ordering in sentence comprehension in languages like English, Finnish, German (SWINNEY, 1998; Weyerts et al., 2002; Kaiser and Trueswell, 2004). In this paper, the language concerned is Bangla. Bangla is a descendant of the Eastern Indo-Aryan language family 2 . Typologically, it is an inflexional analytic language. Syntax or Sentence structure of Bangla differs from English in many aspects. Bangla is a head final language where the

[1] The work was done during the authors stay at IIT Kharagpur

[2] http://en.wikipedia.org/wiki/Bengali_language

D S Sharma, R Sangal and A K Singh (Eds.): Proc. of the 13th Intl. Conference on Natural Language Processing, pages 275–284,
Varanasi, India. December 2016.

principle word order is subject-object-verb (SOV). It is also a relatively free word-order language as it permits free word order in its constituent chunk or local word group level. Intra-chunk reordering of words is not always permitted; different surface forms of the same sentence are possible, which are grammatically correct; some surface forms are easy to comprehend and some are difficult. Therefore, even simple sentences in Bangla (Chatterji, 1926) can have different surface forms with different comprehending complexities. Till date, there is no prominent study to computationally model the cognitive load associated with different word orders in Bangla.

In this study, our objective is to develop models to quantify the influence of the syntactic and lexical properties of a sentence in sentence comprehension. We have considered simple sentences i.e. sentences having one finite verb[3]. Simple sentences in Bangla can contain many language specific constructs. We have explored the underlying factors responsible for the differences in complexity among different surface forms, such as relative order of subject(s) and object(s) with respect to the verb and organization of non-finite structures. First, we have conducted an empirical user survey, and then we have developed and enhanced our model to reflect the comprehending difficulty experienced by the readers efficiently. In the due course, we have demonstrated that although average dependency distance measure (ADD) (Oya, 2011) works well for English, it is not a good estimator of sentence difficulty in Bangla. Our proposed model takes into account both the relative position and number of unprocessed dependencies at an instant; it is unprocessed dependencies that give rise to expectation gaps in user's cognition. Thus, it models both storage and integration costs on reader's short-term memory in processing a sentence based on different surface forms. We have found high correlation among user preferences and model predictions.

The paper is organized as follows: In Section 2 we have presented the related works in the area of sentence comprehension. In section 3 we have discussed the empirical experiments on Bangla sentence comprehension. Section 4 presents the model building and result analysis. Finally, in Section 5 we conclude the paper.

2 Related Works

A handful of researches have been performed on sentence complexity and word order preference in sentence comprehension. Some approaches are based on dependencies such as placement of verbs in a sentence, position of subject and auxiliary verb in a sentence etc. Several psycholinguistic experiments have been performed to study the role of word order in sentence comprehension.

Research on sentence comprehension difficulty have focused on aspects such as T-unit analysis based on (Bardovi-Harlig, 1992), graph-based approach such as average dependency distance (Oya, 2011), effect of referential processing, subject related clause (SRC) and object related clause (ORC) (Gordon et al., 2001), noun phrases (Gordon et al., 2004), singly nested versus doubly nested structures(Gibson, 2000; Vasishth and Lewis, 2006), effect of semantic context (Tyler and Marslen-Wilson, 1977), influence of hierarchy (Bornkessel et al., 2005), memory interference during language processing like expectation and surprisal (Levy, 2008; Hale, 2006).

The study on comprehension of garden-path sentences started in 1970 (Bever, 1970). Emphasis has also been given on the relationship between linguistic structures in a sentence and language learning (Lachter and Bever, 1988). Within a decade, Bayesian approach towards sentence comprehension based on probabilistic context free grammar (PCFG) (Jurafsky, 1996; Jurafsky, 2002) and competition integration model arrived (Spivey and Tanenhaus, 1998). A different version of competition-integration model was proposed later (Hare et al., 2003) to account for effects of semantic contexts and verb sense effects in sentential complement ambiguity resolution.

