Readability Analysis of Bengali Literary Texts

Hadiuzzaman

A thesis presentation of Digital Bangladesh

Computer Science and Engineering
Bangladesh Army university of Science and Technology

1.Introduction

Although the definition of readability is quitevague, itisgenerally unders to odto be the ease of reading and comprehending texts. The acceptability of a writer's work depends upon how much the text is readable and understandable to the mass. This encouraged research in the fields of people's reading skills and quantification of readability. Although some researchers like Pikulski (2002) havequestionedsuchquantificationintheformofautomaticreadabilityscoring formulas(36),theydo findgooduseindifferent fields,includingbutnotlimited togradationofschoolbooksatdifferentgradelevels,andautomaticassessme text comprehension material. Readability tests are regularly used in different application areas such as gradation of schoolbooks (Aukerman, 1965; Ayodele, 2013;Spache,1953),me and Taub,1983;Paasche-Orlow, Taylor, and Brancati, 2003); financial texts (Loughran

and McDonald, 2010) and software readability (Buse and Weimer, 2008). Such applications of readability assessment are yet to be seen for relatively under-resourced languages such as Bengali. Readabilityofatextmostnaturallydevolvesonthemotivationandinterestof the reader. Physical characteristics of a text that affect its readability are paper quality, font size, print quality, misspellings and vocabulary. Other than these, readability depends on reader characteristics such as 'speed of perception''perceptibility at a distance', 'perceptibility in peripheral vision', 'visibility', 'the reflexblinktechnique','rateofwork'(i.e.speedofreading),'ey 'fatigue in reading' (Tinker, 1963). Most of the classical readability indices (e.g. the ones proposed by Flesch, Dale and Chall, McLaughlin, and Fry, 1977) deal with lexical features (Dale and Chall, 1948; Flesch, 1948; Lorge,1944; McLaughlin, 1969), such as average sentence length, average word

length, usage of difficult words, polysyllabic words, etc. It was recently shown that language-specific features– such as parts-of-speech (POS)-based features, semantic features and discourse features – can significantly affect readability (Feng, Jansche, Huenerfauth, and Elhadad, 2002).
So far, these studies have mostly been performed in English and other highresourced languages such as Chinese, German and French (François and Fairon, 2012; Hancke, Vajjala, and Meurers,2012; Henry,1975; Pang,2006; Schwartz, 1975;Walters,1966;Yang,1970).Readabilityresework of Das and Roychoudhury in 2000 (Das and Roychoudhury, 2000). Followup studies in 2003 and 2006 modelled Bengali readability as a parametric fit to human-annotated ground truth data (Das and Roychoudhury, 2004, 2006). Although pioneering, Das and Roychoudhury's work was stymied by the lack of a large enough dataset, and inadequate analysis of features used

in their regression models. Later studies by Sinha, Sharma, Dasgupta, and Basu (2012), Islam, Mehler, Rahman, and Text-technology (2012) and Islam, Rahman, and Mehler(2014)improvedthestateoftheartbyintroducingnov models and datasets. To date, however, Bengali readability research remains in its infancy, owing partly to the lack of large enough human-annotated ground truth data. In this article, we have done our study over a dataset of 30 text passages, excerpted from the literature of 4 eminent writers of Bengali language. These texts were rated for readability on a 7-point Likert Scale by seven independent human annotators. We conducted an inter-annotator agreement study on these ratings as shown in Section 4, and modelled Bengali readability on a set of 18 lexical, syntactic and semantic features. Regression analysis was applied over these features to generate our models. We measured the goodness-of-

fit for our models using adjusted R2 and mean squared error (MSE). In terms of MSE, our best model was found to outperform the models proposed by Sinha et al. (2012). The rest of this chapter is organized as follows. We discuss related research in Section 2, followed by a discussion of our problem, dataset and annotation scheme in Section 3. Feature description, feature selection and model construction are detailed in Section 5.2.1. Section 7 contains results and discussions and Section 8 concludes the chapter with contributions, limitations and future research directions

2 Related Work

Many different readability formulas have been proposed for English (Dale and Chall, 1948; Flesch, 1948; Kintsch and Van Dijk, 1978; Lorge, 1944; McLaughlin, 1969). Most of the classical formulas depend on surface-level features such as average sentence length

