New Readability Measures for Bangla and Hindi Texts

Manjira Sinha Sakshi Sharma Tirthankar Dasgupta Anupam Basu
Indian Institute of Technology Kharagpur, Kharagpur 721302
manjira87@gmail.com, sakshisharma4u@gmail.com,
iamtirthankar@gmail.com, anupambas@gmail.

ABSTRACT

In this paper we present computational models to compute readability of Indian language text documents. We first demonstrate the inadequacy and the consequent inapplicability of some of the popular readability metrics in English to Hindi and Bangla. Next, we present user experiments to identify important structural parameters of Bangla and Hindi that affect readability of texts in these two languages. Accordingly, we propose two different readability models for each Bangla and Hindi. The models are tested against a second round of user studies with completely new set of data. The results validate the propose models. Compared to the handful of existing works in Hindi and Bangla text readability, this paper presents the first ever definitive readability models for these languages incorporating their salient structural features.

KEYWORDS : Text Readability, Indian Language Texts , Structural Features, Readability Metrics

Proceedings of COLING 2012: Posters, pages 1141–1150,
COLING 2012, Mumbai, December 2012.

1 Introduction

Readability of a text generally refers to how well a reader is able to comprehend the content of a text, through reading. Studies have shown that easy to read texts improve comprehension, retention, reading speed and reading persistence. In this paper we have used the terms readability and comprehensibility interchangeably. Readability is a complex cognitive phenomenon. The cognitive load of a text for a reader depends on the characteristics of a text like lexical choice, syntactic and semantic complexity, discourse level complexity as well as on the background of the user.

The quantitative analysis of English text readability started with L.A. Sherman in 1880 (Sherman, 1893). Till date, English has got over 200 readability metrics. Now there are formulas for Spanish, French, German, Dutch, Swedish, Russian, Hebrew, Chinese, Vietnamese and Korean (Rabin et al., 1988). The existing quantitative approaches towards predicting readability of a text can be broadly classified into three categories (Benjamin, 2012): **traditional methods** incorporate the easy to compute syntactic features of a text like sentence length, paragraph length etc. The examples are Flesch Reading Ease Score (Flesch, 1948), FOG index (Gunning, 1968), Fry graph (Fry, 1968), SMOG (McLaughlin, 1969) etc. The chronologically newer formulas like new Dale-Chall index (Chall, 1995), lexile framework(Stenner, 1996), ATOS-TASA(Learning, 2001), Read-X (Miltsakaki and Troutt, 2007) consider the readers' background and text semantics; **cognitively motivated methods** use high level text parameters like cohesion and cognitive aspects of the reader. Proposition and inference model (Kintsch and Van Dijk, 1978), prototype theory (Rosch, 1978), latent semantic analysis (Landauer et al., 1998), semantic networks (Foltz et al., 1998) are examples of this category. This type of approach introduced text levelling or text revising methods (Kemper, 1983; Britton and Gülgöz, 1991). Two distinguished instances of this class are Coh-metrix (Graesser et al., 2004), and the DeLite software (vor der Brück et al., 2008); the third class of approaches incorporate the power of **machine learning methods** and probabilistic analysis. They are useful in determining online readability based on user queries (Liu et al., 2004) and predicting readability of web texts (Collins-Thompson and Callan, 2005; Collins-Thompson and Callan, 2004; Si and Callan, 2003). Sophisticated machine learning methods like support vector machines have been used to identify grammatical patterns within a text and classification based on it (Heilman et al., 2008).

However, we posit that language plays an important role in the study of readability and the corresponding measures. It has been seen that the first language proficiency increases learning skill and comprehension (Oakland and Lane, 2004). Every language has its own unique properties and any effective metric of readability should be tailored to address language specificities. Some of the specialties of Bangla and Hindi, as compared with English are that these languages are very reach in morphology; they have different grapheme characteristics and their orthography is more phonemic than English; they are head-final and allow free order sentence generation.

