Text Readability Classification of Textbooks of a Low-Resource Language

Zahurul Islam, Alexander Mehler and Rashedur Rahman
AG Texttechnology
Institut für Informatik
Goethe-Universität Frankfurt
zahurul,mehler@em.uni-frankfurt.de, kamol.sustcse@gmail.com

Abstract

There are many languages considered to be low-density languages, either because the population speaking the language is not very large, or because insufficient digitized text material is available in the language even though millions of people speak the language. Bangla is one of the latter ones. Readability classification is an important Natural Language Processing (NLP) application that can be used to judge the quality of documents and assist writers to locate possible problems. This paper presents a readability classifier of Bangla textbook documents based on information-theoretic and lexical features. The features proposed in this paper result in an *F-score* that is 50% higher than that for traditional readability formulas.

1 Introduction

The readability of a text relates to how easily human readers can process and understand a text as the writer of the text intended. There are many text related factors that influence the readability of a text. These factors include very simple features such as type face, font size, text vocabulary as well as complex features like grammatical conciseness, clarity, underlying semantics and lack of ambiguity.

Nowadays, teachers, journalists, editors and other professionals who create text for a specific audience routinely check the readability of their text. Readability classification, then, is the task of mapping text onto a scale of readability levels. We explore the task of automatically classifying documents based on their different readability levels. As input, this function operates on various statistics relating to lexical and other text features.

Automatic readability classification can be useful for many Natural Language Processing (NLP) applications. Automatic essay grading can benefit from readability classification as a guide to how good an essay actually is. Similarly, search engines can use a readability classifier to rank its generated search results. Automatically generated documents, for example documents generated by text summarization systems or machine translation systems, tend to be error-prone and less readable. In this case, a readability classification system can be used to filter out documents that are less readable. The system can also be used to evaluate machine translation output. A document of higher readability tends to be better than a document that belongs to a lower readability class.

Research in the field of readability classification started in 1920. English is the dominating language in this field although much research has been done for other languages like German, French, Chinese and so on. These languages are considered as high-density languages as many language resources and tools are available for them. However, many languages are considered to be low-density languages, either because the population speaking the language is not very large or because insufficient digitized text material is available for the language even though it is spoken by millions of people. Bangla is such a language. Bangla, an Indo-Aryan language, is spoken in Southeast Asia, specifically in present day Bangladesh and the Indian state of West Bengal.

26th Pacific Asia Conference on Language,Information and Computation pages 545–553

With nearly 230 million speakers, Bangla is one of the largest spoken languages in the world, but only a very small number of linguistic tools and resources are available for it. For instance, there is no morphological analyzer, POS tagger or syntax parser available for Bangla.

To create a supervised readability classification, it is important to use a corpus that is already classified for the different levels of readers. In this work, the corpus is collected from textbooks that are used in primary and middle school in Bangladesh. The collected documents are classified according to their readability. So the extracted corpus is ideal for a readability classification task.

In this paper, we present a readability classification based on information-theoretic and lexical features. We evaluate this classifier in comparison with traditional readability formulas that, even though they were proposed in the early stages of readability classification research, are still widely used.

The paper is organized as follows: Section 2 discusses related work followed by an introduction of the corpus in Section 3. The features used for classification are described in Section 4, and our experiments in Section 5 are followed by a discussion in Section 6. Finally, we present our conclusions in Section 7.

2 Related Work

There is no standard approach to measuring text quality. According to Mullan (2008), a good readable English sentence should contain 14 to 22 words. He also stated that if the average sentence length is more than 22 words then the content is not clear. If the average sentence length is shorter than 14 then it is probable that the presentation of ideas is discontinuous.

Much work was done previously in this field and many different types of features were used. We summarize the related research grouped by type:

Lexical Features: In the early stage of readability research fairly simple features were used due to the lack of linguistic resources and computational power. *Average Sentence length* (ASL) is one of them. The ASL can be used as a measure of grammatical complexity assuming that a longer sentence has a more complex grammatical structure than a shorter one. Dale and Chall (1948; 1995) showed that reading difficulty is a linear function of the ASL of the percentage of rare words. They listed $3,000$ commonly known words for the 4^{th} grade.

