Multilingual Text Classification using Information-Theoretic Features

Dissertation
for attaining the PhD degree
of Natural Sciences

submitted to the Faculty
of the Johann Wolfgang Goethe University
in Frankfurt am Main

Department of Computer Science
Goethe University Frankfurt

by
© *Mohammad Zahurul Islam*
from Dinajpur

Frankfurt, 2014

accepted by the Faculty _____ of the

Johann Wolfgang Goethe University as a dissertation.

Dean:

Expert assessor:

Date of disputation

Acknowledgments

I would like to express my gratitude to all the people who supported and accompanied me during the progress of this thesis. Special thanks to my family, especially my parents, my wife, my daughter, my sisters, my brother and brothers in law, for being part of every bit of my life. Their endless motivation, constant mental support and unconditional love have been influential in whatever I have achieved so far.

I would like to thank my first supervisor Professor Dr. Alexander Mehler. I would always be grateful for the insightful guidance he has provided me in order to cross the final hurdle. He has given me the idea of exploring *information-theoretic* for text classification. He has listened to all the problems I faced during this thesis and showed me the way to overcome them. Thank you Professor Mehler.

I would also like to thank my second supervisor Prof. Dr. Visvanathan Ramesh. He has given many valuable comments. The thing I like best about him, whenever he see me he always asks me about the progress of my thesis.

I would also like to thank my colleagues at the Text Technology Working Group, Dr. Andy Lücking, Rüdiger Gleim, Armin Hoenen, Dr. Tim vor der Brück, Paul Warner, Guissepe Abrami and Rashedur Rahman for their support during my thesis. Special thanks to Dr. Andy Lücking for some useful comments and help with different issues of LaTeX. Special thanks also to Paul Warner for checking the English of my thesis. I also would like to thank Prof. Dr. Walt Detmar Meurers and Sowmya Vajjala for sharing their GEO/GEOLino corpus.

I would also like to thank Professor Dr. Mumit Khan from the Center for Research on Bangla Language Processing (CRBLP), BRAC University, Bangladesh. He inspired me to join in the field of *natural language processing* (NLP). He advised and helped me to come in Europe to do further study in the field of NLP.

Finally, all my friends in Bangladesh and here in Europe deserve special thanks for their ongoing support. In the end, I would like to thank almighty God for giving me the strength to achieve whatever I have achieved so far.

Abstract

The number of multilingual texts in the World Wide Web (WWW) is increasing dramatically and a multilingual economic zone like the European Union (EU) requires the availability of multilingual Natural Language Processing (NLP) tools. Due to a rapid development of NLP tools, many lexical, syntactic, semantic and other linguistic features have been used in different NLP applications. However, there are some situations where these features can not be used due the application type or unavailability of NLP resources for some of the languages. That is why an application that is intended to handle multilingual texts must have features that are not dependent on a particular language and specific linguistic tools. In this thesis, we will focus on two such applications: *text readability* and *source and translation* classification.

In this thesis, we provide 18 features that are not only suitable for both applications, but are also language and linguistic tools independent. In order to build a readability classifier, we use texts from three different languages: *English*, *German* and *Bangla*. Our proposed features achieve a classification accuracy that is comparable with a classifier using 40 linguistic features. The readability classifier achieves a classification *F-score* of 74.21% on the English Wikipedia corpus, an *F-score* of 75.47% on the English textbook corpus, an *F-score* of 86.46% on the Bangla textbook corpus and an *F-score* of 86.26% on the German *GEO/GEOLino* corpus.

We used more than two million sentence pairs from 21 European languages in order to build the source and translation classifier. The classifier using the same eighteen features achieves a classification accuracy of 86.63%. We also used the same features to build a classifier that classifies translated texts based on their origin. The classifier achieves classification accuracy of 75% for texts from 10 European languages. In this thesis, we also provide four different corpora, three for text readability analysis and one for corpus based translation studies.

Contents

List of Tables

List of Figures

1 Introduction

1.1 Motivation

A language is a medium of human communication, a way to encode and transmit information between humans. All natural languages follow the laws of natural evolution (Nowak, Komarova, and Niyogi 2002; Lieberman, Michel, Jackson, Tang, and Nowak 2007). Although the encoding of information is highly complex, it can be projected as a sequence of words. These sequences form sentences and collections of sentences become a text. A text is therefore a medium of communication between authors and their intended readers. Similarly a translated text is also a communication medium, but here a translator is also involved.

A text can be monolingual or multilingual. Nowadays, most of the documents that are generated in Europe are multilingual (Carson 2003). Also, a multilingual economic zone like the European Union (EU) imposes the necessity of multilingual Natural Language Processing (NLP) tools in order to cope with the giant translation effort of governmental documents. It has been seen that many NLP tools are available for different European languages. Due to this rapid development of NLP tools, many lexical, syntactic, semantic and other linguistics features have been used in NLP applications. However, there are some situations where these features can not be used due the application type or to the unavailability of NLP resources for some languages. In this thesis, we will focus on two such applications: *text readability* and *source and translation* classification. Both applications can be monolingual or multilingual. In this thesis, both of the applications are considered as multilingual.

In the field of text readability research, English is the dominating language. Since the early twentieth century, different readability formulas (Gunning 1952; Dale and Chall 1948; Dale and Chall 1995; Kincaid, Fishburne, Rodegers, and Chissom 1975; McLaughlin 1969) have been proposed for English texts. Nowadays, these formulas are called traditional readability formulas. These formulas use very rudimentary features such as average *sentence length* (SL), average *word length* (WL) and number of *syllables* in a word. When it comes to readability analysis of text from other languages, only very few tools are available even for many well-resourced[1] European language such as German (Hancke, Vajjala, and Meurers 2012), French (François and Fairon 2012) and Italian (Dell'Orletta, Montemagni, and Venturi 2011).

Many languages are considered to be low-resourced, 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 (Islam, Tiedemann, and Eisele 2010). Most of the low-resourced languages lack a tool for readability analysis. A readability classification for any of these languages will require features that are language-independent and do not require any kind of linguistic pre-processing. Most of the traditional formulas do not require any linguistic pre-processing. That is why these can be used to build a readability classifier for any of these languages.