Research towards development of readability measures for Bangla and Hindi is still in its infancy. No definitive model of predicting readability in Hindi or Bangla has been proposed in literature yet. Bhagoliwal (1961) applied the Johnson (Johnson and Bond, 1950), Flesch Reading Ease, Farr-Jenkins-Paterson (Farr et al., 1951), and Gunning FOG formulas to 31 short stories in Hindi. In 1965, he examined the features of Hindi typography affecting the legibility of Hindi texts (Bhagoliwal, 1965). Agnihotri and Khanna (1991) applied the classical English formulas to

Hindi textbooks and studied the relative ordering of the predictions against user evaluations. They concluded that along with surface features, readability of a text depends on its linguistic and conceptual organisation. In Bangla, Das and Roychoudhury (2006) studied a miniature model with respect to one parametric and two parametric fits with respect to two structural features of a text: average sentence length and number of syllables per 100 words. Seven paragraphs for seven different texts were used. They found the two-parametric fit as better performer.

In this paper, we first show that the distinguishing features of Bangla and Hindi render the readability models for English untenable for these languages. We next proceed to develop readability indicators for Bangla and Hindi to predict overall difficulty of a text perceived by a native user of the concerned language. Our study is based on the structural features of a text. We have identified three major parameters that contribute to text readability in these languages. Finally we propose two models for each of Bangla (RB1, RB2) and Hindi (RH1, RH2) involving those features.

The organization of the paper is as follows: section 2 defines the features of a text considered in this study, section 3 details the Indian language texts used; section 4 shows the problem of using English readability model in Bangla or Hindi, section 5 deals with the details of user studies and model building. Finally section 6 offers validation and discussion followed by conclusion and perspective.

2 Structural parameters of a text considered in the study

We have considered the following standard structural parameters of a text but customized them to accommodate the specificities of Hindi and Bangla:

1. **Average Sentence Length (ASL):** Total number of words divided by total number of sentences.
2. **Average Word Length (AWL)**: in terms of visual units: Along with dedicated graphemes for consonants and vowels, Bangla and Hindi scripts have some additional graphemes corresponding to the vowel modifiers (diacritic) and consonant conjuncts (jukta-akshars). We consider each kind as a separate visual unit of a word which is equivalent to each alphabet in an English word. The length of a word corresponds to total number of visual units in that word. Average Word Length is equal to total word length divided by number of words.

   Example: शारीरिक = शा + री + रि + क Length = 4

   দিবানিশি=দি + বা+ নি+ শি Length = 4
3. **Average number of Syllables per Word (ASW):** Average number of Syllable per word is equal to total syllable count divided by number of words.
4. **Number of PolySyllabic Words (PSW):** Polysyllabic words are the words whose count of syllable exceeds 2.
5. **Number of PolySyllabic Words per 30 sentences (PSW30):** PSW normalized for 30 sentences.
6. **Number of Jukta-akshars (JUK):** jukta-akshar or consonant-conjunct is consonants occuring together in clusters. When a consonant with a halant (hasanta) is followed by

another consonant, we consider it as one jukta-akshar. The number of jukta-akshars count is the total number jukta-akshars present in the text. The measure is normalized for 50 sentences. Jukta-akshars are not present in English, so the relation between juktakshars and text readability has not been examined before.

Example: साक्षी = स + ◌ा + क + ◌् + ष + ◌ी (jukta-akshars count=1)
শিক্ষা =◌ি + শ + ক +◌্ +ষ +◌া (jukta-akshars count=1)

3 Text selection

Sixteen Hindi and sixteen Bangla texts are selected for the experiment (11 texts) and validation (5 texts) purpose. They cover a broad range of documents types starting from new paper article, short stories, interviews, and blogs to philosophical articles. So we can generalize the model for a variety of text types. Excerpts of length varying from 400 to 1000 words are chosen randomly from the texts to examine the parameters responsible for text readability in case of short as well as long documents. The texts are numbered from 1 to 16 arbitrarily and henceforth will be referred by the text number only.