Dependency Locality Theory (DLT) (Gibson, 2000) has suggested that during sentence processing, working memory experiences two types of costs. A storage cost to keep in memory what has been encountered so far and an integration cost to assemble sentence components appropriately in order to understand the meaning. The more the distance between an upcoming word and the head it belongs too, the more are the costs. The theory explained the lower comprehension difficulty of SRC than ORC, difficulty of multiple center-embedded structures, ease of cross-referential processing in center-embedded

[3] http://en.wikipedia.org/wiki/Simple_sentence

structure, heaviness effect and sentential ambiguities that were earlier explained using Active Filler Hypothesis (Clifton Jr and Frazier, 1989). Inspired from DLT, average dependency distance based sentence complexity metric has been effective in English and Japanese (Oya, 2011).

Notable works on Sentence comprehension in Hindi have been done in recent times (Vasishth and Lewis, 2006). With the help of SPRT, it has been demonstrated that contrary to the findings in English that center embedding sentences always hinders sentence comprehension speed and efficiency, in certain cases for Hindi, a center embedded doubly nested structure can be beneficial to sentence comprehension depending on the nature of intervening discourse particle between a noun and corresponding verb such as adverb or prepositional phrase. The phenomenon has been termed as anti-locality effect. The word by word reaction time in sentence comprehension has been modeled by spreading activation theory (Anderson, 1983). A study on word by word reading pattern in Hindi has revealed that in head final structures like Hindi, strong expectation overcome the locality effect but not weak expectation (Husain et al., 2014).

3 Empirical user study

Given the subjective nature of text difficulty perception, in order to understand how the different cognitive processes vary across different user groups, two categories of users have been considered for each user study. The choice of participants represents average Indian population. Group 1 consists of 25 native users of Bangla in the age range 21-25 years, who are pursuing college level education and group 2 consists of 25 native users in the age range 13 to 17 years. The former is referred to as the *adult* group and the latter is termed the *minor* group(refer to table 1.1).In this thesis, only the variations in age and years of education have been taken into account. Therefore, I have not fixated on a specific social background and have considered a distribution over a moderate range. The Socio-Economic Classification (SEC) guideline by the Market Research Society of India (MRSI) [4] has been primarily used to determine the social background of the participants. MRSI has defined 12 socio-economic strata: A1 to E3 in the decreasing order. The appropriate class index against a household is assigned by the output of a survey questionnaire collecting data on household items and the education level of the chief wage earner of the family. As can be viewed from the SEC distribution pie-chart (refer to figure 1.5) below, our user group ranges from classes B2 to D2 with only 3 persons from E1 section. The range represent medium to low social-economic sections. The monthly household income ranges from Rs. 4500 to Rs. 15000. To capture the first-language skill, each native speaker was asked to rate his/her proficiency in Bangla on a 1-5 scale (1: very poor and 5: very strong). The distribution has been presented in figure1.4. A significant section of the participants poses medium to poor native language skill. In the backdrop of a country like India it is not exceptional that a person pursuing graduation or higher education is from a medium to low economic background (primarily due to the comparatively low cost of education and a well in place reservation policy) and is not so proficient in the native language.

Type	Background	Mean age (Standard deviation)
Group 1 (adult): 25 native speakers of Bangla	Education: pursuing graduation	22.8 (1.74)
	Socio Economic Classification: B2-D2	
Group 2 (minors): 25 native speakers of Bangla	Education: pursuing school education	15 (1.24)
	Socio Economic Classification: B2-D2	

Table 1.1: User details

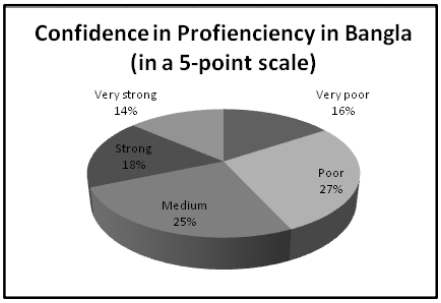

Figure1.1: Proficiency in Bangla

[4] http://imrbint.com/research/The-New-SEC-system-3rdMay2011.pdf

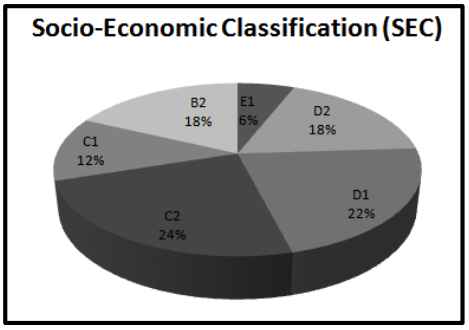

Figure 1.2: Social and economic background of the users.