Gunning (1952) also considered the numbers of sentences and complex words to measure text readability. The formula uses similar lexical features as (Dale and Chall, 1948; Dale and Chall, 1995) with different constants. The Flesch-Kincaid readability index (Kincaid et al., 1975) considers the average number of words per sentence and the average number of syllables per words. They proposed two different formulas, one for measuring how easy a text is to read and the other one for measuring grading level. Senter and Smith (1967) also designed a readability index for the *US Air force* that uses the average number of characters in a word and the average number of words in a sentence. Many of the other readability formulas are summarized in (Dubay, 2004).

English has a long history of readability research, but there is very little previous research in Bangla text readability. Das and Roychudhury (2004; 2006) show that readability formulas proposed by (Kincaid et al., 1975) and (Gunning, 1952) work well for Bangla text. The readability formulas were tested semi-automatically on seven documents, mostly novels. Obviously this data set is small.

Petersen & Ostendorf (2009) and Feng et al. (2009) show that these traditional methods have significant drawbacks. Longer sentences are not always syntactically complex and the syllable number of a single word does not correlate with its difficulty. With recent advancements of NLP tools, a new class of text features is now available.

Language Model Based Features: Collins-Thompson and Callan (2004), Schwarm and Ostendorf (2005), Alusio et al. (2010), Kate et al. (2010) and Eickhoff et al. (2011) use statistical language models to classify texts for their readability. They show that trigrams are more informative than bigram and unigram mod-

els. Combining information from statistical language models with other features using Support Vector Machines (SVM) outperform traditional readability measures. Pitler and Nenkova (2008) also used a unigram language model and found that this feature is a strong predictor of readability.

POS-based Features: Parts of Speech (POS)-based grammatical features were shown to be useful in readability classification (Pitler and Nenkova, 2008; Feng et al., 2009; Aluisio et al., 2010; Feng et al., 2010). In the experiment of (Feng et al., 2010), these features outperform language-model-based features.

Syntax-based Features: Text readability is affected by syntactic constructions (Pitler and Nenkova, 2008; Barzilay and Lapata, 2008; Heilman et al., 2007; Heilman et al., 2008). In this line of research, Barzilay and Lapata (2008) show, for example, that multiple noun phrases in a single sentence require the reader to remember more items.

Semantic-based Features: On the semantic level, a paragraph that refers to many entities burdens the reader since he has to keep track of these entities, their semantic representations and how these entities are related. Texts that refer to many entities are extremely difficult to understand for people with intellectual disabilities (Feng et al., 2009). Noemie and Huenerfauth (2009) show how working memory limits the semantic encoding of new information by readers.

Researchers also experimented with semantic features like *lexical chains*, *discourse relations* and *entity grids* (Feng et al., 2010; Barzilay and Lapata, 2008). It has been shown that these features are useful for readability classification.

In this paper, we do not compare our work with any previous work that explores linguistic features. Due to the unavailability of a Bangla syllable identification system, we could not compare our work with readability formulas that use syllable information. We will only compare our proposed features with a baseline system that uses three traditional readability formulas proposed by Gunning (1952) , Dale and Chall (1948; 1995) and Senter and Smith (1967). These traditional formulas are widely used in many readability classification tools.

3 Corpus Extraction

The government agency *National Curriculum and Textbook Board, Bangladesh*[1] makes available textbooks that are used in public schools in Bangladesh. The textbooks cover many different subjects, including Bangla Literature, Social Science, General Science and Religious Studies. These textbooks are for students from grade one to grade ten. All of the textbooks are in Portable Document Format (PDF). Some of them are made by scanning textbooks and some of them are converted from typed text. There is a Bangla OCR (Hasnat et al., 2007) available but it is unable to extract text from the scanned PDF books. Therefore, we only considered textbooks that were converted to PDF from typed text. The *Apache PDFBox*[2] is used to extract text from PDFs. Note that 24 textbooks were extracted from class *two* to class *eight*. After text extraction, it was observed that the text was not written in Unicode Bangla. A non-standard Bangla input method called *Bijoy* is used to type the textbooks. This is an ASCII based Bangla input method that was widely used in the 1990s. The next challenge was to convert non-standard text to Bangla Unicode.