Currently, many of the *state-of-the-art* text readability tools for English texts use traditional formulas. *The Readability Test Tool*[2] is one of them. The tool measures the readability of a *text* or a *web site* by means of five traditional readability formulas. Using this tool, the English news site of the *British Broadcasting Corporation* (BBC), `www.bbc.co.uk` gets a grade level of 6. The grade level means, the content of the BBC will be easily understandable by a person of 11 to 12 years of age. The German news site *Frankfurter Rundschau* gets a grade level of 10. News from this site will be understandable by a person of 15 or 16 years of age. However, a Bangla news site `www.prothom-alo.com` gets a grade level of 0. That means the site's content will be easily understandable by a child of 5 to 6 years of age. We get similar results using another tool called *Readability-score.com*[3]. Both of the tools show that these traditional formulas work well for English and work reasonably well for German. However, the tools that use traditional formulas do not work for a low resource language like Bangla. The same observation would be noticed for any language that has different linguistic properties than English. Petersen and Ostendorf (2009) and Feng, Elhadad, and Huenerfauth (2009) show that the traditional formulas have significant drawbacks even for English. That is why it is a challenging task to measure reading difficulty of texts from a low-resourced language. The first challenge is that the choice of tools for exploring features for any of these languages is limited. Only features that do not require any kind of linguistic pre-processing are candidates. The second challenge is to achieve a reasonable

[1] a language is considered as being well-resourced if there are quality linguistic resources and tools and resources available for that language

[2] The web link of the tool: `http://read-able.com`

[3] The web link of the tool: `http://www.readability-score.com/`.

readability measurement accuracy. By considering these challenges, we have proposed some information-theoretic and lexical features that are language and linguistic tools independent and that achieve reasonable readability measurement accuracy for texts from three different languages.

The second application we focus on in this thesis are *corpus based translation studies* (CBTS). The studies are relatively new and have gained popularity among researchers from the field of *computational linguistics* (CL) and *translation studies* (TS). There are many texts available that are written by a *native author*[4] while many of them are translated from different languages. Translation is one of the oldest text operations. The history of translation goes back many centuries. The Gospels provide a very good example of historical translation. Nowadays, it is sometimes unclear what the original language of these historical documents was. A source and translation classification tool is necessary to be able to identify source texts and translated texts. This classification tool can be used to find the origin of historical translated documents. Recently, in the field of translation studies, translation scholars proposed different properties of translated texts that make them distinguishable from the original or source texts. Translation scholars are limited to using monolingual or bilingual corpora in order to find these properties of original and translated texts. The field also lacks a corpus that contains original and translated texts from a variety of languages including sentences that are properly annotated and aligned.

Both of the applications require corpora of different languages and some useful features. Many corpora are available that can be used to learn the characteristics of multilingual texts automatically in terms of the above mentioned applications. In this thesis, we propose some *information-theoretic* and *lexical* features that are language independent and achieve satisfactory accuracy for both applications. This is explained in the next section.

1.2 Goals and contributions

1.2.1 Thesis goals

Due to the growing power and speed of computers and the availability of linguistic resources, different linguistically-motivated features have been proposed for different NLP applications including text readability classification. That also creates a digital divide

[4] Native speaker of the language of the text

between languages, cultures and countries. There are many resources and applications available for some languages which are spoken by only a few million speakers (e.g., Dutch). There are also many other languages that are considered to be mostly spoken languages, even though they are used by many millions of people (e.g., Bangla). For these languages, there are not many linguistic tools or resources available.

In this thesis, our goal is to provide some features for *text readability* and *source and translation* classifications that are language independent and do not require any kind of linguistic pre-processing. So, the same feature set will be used for both the applications. There are different properties that make a text *easy* or *difficult* to read and other properties (See Section 9.1 in Chapter 9) that make a translated text distinguishable from an original text. However, these properties can be captured by similar features that are proposed in this thesis. Also, in general a translated text is more readable than its original counter part due to a translator's tendency to make the translated text simple and readable.

It is important to show that for readability analysis the proposed features are useful not only for low-resourced languages but for well-resourced languages as well. For a well-resourced language (e.g., English), these features achieve a satisfactory accuracy is that comparable to many state-of-the-art linguistic features. In order to show the usefulness of our proposed features, we need some corpora that can be used to build an application such as a *text readability* classifier and a *source and translation* classifier. It is also our goal to provide corpora to the research community that are useful for both tasks. The following sub-section describes the contributions of this thesis.

1.2.2 Thesis contributions

The main contribution of this thesis is the examination of the impact of *information-theoretic* features. We propose eighteen *lexical* and *information-theoretic* features and show that a classifier using these features can achieve satisfactory classification accuracy for both tasks. The cognitive model of readability (See Section 3.4 in Chapter 3) and translation process (See Section 9.2 in Chapter 9) are considered in order to propose these features. We will build in total six classifiers using proposed features for both tasks and each of the classifiers will achieve a reasonable classification accuracy. This thesis also contributes three corpora for text readability such as the *English Wikipedia corpus*, the *English textbook corpus* and the *Bangla textbook corpus.* In addition to these corpora, this thesis also provides a customized version of the Europarl corpus that is suitable for corpus-based translation studies. The following sub-sections give a brief summary of these resources and proposed features.

1.2.2.1 The Wikipedia corpus

The corpus is compiled by considering the output of the *Wikipedia article feedback tool version* 4[5]. Wikipedia's readers can give feedback based on different qualities of an article. The corpus contains 641 English Wikipedia articles that are divided into four difficulty levels based on scores given by readers. The difficulty labels are: *very easy*, *easy*, *medium* and *difficult*. The corpus consists of 86,105 sentences, 5,945,799 tokens (total words) and 324,730 types (unique words). The corpus is freely available at: `http://www.hucompute.org/ressourcen/81-english-wikipedia-corpus`.

1.2.2.2 The Bangla textbook corpus

The best resource for a corpus compilation for text readability analysis is the textbooks that are used to teach students from different grade levels in schools. The *Bangla readability corpus* is collected from textbooks that have been used in all public schools in Bangladesh. This resource is provided by the *National Curriculum and Textbook Board* (NCTB), Bangladesh. The corpus contains texts from 51 textbooks from grade *two* to grade *ten*. Similar to the *English Wikipedia corpus*, the corpus is divided into four difficulty classes. The corpus consists of 582 documents, 96,625 sentences, 937,735 tokens and 132,927 types. This can be an ideal resource for text readability research by example of a low-resourced language. This corpus is also freely available at: `http://www.hucompute.org/ressourcen/79-bangla-textbook-corpus`.