No.	Feature	Code
1	**Average Dependency Distance** [Total distance among the dependencies in a sentence divided by number of dependencies]	**(ADD)**
2	**Total Word Length** [total length of words in a sentence in terms of visual units (*akshars*)	**(WLN)**
3	**Syllable Count** [total syllable count in a sentence]	**(SLC)**
4	**Sentence Length** [length of the sentence in terms of words]	**(SLN)**
5	**Number of Dependencies** [number of dependencies in a sentence]	**(DCT)**
6	**Average Word Length** [total word length divided by sentence length]	**(AWL)**
7	**Average number of Syllables** [syllable count divided by sentence length]	**(ASW)**
8	**Number of Polysyllabic Words** [number of words in a sentence with more than two syllables]	**(PSW)**
9	**Number of Juktakshars** [number of consonant conjuncts in a sentence]	**(JUK)**
10	**Number of Clauses** [total number of dependent and independent clauses in a complex sentence]	**(CLU)**

Table 1.2: Sentence features studied with respect to users' perception

300 Bangla sentences (both simple and complex) were selected from a Bangla corpus of size 30 million words. The selection was done in a manner to accommodate many varieties of sentence constructions, such as the organization of nouns and verbs within a sentence, lexical combinations within the sentence such as presence of uncommon words that may be perceived as difficult to comprehend by the user, number of non-finite structures in a sentence, number of clauses etc. The sentence lengths vary from 5 to 25 words per sentence.

The participants were asked to perform the following two tasks for each of the 300 sentences. Each participant was allowed maximum 2 minutes time to rate each sentence. Rs.100/- were offered to each of them as a token of appreciation.

- Rate the sentence on a ten point scale (1=easiest, 10=hardest) depending on its overall comprehension difficulty as perceived by the reader.
- Mark words and phrases perceived as ***hard*** *by* the reader.

Inter-annotator agreement was measured by Krippendorf's alpha and it was found to be more than 0.7 for to each group. A paired t-test revealed significant difference ($p<0.05$) between the rating of two user groups.

The primary sentence attributes studied with respect to the user ratings has been given below in table 1.3. features like, average dependency distance and number of dependencies have been studied due to their importance in sentence comprehension.

4 Result analysis

In the first step, we have studied the Spearman rank correlation coefficients between sentence features and user ratings. The coefficients have been shown in figure 1.3.

From the above chart, it can be observed that total word length (in akshars), total syllable count and sentence length is highly correlated with the user ratings. However, the average word length and syllable distribution do not follow the same trend as user ratings. The same is true for dependencies in a sentence; while total number of dependencies posses a high correlation with user data, average dependency distance does not. These may be indicative of the fact that at sentential level, the overall syntactic nature of a sentence is important to the extent of effort required in comprehension rather than the average distribution of the features in a sentence. As expected, in Bangla, number of consonant conjuncts or jukta-akshars has a high correlation with user perception of difficulty. For complex

sentences, user ratings for sentences vary strongly with the number of clauses in the sentence as is evident from the correlation coefficient.

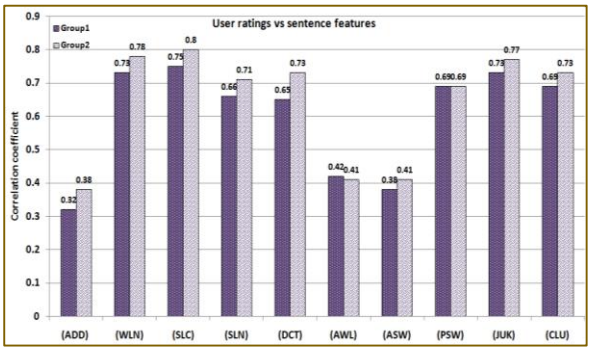

Figure 1.2: User rating versus sentence features

From the above chart, it can be observed that total word length (in akshars), total syllable count and sentence length is highly correlated with the user ratings. However, the average word length and syllable distribution do not follow the same trend as user ratings. The same is true for dependencies in a sentence; while total number of dependencies posses a high correlation with user data, average dependency distance does not. These may be indicative of the fact that at sentential level, the overall syntactic nature of a sentence is important to the extent of effort required in comprehension rather than the average distribution of the features in a sentence. As expected, in Bangla, number of consonant conjuncts or jukta-akshars has a high correlation with user perception of difficulty. For complex sentences, user ratings for sentences vary strongly with the number of clauses in the sentence as is evident from the correlation coefficient.