The selected text books were written using a font called *SutonnyMJ* that has many different versions, all of which differ slightly in terms of the code point of some *consonant conjuncts*. The freely available open source CRBLPConverter[3] is used to convert these non-standard Bangla texts to Unicode. To cope with the font of the text, the CRBLPConverter required some slight modifications. Text books not only contain descriptive texts but also contain questions, poems, religious hymns, texts from other languages (e.g., Arabic, Pali) and transcription of Arabic texts (e.g., Surah). Manual work was involved to clean these non-descriptive texts and extract each chapter as a document. Class *two* contains only one

[1] http://nctb.gov.bd/book.php
[2] http://pdfbox.apache.org/
[3] http://crblp.bracu.ac.bd/converter.php

Classes	Documents	Avg. Document Length	Avg. Sentence Length	Avg. Word Length
three	123	65.21	8.07	4.31
four	88	126.25	8.63	4.37
five	43	196.72	9.34	4.41
six	62	130.13	11.53	4.85

Table 1: The Bangla Readability Corpus

textbook and class *six*, *seven* and *eight* contain two textbooks each. To avoid a data sparseness problem, class *two* is merged with class *three* and class *seven* and *eight* are merged with class *six*. Each document is tokenized using a slightly modified version of the tokenizer which is freely available in[4]. Table 1 shows the details of the corpus. The *Average Document Length* shows the average number of sentences per document. The *Average Sentence Length* represents the average number of words in a sentence and *Average Word Length* displays the average number of characters in a word.

It should be noted that 80% of the corpus is used for training and the remaining 20% is used as a test set.

4 Features

4.1 Lexical Features

In this paper, we compare a lexical and information-theoretic classifier of text readability with a classifier based on traditional readability formulas. The literature explores some of the linguistic indicators of readability. This includes the avg. sentence length, avg. word length and the avg. number of difficult words (of more than 9 letters). We develop a classifier of text readability based on lexical and information-theoretic features. We first describe lexical features used by this classifier.

The *Average Sentence Length* is a quantitative measure of syntactic complexity. In most cases, the syntax of a longer sentence is more difficult than the syntax of a shorter sentence. However, children of lower grade levels are not aware of syntax. In any event, a longer sentence contains more entities and children have to remember all of these entities in order to understand the sentence, which makes a longer sentence more difficult for them. As an example, Table 1 shows that the *Average Sentence Length* rises in the text of higher readability classes. The *Average Word Length* is another lexical feature that is useful for readability classification. A longer word carries some difficulties for children at a lower grade level. For example: the word *biodegradable* will be harder to pronounce, spell and understand for children at a lower grade level. This characteristic is reflected in our readability corpus that is shown in Table 1. The *Average Word Length* will be more useful for agglutinative languages such as German, which allows concatenation of morphemes to build longer words.

The *Average Number of Complex Words* feature is related to the *Average Word Length*. The average length of English written words is 5.5 (Nádas, 1984). Table 1 shows that the average word length in our corpus is below 5. Dash (2005) showed that the average word length in the CIIL[5] corpus is 5.12. Majumder et al. (2006) claimed that the average word length in a Bangla news corpus is 8.62. They have mentioned that the average length is higher due to the presence of many hyphenated words in the news corpus. In this work, any word that contains 10 or more characters is considered a complex word. A complex word will be harder to read for children at a lower grade level. The type token ratio (TTR), which indicates the lexical density of text, has been considered as a readability feature too. Low lexical densities involve a great deal of repetition.

The term *Hapax Legomena* is widely used in linguistics referring to words which occur only once within a context or document. These are mostly content words. Kornai (2008) showed that 40% to 60% of the words in larger corpora are *Hapax Legomena*. Documents with more *Hapax Legomena* generally will contain more information. In terms of text readability, the difficulty level will be higher.

4.2 Entropy Based Features

Recently, researchers have independently made the suggestion that the entropy rate plays a role in human communication in general (Genzel and Charniak, 2002; Levy and Jaeger, 2007). The rate of

[4]http://statmt.org/wmt09/scripts.tgz

[5]http://www.elda.org/catalogue/en/text/W0037.html

information transmission per second in a human speech conversation is roughly constant, that is, transmitting a constant number of bits per second or maintaining a constant entropy rate.

Since the most efficient way to send information through a noisy channel is at a constant rate, Plotkin and Nowak (2000) have shown that this principle could be viewed as biological evidence of how human language processing evolved. Communication through a text should satisfy this principle. That is, each sentence of a text, for example, conveys roughly the same amount of information. In order to utilize this information-theoretical notion, we start from random variables and consider their entropy as indicators of readability.