4 English readability models applied to Hindi and Bangla documents

We have considered the following four models to examine their applicability in Bangla and Hindi, the reason being their high correlation with the established comprehension tests in English (DuBay, 2007; McLaughlin, 1969):

1.	**Flesch Reading Ease** = 206.835 − (1.015 × ASL) − (84.6 × ASW)	3.	**Gunning FOG grade** = 0.4 (ASL+ PSW)
2.	**Flesch-Kincaid Grade-Level** = (0.39× ASL)+(11.8 × ASW)—15.59	4.	**SMOG grading** = 3 + square root of PSW30

TABLE 1-English readability formulas

Although these readability models have been applied to several languages with satisfactory results (Bamberger and Rabin, 1984), in our case, out of bound results are found. As an instance, reading score of Flesch Reading Ease should lie in the range of 0-100, whereas for the Hindi or Bangla texts, its value is more than 150. Grade levels of Flesch-Kincaid Grade Level are not even positive. Grade levels evaluated by Gunning Fog Index and SMOG Index lie far from the expected grades as obtained from user study. The disagreement on the values can be attributed to the significant differences in the language structure of English and Hindi, Bangla as pointed out in introduction. Therefore, we need to start from the scratch in order to develop readability metrics for Bangla and Hindi texts based on structural properties of text.

5 Readability indicator for Bangla and Hindi

As discussed at the end of the previous section, we have developed entirely new readability metrics for Bangla and Hindi based on structural features of a text. In order to achieve this, we have conducted user studies and subsequently built models based on the test results.

5.1.1 Participants

24 native speakers of Hindi and 24 native speakers of Bangla participated in the user studies. Their age group ranges from 24 years to 37 years. 37 of them are from science and engineering background, 10 are from the humanities stream and 1 person is from the commerce stream. 26 of them hold post graduate degrees in their respective fields.

5.1.2 Procedure

Each participant was given the same 16 texts in their native languages in two different sessions: 11 texts during the experiment and 5 texts for the validation. They were asked to rate each on a ten point scale (1=easiest, 10=hardest) depending on its overall comprehension difficulty as perceived by the reader. These results are used to build the readability metrics. Refer to table below (the table contains both experiment and validation texts for sake of convenience):

Text	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Hindi	1.33	5.23	4.44	5.27	3.67	5.21	4.06	4.08
Bangla	3.92	1.54	2.83	1.29	4.23	1.42	2.77	4.83
Text	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Hindi	5.58	4.65	3.35	3.4	4.67	2.31	3.73	3
Bangla	6.08	5.75	5.92	1.38	2.96	2.29	5.33	5.58

TABLE 2- Average grade by each user

5.1.3 User data analysis

The user data have been analysed statistically. To check the degree of variation of different linguistic features to the evaluation done by the users, the Spearman's rank correlate (Zar, 1998) has been computed between them. Table 3 lists correlation between the features and the user study for the 11 experimental texts:

Feature	**Flesch**	**SMOG**	**ASL**	**AWL**	**ASW**	**PSW**	**PSW30**	**JUK**
Hindi	-0.37	0.3	0.4	0.28	0.26	0.21	0.3	0.45
Bangla	-0.76	0.7	0.75	0.8	0.8	0.81	0.75	0.87

TABLE 3-Correlation of textual features, readability scores calculated by Flesch Reading Ease and Smog Index with user evaluation (square-root of PSW30 is omitted as it will have values same as PSW30)

From the above table, it is fairly visible that the best correlated factor with the user's perception of hardness of a text is the number of jukta-akshars present per 50 sentences in the text. There are some interesting findings to be observed. For Hindi the correlation coefficient for ASL is comparatively high but that for ASW is lower. Opposite is the case for Bangla. In both the cases the correlation of user values with the Flesch Reading Ease score is comparatively lower which is based on the assumption that both ASL and ASW are the important factors determining text difficulty. We can see that the assumption does not hold for Bangla or Hindi.