4.1 Models and regression

In the next step, the effects of sentence attributes have been investigated with regression. At first, detail views of the effects of certain sentence attributes on comprehension difficulty have been shown with graphs from figure 1.4 to figure1.8. Only group 1 data has been shown as group 2 data also follow similar trends.

From figure 1.3, it can be observed that the total word length in a sentence is highly correlated with the user feedback; in figure 1.4 below, we have plotted of word length of a sentence versus user rating and fitted the least square regression line. From the plot, it can be observed that user ratings have approximately a linear relationship with the total word length of a sentence and for a given value of word length the corresponding user ratings are scattered over a very small region around the mean. The regression line therefore has a high R^2 value signifying its goodness of fit. Figure 1.5 represents the relation between sentence length (x-axis) and user ratings (y-axis). It can be observed from the figure that for a given sentence length, the user ratings deflects heavily on both sides of the mean; this may be an indication that although sentence length is an important factor determining sentence comprehension complexity, but at the same time other factors also influence the comprehension load. We have explained the phenomena in the following section with some examples. The plot of average word length and average syllable length versus user ratings (figure 1.6 and 1.7) demonstrates that there is apparently no structured relation between these two features and user ratings. This corroborates the poor correlation coefficients obtained in figure 1.3. Another important attribute of a sentence is the average dependency distance (ADD). Although ADD has been found to explain the order of sentence complexity in English in different cases (Oya, 2011), here, the plot of ADD versus user ratings (refer to figure 1.8) demonstrates no apparent relation between these two; the regression equation in figure 1.8 posses a low R^2 value signifying a poor fit.

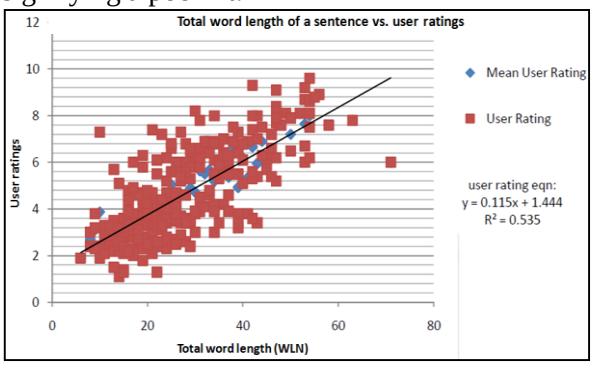

Figure 1.3: Relationship between total word length of sentence and user ratings for Group 1

In the figure 1.4 below, the x-value of a point (represents a sentence) is the total word length and y-value is the user rating of the particular sentence. The similar description also applies for other figures.

Next, we have considered different combinations of the features for predicting sentence complexity using regression with 225 sentences for training and 75 for testing. Results corresponding to the optimal subsets have been presented below in the table 1.2.

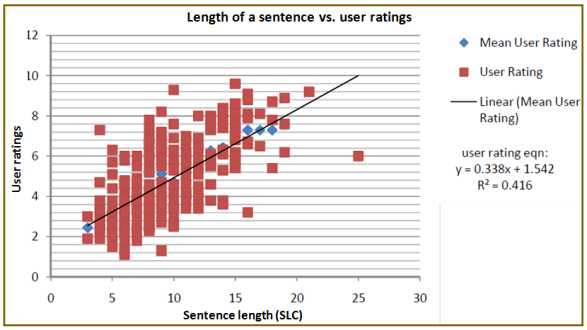

Figure 1.4: Relationship between sentence length and user ratings for Group 1

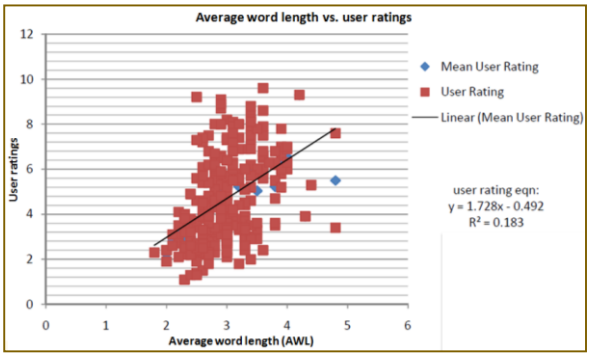

Figure 1.5: Relationship between average word length of sentence and user ratings for Group 1