Shannon (1948) introduced entropy as a measure of information. Entropy, the amount of information in a random variable, can be thought of as the average length of the message needed to have an outcome on that variable. The entropy of a random variable X is defined as

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \qquad (1)$$

The more the outcome of X converges towards a uniform distribution, the higher $H(X)$. Our hypothesis is that the higher the entropy, the less readable the text along the feature represented by X. In our experiment, we consider the following random variables: *word probability*, *character probability*, *word length probability* and *word frequency probability* (or frequency spectrum, respectively). Note that there is a correlation between the probability distribution of words and the corresponding distribution of word frequencies. As we use Support Vector Machines (SVM) for classification, these correlations are taken into consideration.

4.3 Kullback-Leibler Divergence-based Features

The *Kullback-Leibler divergence* or *relative entropy* is a non-negative measure of the divergence of two probability distributions. Let $p(x)$ and $q(x)$ be two probability distributions of a random variable X. The relative entropy of these distributions is defined as:

$$D(p||q) = \sum_{i=1}^{n} p(x_i) \log \frac{p(x_i)}{q(x_i)} \qquad (2)$$

$D(p||q)$ is an asymmetric measure that considers the number of additional bits needed to encode p, when using an optimal code for q instead of an optimal code for p. In other words: $D(p||q)$ measures how much one probability distribution is different from another distribution. More specifically, if the probability distribution of a document p is closer to q than to q' then the document has a smaller distance to q. The document belongs to the category corresponding to q.

In order to apply this method in our framework we start from a training corpus where for each target class and each random variable under consideration we compute the distribution $q(x)$. This gives a reference distribution such that for a text T whose class membership is unknown, we can compute the distribution $p(x)$ only for T in order to ask how much information we get about $p(x)$ when knowing $q(x)$. Since $q(x)$ is computed for each of the four target classes (see Table 1), this gives for any random variable X four features of *relative entropy*.

5 Experiments and Results

5.1 Baseline System

To measure accuracy of our proposed features, a baseline system is implemented that uses three traditional readability formulas, such as: *Gunning fog readability index* (Gunning, 1952), *Dale–Chall readability formula* (Dale and Chall, 1948; Dale and Chall, 1995) and *Automated readability index* (Senter and Smith, 1967). There are more traditional formulas available that use syllable information, these are not considered for this task due to unavailability of a Bangla syllable identification system. The *Gunning fog readability index* and *Dale–Chall readability formula* both use complex or difficult words. The definition of these words varies slightly. Gunning (1952) defines a complex word as a word that contains more than three syllables and Dale and Chall (1948; 1995) introduce 3000 familiar words. Any word not in the list of 3000 words is considered difficult. For this work, both types of words are defined in the same way, described in section 4.1. We consider any word that has 10 or more letters as a difficult or complex word. Table 2 shows the evaluation of the baseline system. The evaluation shows that these features do not perform well. Among

Features	Accuracy	F-Score
Gunning fog readability index	48.3%	36.5%
Dale–Chall readability formula	48.3%	45.0%
Automated readability index	51.6%	46.2%
All together	53.3%	49.6%

Table 2: Evaluation of baseline system with 3 traditional readability formulas.

Features	Accuracy	F-Score
Average sentence length	51.6%	47.3%
Type token ratio	41.6%	30.6%
Avg. word length	50.3%	46.9%
Avg. number of complex Words	46.6%	34.2%
Hapax legomena	40.0%	28.3%
All together	60.0%	56.5%

Table 3: Evaluation of lexical features.

these formulas, *Automated readability index* is the highest performing formula. Das and Roychudhury (2004; 2006) showed that these traditional features nonetheless work well for Bangla novels. Note that we have used the SMO (Platt, 1998; Keerthi et al., 2001) classifier model in WEKA (Hall et al., 2009) together with the Pearson VII function-based universal kernel PUK (Üstün et al., 2006).

5.2 System with Lexical Features

Lexical features use the same kind of surface features as the traditional readability formulas used in the baseline system (see: Section 5.1). Table 1 shows that the *average sentence length* and difficulty levels are proportional. That means that sentence length increases for higher readability classes. *Average word length* exhibits the same characteristics. These characteristics are reflected in the experiment. These two are the best performing features among all of the lexical features. Table 3 shows the evaluation of the system that uses only lexical features. Although the individual accuracy of some of these features is similar to the traditional formulas, the combination of all lexical features outperforms the baseline system.