5.2 Feature selection for model building

To make a selection of the features or text parameters that should be incorporated in our models, we have analysed the Spearman' rank correlation among the structural features.

		Bangla					
		ASL	**AWL**	**ASW**	**PSW**	**PSW30**	**JUK**
Hindi	**ASL**		0.53	0.55	0.58	0.91	0.79
	AWL	0.32		0.93	0.75	0.71	0.69
	ASW	0.24	0.85		0.86	0.76	0.79
	PSW	0.48	0.04	0.28		0.74	0.86
	PSW30	0.84	0.51	0.55	0.67		0.92
	JUK	0.83	0.57	0.32	0.34	0.72	

TABLE 4- correlation among structural features of a text for Bangla and Hindi (square-root of PSW30 is omitted as it will have values same as PSW30)

From table 3, we can see that for Hindi, the four mostly correlated text parameters are JUK, ASL, PSW30 and AWL and for Bangla these are JUK, PSW, ASW/AWL and ASL. From 4, we can see that in case of Hindi, ASL and AWL as well as AWL and JUK are loosely correlated (below 0.8), so we have to consider all three in our model as any one of them cannot well represent the trend for the others. PSW30 will anyway be checked while calculating SMOG equivalence. For Bangla, table 4 shows that except for ASW and PSW, the correlation among JUK, PSW, ASL and AWL is less (below 0.8). Therefore, we have to consider all four of them for the same reason as described in case of Hindi.

5.3 Model Building

We have used regression analysis (Montgomery et al., 2007) for model building. In the previous section, we have identified some text parameters which seem as important contributors towards the comprehensibility of a text; we have checked each parameter to obtain an optimized model while giving preference to those. We have used Coefficient of determination[1] or R^2 and Estimate of the error varience (EEV)[2] as measures of goodness of fit of a model. The table 5 below document the short-listed Models (including Flesch (Model 1) and SMOG (Model 2) equivalence) in Hindi and Bangla for which the fittings are optimal from each category.

Model	Expression	R^2	EEV
Hindi			
Model 1	-3.72+0.078*ASL+3.36*ASW	0.35	1.19
Model 2	2.26 + 0.19 * sqrt(PSW30)	0.25	Not calculated

[1] http://en.wikipedia.org/wiki/Coefficient_of_determination

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

Model 3	-2.34+2.14*AWL+0.01*PSW	0.44	1.02
Model 4	0.211+1.37*AWL+.005*JUK	0.36	1.17
Model 5	2.78-0.21*ASL+0.03*PSW+0.01*JUK	0.50	1.07
Model 6	-2.94+.01*PSW+2.77*ASW+.01*JUK	0.46	1.13
Bangla			
Model 1	-10.4+.11*ASL+5.22*ASW	0.58	1.77
Model 2	0.44*sqrt(PSW30) -1.79	0.53	Not Calculated
Model 3	-5.23+1.43*AWL+.01*PSW	0.80	0.82
Model 4	1.15+.02*JUK-.01*PSW30	0.67	1.40
Model 5	5.37+.01*PSW-2.29*ASW+.01*JUK	0.83	0.83
Model 6	5.71+.18*ASL-1.49*ASW+.01*PSW	0.83	0.84

TABLE 5- First round of readability metrics for Bangla and Hindi

6 Validation and Discussion

To carry out the validation study we took the same 24 users for Bangla and 24 users for Hindi and a completely new set of 5 texts for each of the two languages. The users were asked to perform the same operations on each text as described in the Procedure part. We have applied our 6 shortlisted readability models (refer Table 5) to the validation texts. The comparative analysis of prediction made by our readability models to the actual scores given by the users are summarized below. From the results it can be inferred clearly that root mean square errors for model 3 and model 4 stand out as the bottoms among their respective groups. So, we propose these two models as our readability metric for Bangla and Hindi.

		Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
RMSE(square-root(MSE))	**Hindi**	1.086	1.085	1.04	0.81	2.06	2.23
	Bangla	1.32	1.19	0.85	1.13	1.19	3.51

TABLE 6- Summary of validation results

One interesting thing to be noted here, although the two selected models are the two top fits for both Bangla and Hindi, model 3 in Bangla is the best fit whereas, model 4 is the best fit for Hindi. Model 3 in both the cases comprise of AWL and PSW, but for Hindi model 4 has AWL and JUK, whereas for Bangla it consists of JUK and PSW30. These once again prove our initial assumptions that for different language, different textual features contribute to readability and an effective readability indicator is language dependent.

We name the two models for Hindi as RH1 (model 3), RH2 (model4) and for Bangla they are RB1 (model 3), RB2 (model 4). The figure 1 below graphically represents the comparison of user scores with that of the proposed model for Hindi and Bangla; the straight lines represent the trendlines of the respective curves. We can see that in both cases, the models closely follow the users' response curves. The models differ very slightly in accuracy and they feature different text parameters, so, any model alone may not suffice to correctly predict text difficulty. Therefore, we have decided to keep both the models as measure of how the three different structural dimensions of a text contribute to its comprehensibility.

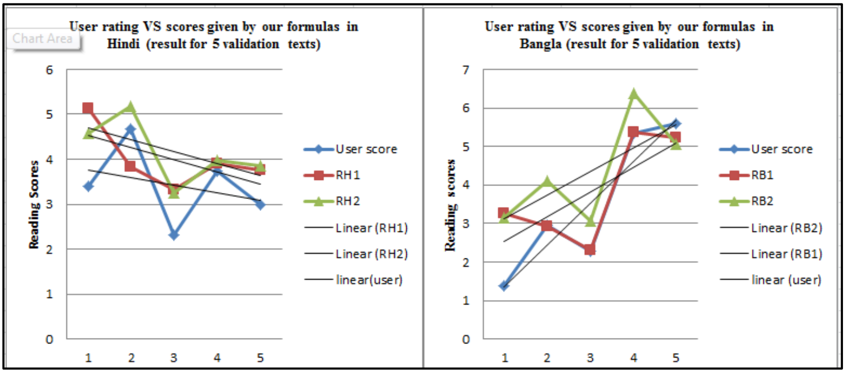

Figure 1: Graph representing the predicted scores versus user evaluation

Conclusion and perspective

In this study, we have developed two new readability measures: RH1, RH2 and RB1, RB2 for Hindi and Bangla text documents respectively. We have also identified AWL, PSW, JUK and PSW30 as major factors affecting readability in Hindi and Bangla. We have shown that for these languages, English readability formulas are not helpful as text difficulty. The two previous studies in Hindi (Bhagoliwai, 1961; Bhagoliwal, 1965; Agnihotri and Khanna, 1991) have applied English readability formulas like Flesch on Hindi passages and school level textbooks, but none of them proposed any definitive model for Hindi text readability like ours. In case of Bangla readability (Das and Roychoudhury, 2006) have compared one and two parametric fits for a miniature model, but they have not considered parameters like AWL, JUK; we have found these parameters to be the major players. The proposed readability models for Bangla and Hindi incorporating features like AWL, JUK have been validated against extensive user studies. In future, we plan to extend this work to different sections of users to obtain readability models, more appropriately related to different user groups.

References

Agnihotri, R. K. and Khanna, A. L. (1991). Evaluating the readability of school textbooks: An indian study. Journal of Reading, 35(4):pp. 282–288.

Bamberger, R. and Rabin, A. T. (1984). New approaches to readability: Austrian research. The Reading Teacher, 37(6):pp. 512–519.

Benjamin, R. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. Educational Psychology Review, 24:1–26.

Bhagoliwal, B. (1961). Readability formulae: Their reliability, validity and applicability in hindi. Journal of Education and Psychology, 19:13–26.

Bhagoliwal, B. (1965). Typographic dimensions affecting the legibility of hindi print: a factorial experiment. Journal of Education and Psychology.

Britton, B. and Gülgöz, S. (1991). Using kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. Journal of Educational Psychology, 83(3):329.