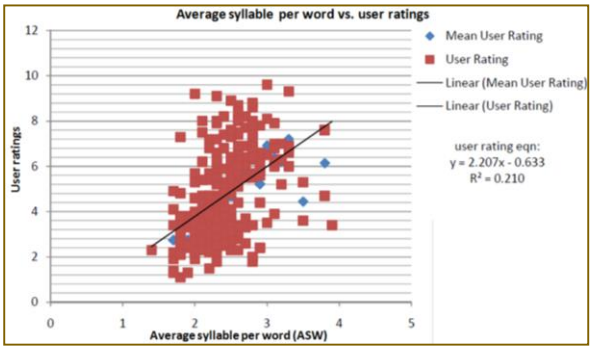

Figure 1.6: Relationship between average syllable per word and user ratings for Group 1

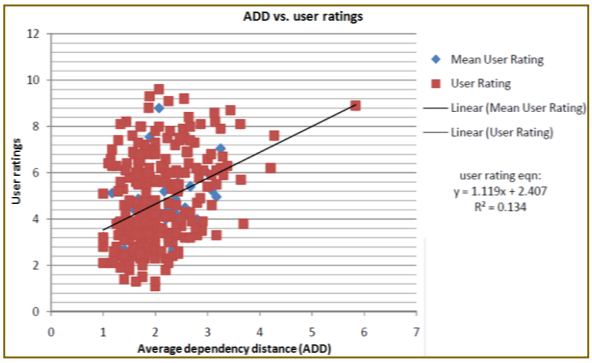

Figure 1.7: Relationship between average dependency distance of sentence and user ratings for Group 1

Below in table 1.3 are the RMSE values for model performances on the training data. From table 1.2 and table 1.3, it can be observed that model 4 for group 1 and model5 for group 2 have relative low error values and higher goodness of fit. Therefore, these models for the two groups can be used as predictive models for sentence complexity for the respective groups given the sentence features. Moreover, these two models have number of dependencies (DCT), total word length (WLN), total syllable length (SLC), number of consonant conjuncts (JUK) and number of clauses (CLU) as independent variables, these features have also found to posses high correlation with user data, therefore, the regression models further assert their contributory roles in sentence comprehension difficulty.

4.2 Effect of lexical and grammatical choice

Apart from the above mentioned syntactic and lexical features of a sentence, the difficulty of individual words also affects the comprehension complexity. To capture this aspect, participants were asked to mark the *hard* words in each sentence during experiment. Now, two outcomes are possible: the hard words will make the sentences difficult to comprehend and therefore they will get relatively higher user ratings than sentences having parameters such as same sentence length and word length; on the other hand, if the reader is able to infer the meaning from the context then the sentences will not be perceived as very difficult to comprehend. Examples of each of the two cases have been provided below in table 1.4 and 1.5 respectively.

In table 1.4, some example sentences with hard words as marked by the readers (shown in bold), have been presented along with user ratings, sentence length and total word length. It can be observed that for both the groups these sentences have user ratings higher than the other sentences with the same sentence length and word length (for a correspondence of group 1 data please refer to figure 1.4 and 1.5 above, group 2 data also follow similar trend).

In table 1.5 below, information about hard words such as their corpus frequency, akshar length, syllable length, number of consonant conjuncts and frequency of their commonly used synonym has been shown. It is apparent that all these words have significantly less corpus presence than their frequent synonyms and therefore are unfamiliar to the general readers. In table 1.5 below is another sets of sentences with hard words marked by the users (shown in bold), but they did not get high user ratings within the same sentence length or word length groups.

Upon further enquiry, it has been revealed that in these sentences, the meaning of the hard words can be resolved from the sentential contexts (the supporting contexts has been underlined).