5.3 System with Entropy Based Features

As noted earlier, entropy measures the amount of information in a document. The entropy rate is constant in human communication (see Section: 4.2).

Features	Accuracy	F-Score
Word probability	53.3%	49.3%
Character probability	48.3%	35.4%
Word length probability	50.0%	36.9%
Word frequency probability	43.3%	32.4%
Character frequency probability	53.3%	47.7%
Entropy features	61.6%	59.8%
Lexical + entropy features	73.3%	72.1%

Table 4: Evaluation of entropy based features.

The documents in this work are assumed to be a medium of communication between writers and readers. Conversely, information flow of a very readable document will differ from that of a less readable document. So, the constants for the corresponding entropy rates of the different readability classes will differ. As a single feature, these entropy based features perform similarly to lexical features. But, collectively this is the best performing feature set. Among all similar features the random variable with *Word Probability* works better than others. Table 4 shows the results of these features. Adding *lexical* features with *entropy* based features improves *accuracy* and *F-score* substantially.

5.4 System with Kullback-Leibler Divergence-based Features

Relative entropy-based features represent the distance between the test document and target classes. The target class with the lowest distance will be the class of the test document. Five different types of random variables are used in this work (see Section 4.3). The random variable based on *character probabilities* is the best performing individual feature among all features used in this work. However, this feature set performs worse than the *lexical* and *entropy* based features set. The evaluation is shown in Table 5. The combination of all, i.e., *lexical*, *entropy* and *relative entropy* based features, gives the best result, namely *accuracy* of 75% and *F-score* of 74.1%.

6 Discussion

Das and Roychudhury (2004; 2006) found that traditional readability formulas are useful for Bangla readability classification. However, the experimental results in this paper show that these formulas are

Features	Accuracy	F-Score
4 Word probabilities	50.0%	50.2%
4 Character probabilities	61.6%	61.1%
4 Word length probabilities	48.3%	46.5%
4 Word frequency probabilities	50.0%	45.6%
4 Character frequency probabilities	43.3%	34.2%
20 Relative entropy based features	56.6%	54.0%
Entropy + relative entropy features	68.3%	65.9%
Lexical + entropy + relative entropy based features	75.0%	74.1%

Table 5: Evaluation of Kullback-Leibler Divergence-based Features.

not useful for studies like the one presented here. This is probably due to the fact that these formulas were specially designed for English. One reason for the poor performance is that Bangla script is a syllabic script that has glyphs representing clusters and ligatures.

It also has to be noted that Bangla is an inflectional language, so that the average word length can be longer than that of many other languages.

The lexical features that are assumed to be good indicators of text difficulty did indeed perform well in classification. The respective feature set performs better than the baseline system. *Average sentence length* and *Average word length* do not perform well, as reflected in Table 1. That shows that the average word and sentence lengths are longer in higher readability classes than in lower readability classes.

As an individual feature, each *entropy* based feature performs similarly to other features. However, the combination of the *entropy* based features are the best performing features among all. The classification performance even increases when *entropy* based features are combined with *lexical* features.

Among all *relative entropy* based features, the random variable based on *character probabilities* performs best. This feature performs better than the baseline system. But the performance drops when this feature is added to other *relative entropy* based features. Although the *relative entropy* based feature set performs better than the baseline system, the *lexical* and *entropy* based feature set performs even better. The performance surpasses the baseline system by 50% when *lexical*, *entropy* based and *relative entropy* based features are combined.

7 Conclusion

In this paper, we have presented features for text readability classification of a low resource language. Altogether we have proposed 30 quantitative features. Twenty-five of them are information-theoretic based features. These features do not require any kind of linguistic processing. Recent advances in NLP tools argue that linguistic features are useful for readability classification. However, our experimental results show that lexical and information-theoretic features perform very well. There are many languages in the *Asia Pacific* region that are still considered as low resource languages. These features can be used for readability classification of these languages. As a future work, we plan to explore many other information-theoretic features like *mutual information*, *point wise mutual information* and *motifs*.