(ASL), average word length (AWL), number of syllables and number of polysyllabic words. Senter and Smith (1967) argued that factors related to word difficulty and sentence difficulty were more important than surface features. Liu, Croft, Oh, and Hart (2004) treated readability estimation as a text categorization problem and used support vector machines (SVM) on syntactic and semantic features. Feng et al. (2010) did a comprehensive feature analysis for readability modelling. They considered several fine-grained feature categories like shallow features, discourse features, POS-based features, language modelling features, lexical chain features, etc. They observed that POS features (nouns in particular) and shallow features (average sentence length) had the best predictive performance.
ical readability formulas (Pikulski., 2002). Pikulski affirms what the International Read-

ing Association points out as the two major clusters of factors that are not assessed by readability formulas. First is the conceptual readability, consisting of factors such as 'density of concepts, abstractness of ideas, text organization and coherence and sequence of ideas', and second is the format or design cluster consisting of 'page format, length of type line, length of paragraphs, and the use of illustrations and color' (Gray and Leary, 1935). Durr, Hillerich, and Mifflin (1986) tried to overcome this problem and incorporated the interaction between readers (students) and readability formulas while implementing the latter in their reading programme. While most readability studies have been performed in English, a few other high-resourced languages also enjoyed a good amount of readability research. In German, Hancke et al. (2012) introduced derivational and inflectional morphology of nouns as fea-

tures for readability classification. Schwartz (1975) developed the 'German Readability Graph' in line with the 'Fry Readability Graph' (Fry, 1977). Number of syllables and number of sentences in 100-word samples were plotted on the graph. Glöckner, Hartrumpf, Helbig, Leveling, and Osswald (2006) overcame the limitations of traditional readability formulas for German by employing powerful NLP techniques to extract causal factors of readability instead of approximating them by surface features such as sentence length. Glöckner et al.'s system further rewrites the text for better readability. In French, Henry (1975) gave an outline of readability measures. François and Fairon (2012) proposed a new 'French as a foreign language' readability formula. They took into account 46 textual features representing the lexical, syntactic and semanti levels, as well as some of the context sensitivity. Franc¸ois and

Fairon showed that maximizing the type of linguistic information does not always yield the best results, and avoiding semantic features did not affect the performance of their best models.

Chinese has seen a good amount of readability research (Chen, Chen, and Cheng, 2013; Jeng, 2001; Lee et al., 2012; Pang, 2006; Yang, 1970). Yang's (1970) study was a pioneering e ffort in Chinese readability assessment. He collected 85 Chinese passages from different genres of writing and found that the number of strokes in a character, the presence of words in a basic word list, and the proportion of full sentences in a passage had the most predictive power for text reading difficulty. Yang derived two readability formulas with 31 and 7 predictors respectively. More recently, Jeng (2001) compared linear and nonlinear approaches to measure readability of children's literature. His corpus had

223articleswithatotalofmorethan82,000wordsselectedfron
sbooks fornativeChinesespeakersingrades1to6(Yang,1970
neural networks and linear regression on this corpus showed their comparable effectivenessinreadabilityestimation.Pang(2006)modelledChi
as a support vector regression problem. He compared his approach with linear regression,andfoundthatsupportvectorregressionoutperfor
under a cross-validation setting. Pang further showed the utility of his approach in modelling readability of Chinese web pages. Lee et al. (2012) combined principal component analysis (PCA) with genetic programming (GP) for modelling Chinese readability, and Chen et al. (2013) applied t fidf and lexical chains as features for readability classification using SVM. For the under-resourced Indian languages, especially Bengali, very few such working readabilityindicesexist.DasandRoychoudhury(2000)have
of readability indices on Bengali texts. They considered the density of polysyllabic words,

average sentence length (ASL), number of syllables per 100 words, and a few other surface features in their studies (Das and Roychoudhury, 2004, 2006). Their sample size was very small (seven passages) and the dataset was created only from literary novels

More recently, Sinha et al. (2012) presented readability models for Hindi and Bengali. Their dataset consisted of 16 human-annotated passages. They incorporated structural features such as average sentence length (ASL), average word length (AWL), average number of syllables per word (ASW), number of polysyllabic words (PSW), number of polysyllabic words per 30 sentences (PSW30), and number of juktakkhors (JUK). They identified AWL, PSW, JUK and PSW30 as the four most important features that affect readability in Hindi and Bengali. They further used machine learning approach (Sinha andIslam

et al. (2012) proposed three broad categories of features; lexical features, entropy-based features and Kullback-Leibler [KL] Divergencebased features. Combining all three categories gave the best results (in terms of Fscore) in Bengali readability classification on a corpus of school textbooks. Islam and others extended their work in Islam et al. (2014) where 18 carefully curated features were used to achieve state-of-the-art results (86.46

While all the above studies are very important, we feel that more research needs to be done in Bengali readability analysis. We have chosen regression modelling to construct our readability scores, which is a more fine-grained modelling technique than Islam et al.'s(2012, 2014) multi-class classification. Regressioncanaffordacontinuousresponseva classclassification can only afford k discrete labels. In readability modelling, therefore, we feelthattheregressionsettingisamore-

appropriatechoicethantheclassification setting. Whereas the former yields a continuous scale of values on which a document can be placed, discrete binning of the latter seems to us to be less satisfactory. Furthermore, we had seven labels for annotation (cf. Section 3.1), which is greater than the four labels used by Sinha et al. (2012). A continuous responsevariablewascreatedbytakingthemeanofthesesevendiscreteratings. While all the above studies are very important and insightful, none of them explicitly performed an inter-rater agreement study. An inter-rater agreement studyisveryimportantwhenwetalkaboutreadabilityassessment.Further,none ofthesestudiesmadeavaila annotatedgoldstandarddatasets, therebystymieingfurtherresearch.Weattempttobridgethesegapsinourwork