Model –Expression		Group 1		Group 2	
		R^2	**Err**	**R^2**	**Err**
Model 1	-0.033*DCT + 0.106*JUK+ 0.104*WLN + 1.185	0.57	0.97	0.55	1.10
Model 2	0.114*DCT + 0.225*JUK + 0.137*SLN+ 1.312	0.59	0.81	0.63	0.92
Model 3	**-0.073*DCT -0.025*SLN + 0.081*WLN + 0.071*SLC - 0.015*PSW+ 1.120**	**0.61**	**1.2**	**0.59**	**1.10**
Model 4	0.097*JUK -0.051*SLC + 0.077*WLN + 0.044*CLU - 0.010*DCT+ 1.221	0.67	0.91	0.60	1.30
Model 5	**-0.051*DCT + 0.095*JUK + 0.077*WLN + 0.103*CLU -0.008*SLC + 1.165**	**0.61**	**1.1**	**0.65**	**0.87**

Table 1.2: Regression equations involving sentence features and user ratings (Err means RMSE (root mean square error)

Model	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**
Group 1 RMSE	1.2	1.1	0.97	0.79	0.85
Group 2 RMSE	1.1	0.89	1.3	0.91	0.81

Table 1.3: RMSE values for regression testing

No.	**Sentence (difficult words in bold)**	**Rating**		**Sentence length**	**Word length**
		Gr 1	**Gr 2**		
1.	উভয়ে রোদন **সম্বরণ** করিয়া চক্ষু মুছিলেন *ubhYe rodana **sambaraNa** kariYA cakShu muchilena* (Both of them hold their tears and wipe their eyes.)	5.9	5.5	6	19
2.	**ধাত্রীক্রোড়স্থ** শিশু মার সঙ্গে সঙ্গে কাঁদিতে কাঁদিতে গেল ***dhatrIkroRastha** shishu mAra sange sange k.NAdite k.NAdite gela* (Child went with the mother while crying in her nanny's lap)	6.8	6.5	8	21
3.	কিন্তু কালক্রমে সেই **আগলও** ভেঙেছে ধনতন্ত্রের **দুর্বার** চাপে *Kintu kAlakrame sei **Agalao** bhe.neche dhanatantrera **durbAra** cApe* (That obstacle too has collapsed with time by the mounting pressure of capitalism)	6.5	6.8	8	23
4.	ইতিহাসের বাঁকে কখনো কখনো সেই **সন্ধিক্ষণের** মুখোমুখি হলে আমরা তাকে সহজেই রূপান্তরের লগ্ন বলে চিনতে পারি *Itihasera b.NAke kakhano kakhano sei **sandhikShaNera** mukhomukhi hale AmrA tAke sahajei rupAntarera lagna bale cinathe pAri* (We can easily identify the beginning of change when we encounter that juncture in the turn of history)	8	8.5	16	50

Table 1.4: Sentences with difficult words and their ratings

Sentence	Rating		Sentence length	Word length
	Group 1	Group 2		
বাদল জড়ভরতের মত ঠায় সেই ভাবে বসিয়া আছে *(bAdala jaRabharatera mata thAYa sei bhAbe basiYA Ache)* **Badal has been sitting still ever since.**	2.5	2.7	8	22
গোরার ওষ্ঠপ্রান্তেঈষৎ একটু কঠোর হাসি দেখা দিল *(GorAra oShthaprAnte IShat ekatu kathora hAsi dekhA dila)* **A brief and cruel smile appeared at the corners of Gora's lips.**	2.9	3.2	8	22
বৃক্ষলতাকণ্টক ভেদ করিয়া কল্যাণী বনমধ্যে প্রবেশ করিতে লাগিলেন *(brrikShalatAkanTaka bheda kariYA kalyANI banamadhye probesh karate lAgilena)* **Kalyani started entering the forest penetrating the thorny bushes.)**	4	3.4	8	29
কিন্তু একটি দ্বারেও কপাট বা অর্গলনাই *(kintu ekaTi dbAreo kapATa bA argala nAi)* **(No door has any knob or locking meachanism.**	2.6	2.1	7	17
শুদ্ধা চারী ব্রাহ্মণ বাঘকে খাঁচায় আটক দেখেছিল *(shuddhAcArI brAkShmaNa bAghake kh.NAcAYa ATaka dekhechila)* **The pious bramhin saw the tiger locked in a cage.**	2.5	2.7	6	20