1.2.2.3 The English textbook corpus

This corpus is compiled from the same source as the Bangla textbook corpus. Generally, these textbooks are English translations of Bangla textbooks. These textbooks are used in Bangladesh for teaching students from grade two to grade ten who want to study in English. The corpus is extracted from 48 textbooks. Extracted documents are divided into four difficulty classes, similarly to the two corpora described above. The corpus contains 519 documents, 95,470 sentences, 1,183,031 tokens and 60,359 types. This corpus is also available freely at: `http://www.hucompute.org/ressourcen/80-english-textbook-corpus`. This resource could also be useful for research on Second Language Acquisition (SLA).

[5] http://en.wikipedia.org/wiki/Wikipedia:Article_Feedback_Tool

1.2.2.4 The Customized Europarl corpus for translation studies

The field of corpus based translation studies lacks a multilingual corpus where original and translated texts are available from different languages. There are some resources that are suitable for the study but require some customization in order to extract original and translated sentences. This thesis provides such a corpus that contains original and translated texts from 21 European languages. The corpus is extracted from the well known corpus *Europarl* (Koehn 2005). The *Europarl* corpus contains some erroneous annotations of speaker names and their spoken language. We have performed extensive studies to filter out those sentences with erroneous annotation, and after this processing is completed, the corpus contains 2,646,765 parallel sentences from 412 language pairs of 21 European languages. This corpus is also available freely at: `http://hucompute.org/ressourcen/56-customized-europarl`.

1.2.2.5 Information-theoretic features

Classical entropy was proposed back in 1948 by Shannon (Shannon 1948). The entropy of a random variable is related to the difficulty of correctly guessing the value the variable. A text is more readable when the next words are easier to guess. The guessing uncertainty of different text properties can be measured by entropy. In this thesis, we provide some entropy-based features that have good predictive power to measure reading difficulty of a text. Joint and conditional entropies of different random variables are also useful for readability classification. We also have provided two features based on information transmission features that use joint and conditional entropies. Along with these information-theoretic features we also used some of the TTR formulas for these kind of tasks for the first time.

1.2.3 Publications

Some parts of this thesis (partially or in full) were published in well known international peer-reviewed conference proceedings in the field of computational linguistics and natural language processing. These papers are published in collaboration with other co-authors. The following sub-sections describe my contributions within these papers.

1.2.3.1 Chapter 3, Chapter 4, Chapter 5, Chapter 6, Chapter 7, Chapter 8

Some parts of Chapter 3, Chapter 4, Chapter 5, Chapter 6, Chapter 7, Chapter 8 were published four conference proceedings. Islam, Mehler, and Rahman (2012) provide a readability corpus of the Bangla language and a classification of the corpus using lexical and information-theoretic features. This paper includes several *Kullback-Leibler divergence-based* features. I was responsible for feature selection, algorithm design, and the implementation of experiments. The corpus was extracted collaboratively with *Rashedur Rahman*. I have written different sections that are related with these tasks.

Islam and Mehler (2013) was a collaboration with *Alexander Mehler*. I was involved in corpus compilation, design, feature selection and implementation of different experiments. I also have written different sections related my tasks.

Islam, Rahman, and Mehler (2014) is a follow-up paper of Islam, Mehler, and Rahman (2012). The paper provides a corpus that contains more texts than Islam, Mehler, and Rahman (2012) and has been cleaned extensively. The corpus cleaning was performed collaboratively with *Rashedur Rahman*. I performed feature selection, experiment design and implementation.

Islam and Rahman (2014) is about evaluating a classifier using our proposed feature to classify texts that are targeted for children. The paper evaluates children's news texts from five different news sites from Bangladesh and India. I was responsible for experiment design, implementation and evaluation. I have also written the respective sections related my tasks.

1.2.3.2 Chapter 9, Chapter 10, Chapter 11, Chapter 12

There are two publications namely Islam and Mehler (2012) and Islam and Hoenen (2013) that are parts of this chapter. Islam and Mehler (2012) provides a customized version of the *Europarl* corpus. I have designed and implemented the customization tool, collect corpus and performed experiments. All of these tasks were supervised by *Alexander Mehler*. I wrote some of the sections of the paper.

Islam and Hoenen (2013) was collaboration with *Armin Hoenen*. I performed corpus extraction, experiment design and implementation. *Armin Hoenen* discussed our findings and provided a linguistic explanation. I wrote the whole paper except for the discussion and conclusion sections, which were written by *Armin Hoenen*.

1.3 Background

1.3.1 Text categorization and support vector machine (SVM)

Text categorization is an activity of labeling natural language texts with thematic categorizes from a predefined set (Sebastiani 2002). Researchers began to look seriously at *text categorization* in the early '60s (Sebastiani 2002). Due to applicative interest and rapid development of computing power, this field become one of the major sub-field in NLP in the early '90s (Sebastiani 2002).

Text categorization has been applied in many different contexts such as *document indexing*, *word sense disambiguation*, *generating hierarchical categories of web resources* and other applications of document organization. In the early '80s researchers built *text categorization* applications using rules. One of the examples is the *CONSTRUE* system (Hayes, Andersen, Nirenburg, and Schmandt 1990). It uses *disjunctive normal form*[6] (DNF). The flaw of this type of application is that the DNF rules must be manually defined by experts. If there is any update of categories or domains, an expert must start from scratch and define new DNF rules.

On the other hand, the ML based *text categorization* is an inductive process where a *categorizer* can be learned from a set of predefined documents that belong to categories of interests. These predefined documents are manually categorized by experts. The inductive process gleans different characteristics from these predefined documents that an unseen document should have in order fall under a category. The *text categorization* task is a *supervised* learning process, since the learning process is *supervised* by the knowledge of the categories and the training instances that belong to them (Sebastiani 2002). The advantages of this approach is that the process can achieve classification accuracy comparable to the accuracy achieved by human experts. The process saves expert manpower because it does not require intervention from knowledge engineers and domain experts. The advantages of ML based *text categorization* over *rule based* approaches are evident. When the off-the-shelf learner is available, it is an inductive process to build a *categorizer* from a set of manually classified documents. If the category list is updated and ported to a completely different domain, it is only necessary to retrain the *categorizer* on new training set. Here the predefined documents are the key resources.