Chall, J. (1995). Readability revisited: The new Dale-Chall readability formula, volume 118. Brookline Books Cambridge, MA.

Collins-Thompson, K. and Callan, J. (2004). A language modeling approach to predicting reading difficulty. In Proceedings of HLT/NAACL, volume 4.

Collins-Thompson, K. and Callan, J. (2005). Predicting reading difficulty with statistical language models. Journal of the American Society for Information Science and Technology, 56(13):1448–1462.

Das, S. and Roychoudhury, R. (2006). Readability modelling and comparison of one and two parametric fit: A case study in bangla*. Journal of Quantitative Linguistics, 13(01):17–34.

DuBay, W. (2007). Smart Language: Readers, Readability, and the Grading of Text. ERIC.

Farr, J., Jenkins, J., and Paterson, D. (1951). Simplification of flesch reading ease formula. Journal of applied psychology, 35(5):333.

Flesch, R. (1948). A new readability yardstick. Journal of applied psychology, 32(3):221.

Foltz, P., Kintsch, W., and Landauer, T. (1998). The measurement of textual coherence with latent semantic analysis. Discourse processes, 25(2-3):285–307.

Fry, E. (1968). A readability formula that saves time. Journal of reading, 11(7):513–578.

Graesser, A., McNamara, D., Louwerse, M., and Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. Behavior Research Methods, 36(2):193–202.

Gunning, R. (1968). The technique of clear writing. McGraw-Hill NewYork, NY.

Heilman, M., Collins-Thompson, K., and Eskenazi, M. (2008). An analysis of statistical models and features for reading difficulty prediction. In Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications, pages 71–79. Association for Computational Linguistics.

Johnson, R. and Bond, G. (1950). Reading ease of commonly used tests. Journal of Applied Psychology, 34(5):319.

Kemper, S. (1983). Measuring the inference load of a text. Journal of educational psychology, 75(3):391.

Kintsch, W. and Van Dijk, T. (1978). Toward a model of text comprehension and production. Psychological review, 85(5):363.

Landauer, T., Foltz, P., and Laham, D. (1998). An introduction to latent semantic analysis. Discourse processes, 25(2-3):259–284.

Learning, R. (2001). The atos readability formula for books and how it compares to other formulas. Madison, WI: School Renaissance Institute.

Liu, X., Croft, W., Oh, P., and Hart, D. (2004). Automatic recognition of reading levels from user queries. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pages 548–549. ACM.

McLaughlin, G. (1969). Smog grading: A new readability formula. Journal of reading, 12(8):639–646.

Miltsakaki, E. and Troutt, A. (2007). Read-x: Automatic evaluation of reading difficulty of web text. In Proceedings of E-Learn.

Montgomery, D., Peck, E., and Vining, G. (2007). Introduction to linear regression analysis, volume 49. John Wiley & Sons.

Oakland, T. and Lane, H. (2004). Language, reading, and readability formulas: Implications for developing and adapting tests. International Journal of Testing, 4(3):239–252.

Rabin, A., Zakaluk, B., and Samuels, S. (1988). Determining difficulty levels of text written in languages other than english. Readability: Its past, present & future. Newark DE: International Reading Association, pages 46–76.

Rosch, E. (1978). Principles of categorization. Fuzzy grammar: a reader, pages 91–108.

Sherman, L. (1893). Analytics of literature: A manual for the objective study of english poetry and prose. Boston: Ginn.

Si, L. and Callan, J. (2003). A semisupervised learning method to merge search engine results. ACM Transactions on Information Systems (TOIS), 21(4):457–491.

Stenner, A. (1996). Measuring reading comprehension with the lexile framework.

vor der Brück, T., Helbig, H., Leveling, J., and Kommunikationssysteme, I. (2008). The Readability Checker Delite: Technical Report. FernUniv., Fak. für Mathematik und Informatik.

Zar, J. (1998). Spearman rank correlation. Encyclopedia of Biostatistics.