Table 1.5: Sentences with difficult words where contextual resolve is possible

Word	CF	AL	SL	N1	N2
সম্বরণ(*sambaraNa)*	160	4	3	1	আটকানো*(ATkAno)* 603
ধাত্রীক্রোড়স্থ (*dhatrIkroRastha)*	1	5	5	3	কোলে*(kole)* 1728
আগলও(*Agalao)*	59	4	3	0	বাঁধ, বাঁধা, বাঁধন*(b.Nadh/b.NdhA/b.Ndhana)* 510
দুর্বার (*durbAra)*	140	3	2	2	প্রবল, প্রচণ্ড*(prabala/ praconDa)* 5053/3093
সন্ধিক্ষণ(*sandhikShaNa)*	16	4	3	2	সঙ্গে*(sa.nYage)* 147
জড়ভরতের (*jaRabharatera)*	2	6	6	1	স্থির*(sthira)* 4700
ওষ্ঠপ্রান্তে (*oShthaprAnte*)	8	4	4	4	ঠোঁটে*(Th.NoTe)*42
বৃক্ষলতাকণ্টক (*brrikShalatAkanTaka*)	1	7	6	4	কাঁটাবন, কাঁটাঝোপ,কাঁটাগুল্ম,কাঁটাগাছ,কাঁটাতরু *(k.NATAbana, k.NATAjhopa, k.NATAgulma, k.NATAgAcha, k.NATAtaru)*

অর্গল (*argala*)	27	3	2	1	খিল*(khila)* 131
শুদ্ধাচারী (*shuddhAcArI*)	23	4	4	4	ধার্মিক*(dhArmika)* 135

Table 1.6: Information about the difficult words (CF: Corpus frequency, AL: Akshar length, SL: Syllable length, N1: Number of jukta-akshars/ vowel diacritic, N2: Frequency of commonly used synonym)

Sentences	Rating		Sentence length	Word length
	Group 1	Group 2		
চুরি করা মহা পাপ *(curi karA mhApApa)* [Stealing is a grave sin/To steal is a grave sin]	1.2	1.5	3	8
পড়ার চেয়ে লেখা কঠিন *(paRAra ceYe lekhA kathina)* [Writing is harder than reading]	1.9	1.3	4	10
আমরা জিনিস বহন করার জন্য ব্যাগ ব্যবহার করি *(AmrA jinisa bahana karARra janYa bYAga bYabahAra kari)* [we use a bag to carry things]	2.4	3.2	8	22
দয়া করে ঘুমিয়ে থাকা শিশুকে জাগিয়ো না *(daYA kare ghumiYe thAkA shishuke jAgiYo nA)* [please, do not awake the sleeping baby]	1.8	2.2	7	16

Table 1.7: Sentences with different intransitive

In these cases also the relevant information about the hard words has been provided in table 1.6.

In English, non-finite structures have been found to have effect on sentence comprehension (**Error! Reference source not found.**). Non-finite structures are infinitive, gerund, participle, nonfinite complement, nonfinite adjunct etc. In Bangla, no such nonfinite structures like gerund, participle or infinitive exists. Verbs are used in different grammatical formats to serve the purpose of nonfinite forms (Thompson, 2012). To examine the effect of such grammatical properties on sentence comprehension in Bangla, some sentences with such non-finite verbs were included in the experimental data group (shown in table 1.7). However, no significant variations in user ratings have been observed against these sentences as compared to sentences with same length and word length.

5 Conclusion

This paper has presented studies on sentence comprehension in Bangla. In the first part of the paper, relations between sentence features and user ratings have been studied. Subsequently regression formulae based on sentence syntactic and lexical attributes have been developed for both user groups for predicting sentence comprehension difficulty in Bangla. In addition, case by case study of effect of difficult words and presence of non-finite structures on sentence comprehension has also been studied. It has been observed that the way difficult to understand words in a sentence affect the sentence comprehension complexity depends on whether a contextual resolve of the meaning of the sentence is possible or not. Non-finite structures have been found to have no significant effect on sentence comprehension difficulty in Bangla. However, due to the small size (1200 sentences) of dependency annotated corpus in Bangla, the set of possible dependencies is not exhaustive and the probability values for the dependency rules may not always accurately represent the familiarity encountered in practice. This is a limitation of the present approach. Moreover, dependency rules are not formulated as recursive productions in a PCFG, therefore, a hierarchical relation between the increase or decrease in probabilities from one word to other in a sentence has not possible.