Text categorization shares characteristics with *information retrieval* and *text mining* (Knight 1999; Pazienza 1997; Sebastiani 2002; Hotho, Nürnberger, and Paaß 2005; Mehler and Wolff 2005). *Text mining* denotes the task of analyzing large quantities of

[6] http://en.wikipedia.org/wiki/Disjunctive_normal_form

texts, detecting patterns and extracting useful information. Sometimes, *text categorization* is considered as an instance of *text mining.*

Sebastiani (2002) defines *text categorization* as a task of assigning a Boolean value to each pair $< d_j, c_i > \in DXC$, where D is the domain of documents and $C = \{c_1, \ldots, c_n\}$ is the set of predefined categories. A *true* value can be assigned to $< d_j, c_i >$ when the document d_j belongs to category c_i. This is single label *text categorization* where one document will get only one label. It is also possible that the same document $d_j \in D$ will be assigned with any number of overlapped categories from 0 to $|C|$. This technique is called multi-label *text categorization.* The tasks we are going to handle in this thesis are single label categorization tasks.

There are many different ML algorithms such as *Expectation Maximization* (EM)(Nigam, McCallum, Thrun, and Mitchell 2000), *Naive Bayes classifier* (Mccallum and Nigam 1998), *Latent Semantic Indexing* (Deerwester, Dumais, Landauer, Furnas, and Harshman 1990), *Neural Networks* (NN) (Dagan, Karov, and Roth 1997), *Support Vector Machine* (SVM) (Vapnik 2000) and more available for *text categorization.* The *support vector machine* (SVM) is one of the most widely used machine learning algorithms. The algorithm is based on *structural risk minimization,* which was introduced by Vapnik (2000). Joachims (1998) and Joachims (1999) use the SVM for *text categorization.* Joachims (1998) also shows two advantages of using a SVM for *text categorization.* The first one is that, SVMs are fairly robust to overfitting and can scale up to considerable dimensions. The second one is that no human and machine effort are necessary in parameter tuning on the validation set, as there is a theoretically motivated, *default* choice of parameter setting shown to provide the best effectiveness (Sebastiani 2002). The main idea of the SVM is to find a hypothesis h that can guarantee the lowest true error (Joachims 1998). The algorithm learns the hypothesis from training examples. In SVM documents $d_j \in D$ are represented by – possibly a weighted vector $(t_{d_1}, t_{d_2}, \ldots, t_{d_N})$ of counts of its features. A SVM separates two classes: a positive class (indicated by +1) and a negative class (indicated by −1). A new document with weighted vector $\vec{t_d}$ will belong to class +1 if the the value $f(\vec{t_d}) > 0$ and belong to −1 otherwise.

A *kernel* function is used to map the points in higher dimensional features space. The goal is to fit a hypothesis h between positive and negative examples in the high dimensional feature space in a way that maximizes the distance between the hypothesis h and data points (Joachims 1998). Now the true error of the hypothesis is the probability that the hypothesis will make an error on a text example. The distance between the hypothesis and a data point is called *margin.* The closest training examples to the hypothesis are called support vectors. If there is no hypothesis found that can separate all training examples correctly, then a hypothesis h can be chosen that performs best to separate training examples. This method is called *soft margin* (Cortes and Vapnik 1995).

Here we focus on two different NLP tasks: *readability* classification and *source and translation* classification. Sometimes *text classification* can be substituted by *text categorization*, *document classification* and *document categorization.*

A ML based *text categorizer* requires three components (Collins-Thompson 2014). The first one is the gold standard training corpus that is representative of the target language, genre and other aspects of the task. Each text in the training corpus is assigned with a label that belongs to categories of interests. Typically, an expert human annotator annotates a text with a label. However, it is a significant issue in data-driven modeling and prediction becomes a time-consuming and expensive task to obtain labels assigned by experts. That is why both the tasks we focus here lack labeled corpora. The learning framework requires a large number of expert labeled training examples.

The second component is a list of *features* that are extracted from texts. The *Bag-of-words* is one of the most widely used features for *text categorization* where a document is presented as a vector of total of word occurrences (Lodhi, Saunders, Shawe-Taylor, Cristianini, and Watkins 2002). However, the *bag-of-words* representation does not preserve information about the sequence of words which is essential for the tasks we are addressing here. Also, features such as *bag-of-sequences*, *n-grams*, *skip n-grams* have been explored for different text categorization tasks. Due to the rapid development of different linguistic tools, new types of linguistics features have been explored for different NLP tasks. These features include *syntactic*, *semantic* and other attributes that are salient for the specific NLP task. We discuss a variety of features in Chapter 6.

The third component is a machine learning model to train a model to predict the correct label of a given text. Generally, the training corpus is divided in three portions. The biggest portion (typically 70%) is used as train data and the remaining corpus is divided into a *tuning corpus* and a *test corpus.* A features vector is generated for each document from the training data along with their corresponding gold standard label. The model parameters are adjusted using the *tuning corpus.* At the end, the model is used to predict the difficulty level of an unseen document. The unseen document is represented in a similar way to the training data with the gold standard label. It is the task of the trained model to predict the label of the unseen document.

2 Research Hypothesis

2.1 Methods

In this thesis, we explore different properties of text that are useful for *text categorization.* We do this in the context of two NLP applications: *text readability* classification and *source and translation* identification. In *text readability*, we explore different text properties that influence reading difficulty. A set of features were selected based on their usefulness for the task. The set of features we explore for readability also reflects different properties of a translated text. That is why we use the same set of features also for the second application, as well. The experimental results in Chapter 7 and in Chapter 12 show that the selected feature set is not only useful for readability classification but also for *source and translation* identification.

We described the related work of *text readability* classification in Chapter 4. We found no previous work that addresses readability of texts from more than one language. This is therefore the first initiative to use the same feature set to categorize text based on readability from three different languages. The languages we considered for this task have distinct linguistic properties. The second application is also to our knowledge the first time that anyone tried to identify original and translated texts from different languages.

The previous chapter established that a ML technique requires gold standard training data (See Section 1.3.1 in Chapter 1). Unfortunately both of the field lack such free gold standard corpora. That is why it was necessary to collect corpora from freely available sources so that we can share these resources within research community for free.