3.Problem Description

Bengali is an Indic language spoken by native speakers in Bangladesh and the east-

ern part of India comprising West Bengal, Tripura and Assam. It is written in the Bengali script. With about 220 million native and about 30 million non-native speakers, Bengali is one of the most widely spoken languages, ranked seventh in language usage among all languages in the world ('Language Difficulty Ranking' [n.d.]). Three countries – India, Bangladesh and Sri Lanka – have their national anthems written in Bengali. Bengali has been classified as a Category IV language in terms of learning difficulty for somebody whose first language is English (Summary by Language Size [n.d.]). It shows diglossia, with one form reserved for formal settings and literature, and the other for everyday usage. Furthermore, spoken Bengali has several dialects. The principal dialect uses two forms, a written form, sadhu bhasha (formal) and a spoken form, chalit bhasha (colloquial).

Bengali is a highly inflected language, and has special features such as 'juktakkhor' (compound characters), 'sandhi' (phonetic ligature), and 'samas' (word compounding), thereby making it relatively difficult for non-natives to read and understand Bengali text. In this work we propose a model for measuring readability for the Bengali language incorporating language-specific features such as morphological variants of words, semantic variations, presence of stop words, word senses, etc.

3.1. Dataset

Bengali being an under-resourced language, there is a noticeable dearth of standardized corpora and language processing tools. We collected 30 passages with the help of the Society for Natural Language Technology Research (SNLTR). We considered the works of four famous Bengali authors:

Bankim Chandra Chattopadhyay (1838–1894), Bibhutibhushan Bandyopadhyay (1894–1950), Rabindranath Tagore (1861–1941) and Sarat Chandra Chattopadhyay (1876–1938). The works of Bibhutibhusan Bandyopadhyay are yet to come online, so we manually typed in the text. These four authors have distinct styles of writing, in different genres. Table 1 gives overall statistics of our dataset, categorized by children's literature and adult literature,1 and also by the two diglossic variants of Bengali language. We sampled 30 random passages from the works of the 4 litterat were carefully selected to ensure that they cover all genres – from children's texts to adult literature, from novels to short stories to articles. Sampled passag given to seven annotators (three males, four females) for readability judgment. Annotators were all native Bengali speakers of the same academic background and same age-group (between 30 and 35 years) and all of them were avid reade of Bengali literature. Annotators were required to rate the pass

the ease of reading the text and comprehending it, on a 7-point Likert Scale (Likert, 1932). The rating scale was:[**?**]

Table 1.Children's-adult literature and sadhu bhasa chalid bhasha text count

Author	kids	adult	total
Rabindra Nath Tagore	4	12	4
Bankim Chandra Chattopadhyay	0	6	6
Bibhuti Bhusan Bandopadhyay	1	2	3

- 1 - Very easy to read
- 2 – Easy to read
- 3 – Somewhat easy to read
- 4 – In-between
- 5 – Somewhat difficult to read

This rating scale reflects the fact that readability is not a binary/ternary variable; it is an ordinal variable. We further collected the data on whether the annotators were avid readers of Bengali or

not. Each annotator rated every passage. It is worth mentioning that readability annotation in Bengali is challenging because passages written in sadhu bhasha tend to be harder to read than those written in cholit bhasha. Since our dataset contains both sadhu bhasha and cholit bhasha, maintaining consistency in readability rating becomes a big issue.

4. Inter-Annotator Agreement

Beforeproceedingtomodelgenerationastudyofinter-annotator agreementbetweentheraterswas-done.Onthebasisofratingsgivenbytheseven annotators, we studied whether they agreed among each other. We used Spearman's rank correlation[2] coefficient as shown in Equation and Cohen's kappa as shown in Equation . Results indicate that there is[2] moderate to fair agreement among the seven annotators (Phani, Lahiri, and

Biswas, which in turn indicates that [3]. human annotators agreed pretty well with respect to Bengali readability scoring. Mean ratings were then used for readability modelling [4].

$$(R) = 1 - (6 * \Sigma\, d^2)/(n^3 - n)$$

where,
R is the Spearman's Rank correlation coefficient,
d is the difference in ranks, and
n is the number of annotators.

References

1.hadiuzzaman ljhj khhkj khkj.

2. bangladesh.

3. Baust

4. Buet