8 Acknowledgements

We would like to thank Mr. Munir Hasan from the Bangladesh Open Source Network (BdOSN) and Mr. Murshid Aktar from the National Curriculum & Textbook Board Authority, Bangladesh for their help on corpus collection. We would also like to thank Andy Lücking, Paul Warner and Armin Hoenen for their fruitful suggestions and comments. Finally, we thank three anonymous reviewers. This work is funded by the LOEWE Digital-Humanities project in the Goethe-Universität Frankfurt.

References

Ra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *NAACL-HLT 2010: The 5th Workshop on Innovative Use of NLP for Building Educational Applications*.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 21(3):285–301.

Kevyn Collins-Thompson and James P Callan. 2004. A language modeling approach to predicting reading difficulty. In *HLT-NAACL*.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–20+28.

Edgar Dale and Jeanne S. Chall. 1995. *Readability Revisited: The New Dale-Chall Readability formula*. Brookline Books.

Sreerupa Das and Rajkumar Roychoudhury. 2004. Testing level of readability in bangla novels of bankim chandra chattopodhay w.r.t the density of polysyllabic words. *Indian Journal of Linguistics*, 22:41–51.

Sreerupa Das and Rajkumar Roychoudhury. 2006. Readabilit modeling and comparison of one and two parametric fit: a case study in bangla. *Journal of Quantative Linguistics*, 13(1).

Niladri Sekher Dash. 2005. *Corpus Linguistics and Language Technology: With Reference to Indian Languages*. New Delhi: Mittal Publications.

William H. Dubay. 2004. The principles of readability. *Costa Mesa, CA: Impact Information*.

Carsten Eickhoff, Pavel Serdyukov, and Arjen P. de Vries. 2011. A combined topical/non-topical approach to identifying web sites for children. In *Proceedings of the fourth ACM international conference on Web search and data mining*.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL*.

Lijun Feng, Martin Janche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *The 23rd International Conference on Computational Linguistics (COLING)*.

Dimitry Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40st Meeting of the Association for Computational Linguistics (ACL 2002)*.

Robert Gunning. 1952. *The Technique of clear writing*. McGraw-Hill; Fourh Printing Edition.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations*, 11(1):10–18.

Md. Abul Hasnat, S M Murtoza Habib, and Mumit Khan. 2007. A high performance domain specific ocr for bangla script. In *International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CISSE)*.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2007. Combining lexical and grammatical features to improve readavility measures for first and second language text. In *Proceedings of the Human Language Technology Conference*.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 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 (EANL)*.

Rohit J. Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond J. Mooney, Salim Roukos, and Chris Welty. 2010. Learning to predict readability using diverse linguistic features. In *23rd International Conference on Computational Linguistics (COLING 2010)*.

S.S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy. 2001. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*, 13(3):637–649.

J. Kincaid, R. Fishburne, R. Rodegers, and B. Chissom. 1975. Derivation of new readability formulas for Navy enlisted personnel. Technical report, US Navy, Branch Report 8-75, Cheif of Naval Traning, Millington, TN.

András Kornai. 2008. *Mathematical Linguistics*. Springer.

Roger Levy and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, pages 849–856.

Khair Md. Yeasir Arafat Majumder, Md. Zahurul Islam, and Mumit Khan. 2006. Analysis and observations from a bangla news corpus. In *9th International Conference on Computer and Information Technology (ICCIT 2006)*.

W.M.A Mullan. 2008. Dairy science and food technology improving your writing using a readability calculator.

A. Nádas. 1984. Estimation of probabilities in the language model of the ibm speech recognition system. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(4):859–861.

Sarah E. Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assesment. *Computer Speech and Language*, 23(1):89–106.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

John C. Platt. 1998. *Fast training of support vector machines using sequential minimal optimization*. MIT Press.

Joshua B. Plotkin and Martik A. Nowak. 2000. Language evolution and information theory. *Journal of Theoretical Biology*, 205(1):147–159.

Sarah E. Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *the Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics(ACL 2005)*.

R.J. Senter and E. A. Smith. 1967. Automated readability index. Technical report, Wright-Patterson Air Force Base.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(1):379–423.

B. Üstün, W.J. Melssen, and L.M.C. Buydens. 2006. Facilitating the application of support vector regression by using a universal Pearson VII function based kernel. *Chemometrics and Intelligent Laboratory Systems*, 81(1):29–40.