The development of both applications was done in four steps. These steps are: *corpus collection*, *feature extraction*, *implementation* and *evaluation.* For *readability classification*, we collect corpora of three different languages: *English*, *Bangla* and *German.* The two English corpora were collected from the *English Wikipedia* and English textbooks that have been used in schools in Bangladesh. Nowadays, *crowd-sourcing* is one way to label the gold standard training corpus. There are many open-source platforms where readers voluntarily contribute by giving feedback on different aspects of texts. English Wikipedia is such a platform where readers could rate an article based on different parameters including whether the *article was complete*, *trust-worthy* and *well-written*. The ratings can be used as difficulty labels of the articles, or grade levels can be computed from the ratings. The standard unit for a reading difficulty label is grade level. These grade levels are ordinal values that correspond to discrete ordered reading difficulty levels (Collins-Thompson 2014). For example: American grade levels from *one* through *twelve* can be continuous values to label the training corpus. However, it is not easy to collect a training corpus for each grade level. To circumvent the problem, researchers often use difficulty labels such as *very easy*, *easy*, *medium* and *difficult*. We have followed this procedure.

The Bangla textbook corpus is collected from the same source as the English textbook corpus, but these books are in Bangla. Significant pre-processing was involved to convert these texts into a usable form. Section 5.2 in Chapter 5 describes the corpus in detail. The German corpus was compiled and used by Hancke, Vajjala, and Meurers (2012) for text readability classification. These texts were crawled from a German magazine site called *GEO*.

In our search for forerunners (See Chapter 9) to our translation application, we could find no previous studies that provided a multilingual corpus for translation studies. Scholars in the field of *translation studies* have proposed several translation properties using monolingual or comparable corpora. Obviously there is a need for a multilingual corpus with proper annotation of original and translated sentences. Translation scholars can use the corpus to validate their theoretical findings. We provide such a corpus for the translation studies community. The corpus is compiled from the well known *Europarl* corpus (Koehn 2005).

We have considered both tasks as supervised categorization tasks. That is each input document will get a label from one of the pre-trained classes. The test data is separated from the training data. Our goal was to explore features that are language independent and do not require any linguistic pre-processing. We experimented with a variety of features. In the end, we selected 18 features that are shown to be useful for both applications. Seven of them are from the field of *information theory* and the rest are *lexical* features. A feature was selected if and only if the accuracy of the classifier was improved by adding the feature into the list. At the beginning the feature list contained one feature, that is average SL.

In order to find the most useful features, we compared many state-of-the-art features.

Algorithm	Accuracy	F-Score
Naive bayes	59.59%	59.10 %
Decision tree (J48)	69.57%	69.60%
LibSVM	32.44 %	15.90%
SMO	71.29%	71.60 %

Table 2.1: Performance of different text categorization algorithms.

We describe features in Chapter 6. For comparison, we have used the English Wikipedia corpus and the English textbook corpus. We used NLP techniques to explore different text features at several linguistic levels, *lexical* features that are used in traditional readability formulas, as well as features based on *language modeling*, *part-of-speech*, *syntax-based* features, *semantic-based* features of *coreference chains*, *coreferential inference*, *semantic role* and *semantic frame*. Some of the semantic features we have explored here have been used for the first time for text readability classification.

The classifier using linguistic features is compared with a baseline system. We use five traditional readability formulas in order to build the baseline system. The reason for selecting these five formulas is that they are still being used in different readability measurement tools (De Belder and Moens 2010). Section 6.1 in Chapter 6 describes these formulas in details.

We use the WEKA toolkit (Hall, Frank, Holmes, Pfahringer, Reutemann, and Witten 2009), known for high quality multi-class classification. The toolkit contains many machine learning algorithms. The Sequential Minimum Optimization (SMO) (Platt 1998; Keerthi, Shevade, Bhattacharyya, and Murthy 2001) is one algorithm used to train a SVM using the Pearson VII function-based universal kernel PUK (Üstün, Melssen, and Buydens 2006). We experimented with different classification algorithms and SMO performed the best. See Table 2.1 for the evaluation. It is necessary to find an appropriate *kernel* in order to train a SVM classifier. In this thesis, we use the *Pearson VII function based kernel* (PUK) kernel, because it can function as a generic kernel. We also performed similar experiments to select the best performing kernels for the SMO algorithm. Table 2.2 shows that the PUK kernel performs the best of all. This kernel has been used in many text classification tasks.

Various models have been trained with features implemented in this study and we used standard *Accuracy* and *F-score* to evaluate them. For each experiment the corpus was divided into a *training* corpus and a *testing* corpus. The *training* corpus contains 80% of the original corpus and the rest of the corpus is used as a *test* corpus. We repeat each experiment 100 times where *train* and *test* corpora are randomly selected. Finally

Kernels	Accuracy	F-Score
Poly kernel	71.29%	71.60 %
Poly kernel (normalized)	67.86%	68.10%
PUK kernel	71.45%	71.80%
RBF kernel	55.38%	45.90 %

Table 2.2: Performance of different kernels.

the weighted average of *Accuracy* and *F-score* are calculated from the results of these experiments.

2.2 Hypothesis

Both tasks that we are concentrating on in this thesis are influenced by the theoretical cognitive model. It is important to consider the respective theoretical cognitive framework. Our research is guided by theoretical frameworks that were established in the field of cognitive science to understand cognitive processes while reading a text and translating a text from one to another language. Section 3.4 and in Chapter 3 and Section 9.2 in Chapter 9 describe these cognitive models.

According to Kintsch (1998), the cognitive model of a reader is to build meaning word by word, sentence by sentence and paragraph by paragraph. It is an incremental process. Mahowald, Fedorenko, Piantadosi, and Gibson (2013) stated that “speakers are more likely to choose a short form in a supportive context”. That is, when the cognitive model exists, the speaker is biased to use the shorter form of the word, for example: *exam* instead of *examination.* The same characteristic might be reflected in writing. *Information density* reflects some cognitive features that influence reading difficulty.

From the field of information theory, the most accurate way to measure information was introduced in the 1940s by Shannon (1948). Furthermore the most efficient way to send information through a noisy channel is at a constant rate (Genzel and Charniak 2002; Genzel and Charniak 2003). Plotkin and Nowak (2000) have shown that this principle also correlates with biological evidence of how human language processing evolved. This rule must be retained in any kind of communication to make it efficient. That is, in a text the entropy of an individual sentence increases with its position in the text, if the entropy is measured without considering the context. The increasing rate can vary in terms of the difficulty level of a text, since a more readable text differs from a less readable text in many ways. In order to utilize this *information-theoretic* notion we start from random variables and consider their entropy as indicators of readability. Our hypothesis is that the higher the entropy, the less readable the text along the feature represented by a random variable X. Entropy of a random variable is proportional to the difficulty of correctly guessing the value of the variable. Entropy is highest when the values are equally probable and lowest when the one of the choices has a probability of 1.