In future, it will be interesting to examine whether the accessibility hierarchy (Keenan and Comrie, 1977), observed to be true for English sentence comprehension, also holds in case for

Bangla. Another important path of investigation will be the relative differences in reading behavior among sentences with different relative ordering of noun and verb. Lastly, the word by word reading time study has to be extended for other user groups as well. With this, the study overall sentence readability in Bangla ends; the next and also the last contributory paper in sentence complexity will present study on influence of word ordering of a sentence comprehension in Bangla.

Reference

1. Chatterji, S.-K. (1926). *The origin and development of the Bengali language*, volume 2. Calcutta University Press.
2. Levy, R. (2013). Memory and surprisal in human sentence comprehension.
3. Meltzer, J. A., McArdle, J. J., Schafer, R. J., and Braun, A. R. (2010). Neural aspects of sentence comprehension: syntactic complexity, reversibility, and reanalysis. *Cerebral cortex*, 20(8):1853–1864.
4. Oya, M. (2011). Syntactic dependency distance as sentence complexity measure. In *Proceedings of the 16th International Conference of Pan-Pacific Association of Applied Linguistics*, pages 313–316.
5. SWINNEY, D. A. (1998). The influence of canonical word order on structural processing. *Syntax and semantics*, 31:153–166.
6. Weyerts, H., Penke, M., Münte, T. F., Heinze, H.-J., and Clahsen, H. (2002). Word order in sentence processing: An experimental study of verb placement in german. *Journal of Psycholinguistic Research*, 31(3):211–268.
7. Bardovi-Harlig, K. (1992). A second look at t-unit analysis: Reconsidering the sentence. TESOL quarterly, 26(2):390–395.
8. Oya, M. (2011).Syntactic dependency distance as sentence complexity measure. In Proceedings of The 16th Conference of Pan-Pacific Association of Applied Linguistics.
9. Gordon, P. C., Hendrick, R., and Johnson, M. (2001).Memory interference during language processing.Journal of Experimental Psychology: Learning, Memory, and Cognition, 27(6):1411.
10. Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. Image, language, brain, pages 95–126.
11. Vasishth, S. and Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, pages 767–794.
12. Tyler, L. K. and Marslen-Wilson, W. D. (1977).The on-line effects of semantic context on syntactic processing.*Journal of Verbal Learning and Verbal Behavior*, 16(6):683–692.
13. Bornkessel, I., Zysset, S., Friederici, A. D., von Cramon, D. Y., and Schlesewsky, M. (2005). Who did what to whom? the neural basis of argument hierarchies during language comprehension. *Neuroimage*, 26(1):221–233.
14. Hale, J. (2006). Uncertainty about the rest of the sentence.*Cognitive Science*, 30(4):643–672.
15. Levy, R. (2008).Expectation-based syntactic comprehension.*Cognition*, 106(3):1126–1177.
16. Bever, T. G. (1970). The cognitive basis for linguistic structures.*1970*, pages 279–362.
17. Lachter, J. and Bever, T. G. (1988). The relation between linguistic structure and associative theories of language learning—a constructive critique of some connectionist learning models. *Cognition*, 28(1):195–247.
18. Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation.*Cognitive Science*, 20(2):137–194.
19. Jurafsky, S. N. D. (2002). A bayesian model predicts human parse preference and reading times in sentence processing. *Advances in neural information processing systems*, 14:59.
20. Spivey, M. J. and Tanenhaus, M. K. (1998). Syntactic ambiguity resolution in discourse: modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(6):1521.
21. Hare, M., McRae, K., and Elman, J. L. (2003). Sense and structure: Meaning as a determinant of verb subcategorization preferences. *Journal of Memory and Language*, 48(2):281–303.
22. Clifton Jr, C. and Frazier, L. (1989).Comprehending sentences with long-distance dependencies. In *Linguistic structure in language processing*, pages 273–317. Springer.
23. Kaiser, E. and Trueswell, J. C. (2004).The role of discourse context in the processing of a flexible word-order language.*Cognition*, 94(2):113 – 147.