Shannon showed language to be a communication channel where information is encoded in a sequence of basic symbols. However the statistical structure of a linguistic sequence was not considered in his study. Recently, Montemurro and Zanette (2002) showed that *entropy* of *word distribution* in a text carries information of the specific linguistic role of word classes. According to them, *adverbs* and most of the *verbs* tend to be uniformly distributed. Also, statistical regularities of some words in a text may reflect their linguistic roles in that text. Readability of a text also depends on the predictability of a text. A text with lower entropy will be more predictable and therefore easier to read.

Keller (2004) validates the claim by Genzel and Charniak (2002) and Genzel and Charniak (2003), in which the entropy of sentences increases with the position of the texts. He also showed that this principle holds for individual sentences. The entropy rate effect is influenced by sentence length that correlates with sentence position. We also calculated the *non-parametric* entropy of different readability corpora. Figure 2.1 shows the entropy of different readability classes in the English Wikipedia corpus. In order to estimate the entropy constancy we randomly selected four documents from the four readability classes. Documents from the *very easy* readability class had a higher entropy compared to documents from the *difficult* readability class. The annotation of documents in terms of readability of the English readability corpus might also be influenced by other factors.

Figure 2.2 and Figure 2.3 show the entropy constancy in the *German* readability corpus and the *Bangla* readability corpus. For all of the corpora the entropy increases with the number sentences. This non-parametric entropy estimation shows that a good readable text differs from a less readable text, which satisfies our hypothesis. These figures confirm the claim by Genzel and Charniak (2002) and Genzel and Charniak (2003) about constancy of entropy in texts.

Reading comprehension is also influenced by different aspects of a text. These aspects start with lexical variables such as the frequency of the words in a language, average word length and so on. Different syntactic variables that measure complexity of sentences are

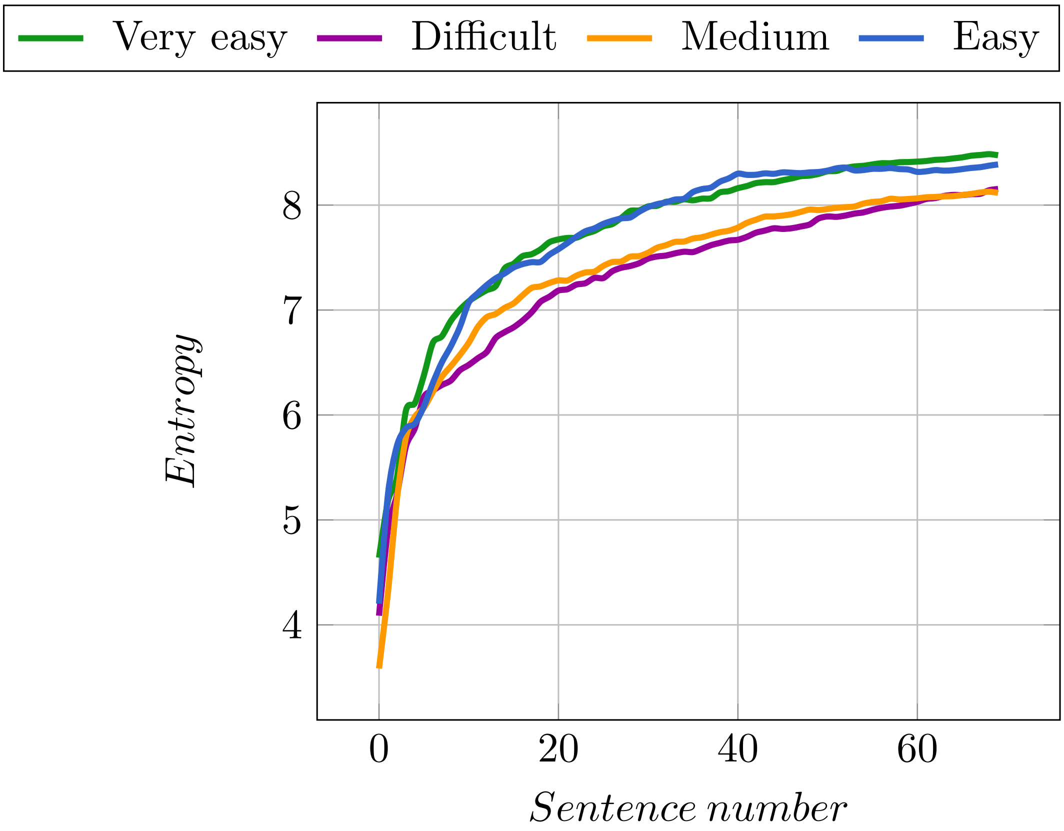


Figure 2.1: Non-parametric estimation of entropy in the English Wikipedia corpus

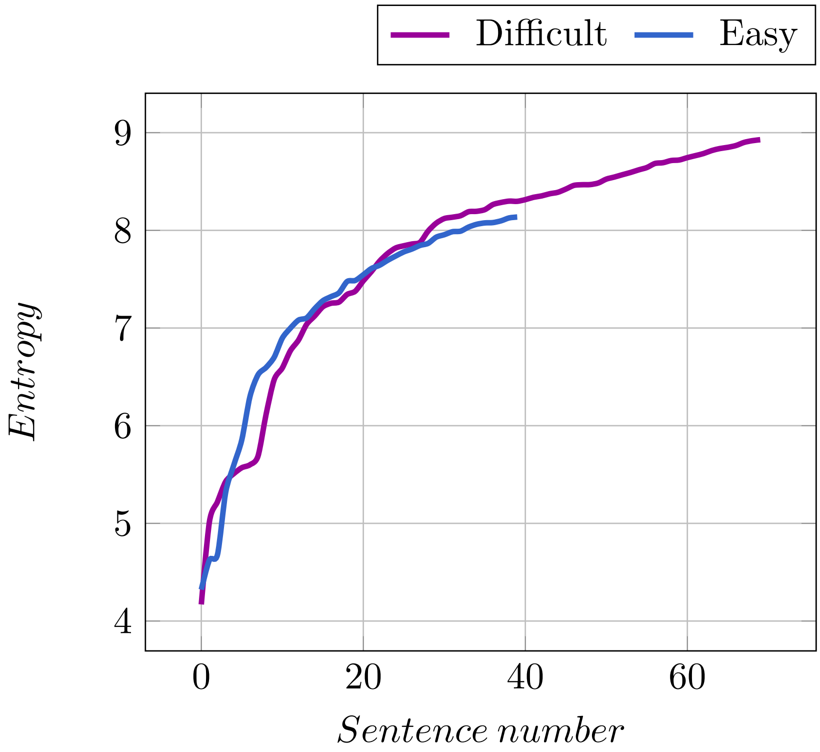


Figure 2.2: Non-parametric estimation of entropy in the German GEO/GEOLino corpus

included such as, average sentence length, average number of phrases in a sentence, average number of clauses in a sentence.

In recent studies in the field of text readability classification, researchers have used different linguistic, language model related features (Si and Callan 2001; Collins-Thompson and Callan 2004; Pitler and Nenkova 2008; Schwarm and Ostendorf 2005; Aluisio, Specia, Gasperin, and Scarton 2010; Kate, Luo, Patwardhan, Franz, Florian, Mooney, Roukos, and Welty 2010; Eickhoff, Serdyukov, and Vries 2011). In order to check the usefulness of our proposed features, it is necessary to compare them with state-of-the-art linguistic features. That is why, we have considered different linguistic based features that are used in previous studies mention in Chapter 4. We agree with Kintsch (1998) and Feng (2010) that the working memory burden inflicted by various semantic and discourse processes is crucial for coherent memory presentation construction. That is why we explored features that reflect semantic and discourse related properties of a text. *Semantic frames* is a crucial linguistic property that influences the readability of a text. Sometimes it is hard to understand the underlying meaning of a single word without considering essential related knowledge about the word. For example, a reader would not be able to understand the word *wedding* without knowing about the social event where different persons such as *bride*, *bridegroom*, *relatives*, *friends* and more are involved. *Semantic roles* is another crucial semantic feature; it shows the actual role a participant plays in some real or imagined situation. For example: it is also important to know the role of a *bridegroom* in a *wedding* to have a better understanding of a sentence with *wedding*.

Although readability of a text is influenced by different linguistics factors, most of them have not been studied as a part of cognitive model of reading comprehension. One reason could be that most of the state-of-the-art linguistic tools are not advanced enough to process natural language texts without producing any error. That is why most of the cognitive models related to reading focus on surface or lexical level features.

The ASL is one of the oldest features used to predict readability difficulty. This feature is related to the syntactical complexity of a sentence (Rayner, Pollatsek, Ashby, and Clifton Jr 2012; Harly 2008). Generally, the syntax of a longer sentence is more complex than the syntax of a short sentence. Sentences in a text that is aimed at children are simple and short. Translators also try to break a complex sentence into two or more simple translated sentences. These simple translated sentences are shorter in length than the corresponding original sentence.

The AWL is also a cognitively motivated factor that influences readability difficulty. Generally, a text that is aimed for children does not contain longer words. A longer word will be harder for children to pronounce and comprehend.

Function words, like *the*, *from*, *and*, *all* and so on, often have little lexical meaning, but they clarify relationships between content words. Function words are more frequent and predictable than content words. Generally, function words carry grammatical information about content words. High frequency function words are relatively shorter than mid/low frequency function words (Zipf 1935; Bell, Brenier, Gregory, Girand, and Jurafsky 2008). Due to their high frequency in texts and their grammatical role, function words also indicate authorial style (Argamon and Levitan 2005). *Lexical density* is the ratio of these *grammatical words* and *function words*. They can be computed using different type-token ratios (TTR). In this thesis we use various TTR formulas that are widely used for various NLP tasks. *Lexical density* is also a useful feature for source and translation classification. Hansen (2003) states that the *lexical density* tends to be lower in a translated text than the corresponding source text. More specifically, a translated text will contain more *function words* and fewer *grammatical words* than corresponding source text. *Most frequent words* and *familiar words* are similar in a sense that most frequent words are familiar words. According to Rayner, Pollatsek, Ashby, and Clifton Jr (2012) and Harly (2008), a more readable text contains more familiar words.

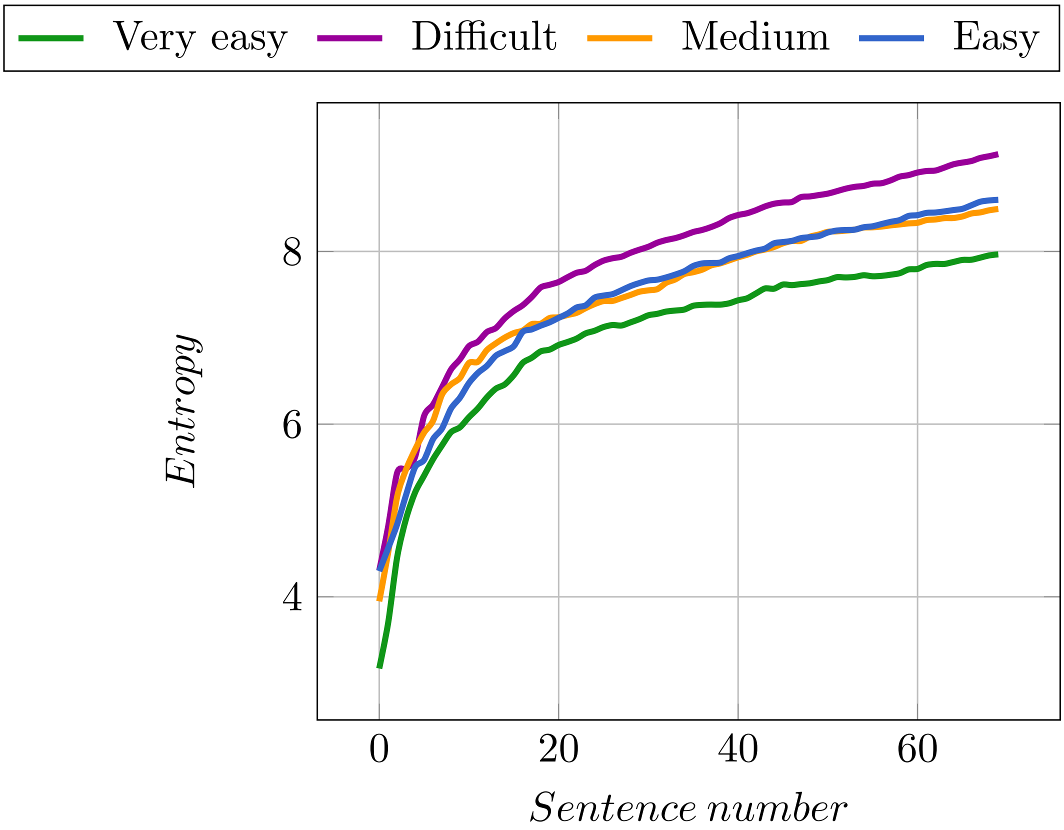


Figure 2.3: Non-parametric estimation of entropy in the Bangla textbook corpus

3 Text Readability

3.1 Text readability

Readability deals with the difficulty or ease with which a text can be read. A text that is easy for a reader to read can be difficult for other readers. Therefore, the problem is to match different levels of reading materials with different reading levels. More specifically, readability means grading materials and fitting them to readers according their levels.

Readability of texts is one of most discussed concepts in reading. The exact time line when people became concern about readability is unknown. However, famous Greek philosophers *Aristotle* and *Plato* were conscious of the comprehensibility of their speeches (Mikk 2005). Nowadays, the comprehensibility of texts, known as *text readability* or *text complexity*, has become increasingly important. In the modern era, people first become concerned about text readability in early twentieth century (Biere 2000).

Many researchers define text readability differently. Dale and Chall (1949) defined readability as: "the sum (including the interaction) of all those elements within a given piece of printed material that affects the success that a group of readers have it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting". McLaughlin (1969) provides a slightly different definition. He defines readability as: "the degree to which a given class of people find certain reading matter compelling and comprehensible". This definition focuses on the relationship between reading difficulty, readers reading skills and readers prior knowledge on the topic. A more specific definition could be that readability of a text is the level of ease with which a text can be understood by a particular reader for a specific purpose. Readability of a text depends on a variety of text factors and different characteristics of readers. Harris and Hodges (1995) describes "text and reader variables interact in determining the readability of any piece of material for any individual reader". The same text could be very easy for reader X but could be very difficult for reader Y. Prior knowledge greatly influence how well a reader can understand texts dealing with a particular topic. A *software engineer* might easily read a technical report on *software design and life-*

cycle. The same report could be very difficult for readers without prior knowledge on the topic. Recently, Montemurro and Zanette (2010) stated that "written language is a complex communication signal capable of conveying information encoded in the form of ordered sequences of words". Text production is the process of information transfer between authors and target readers.

Along with prior knowledge on the topic to be read, there are more factors that influence text complexity. Gray and Leary (1935) listed 288 factors that influence readability of a text. Among these, a single factor may not indicate difficulty. However, a collection of factors can draw a complex relationship between texts and readers. These factors include the syntactic complexity of sentence, semantic, density of concepts, abstractness of ideas, text organization, coherence, sequence of ideas, page format, length of type line, paragraph length, use of punctuation and graphical factors such as use of image, illustration and color (Rubin 1981). The reading ease can not be determined simply by these intrinsic text properties. It is a result of an interaction of a complex language comprehension process between the reader and the text (Miller and Kintsch 1980; Davison and Kantor 1982). In addition to these text properties readers' literacy skills, motivation and interest to read also influence the readability of a text.

From the beginning of the research on text readability, researchers were not aware of the relation between *cognition* and *reading.* After 1970, researchers start thinking about how *cognition* is related to *readability* after 1970 (Biere 2000). In addition to this, the research community started to study *text structure*, *text simplicity* and *text cohesion*. *Text structure* helps the reader to understand texts. For example: research publications have different structure that makes them distinguishable from other kinds of documents. A text becomes more readable with the explanation of any domain specific words. Sometimes, visualization also helps readability.

There are two types of text readability (Tuldava 1993). One is *subjective* readability, which refers to how difficult or easy a text for the targeted readers is. Another way of expressing it is, how easily can a reader comprehend the text and reproduce it. *Subjective* readability can be measured by the speed of reading, and by asking readers questions about the subject of the text. The other one is the *objective* readability that refers to the complexity or sophistication of a text. This complexity is related to the text structure and content of the text. For example, a text will be difficult to read if it contains mostly long sentences and a large number of unfamiliar words.

To measure both types of readability, it is required to use a unit. Most of the studies on readability use grade level as the reading difficulty unit, but other scales of measurement are used. The grade level could be an ordinal value that corresponds to a difficulty level. For example: grade levels in America go from level 1 to level 12. However, difficulty levels can be used directly to measure reading difficulty. As an example, a text could be labeled as *easy* or *difficult*. In this thesis we use this scale.

Readability measurement is a task of mapping texts onto scales of readability. That is why the task can be reconstructed as an automatic text categorization (Sebastiani 2002) application. The application refers to the classification of texts into predefined categories. These predefined categories can be grade levels or difficulty classes.

3.2 Measurement of text readability

From the beginning of the research on text readability, researchers have explored different text features. Their assumption was that the readability of a text could be measured by a simple function of a few text properties that are objective and easy to determine (Miller and Kintsch 1980). There are multiple readability formulas available to measure readability of a text that use these text properties. These formulas are based on correlation analysis where researchers have selected a number of text properties and asked readers to grade readability of various texts on a scale. These studies were focused on average SL, average WL, average number of syllables in a word and number of polysyllabic words. These features are combined with some constants.

Nowadays, different linguistic features have been explored to measure reading difficulty. Linguistic factors play an important role in text readability. These factors can be explored in different linguistic levels. Some linguistic features are able to capture these factors, such as: unfamiliar or ambiguous word can measure the lexical complexity of a text, complex morphological particles can represent morphological difficulties of the corresponding text, syntactical complexity can be measured by grammatical structure of the text. All these features can be explored by a lower level analysis of a text. There are some higher level features that also measure text difficulties. Unusual sense, idioms and subtle annotation represent semantic difficulties. Clear argument structure of texts, inter-sentence relationship, and discourse connectives can measure text difficulties based on discourse and cohesion. Language influenced by genre and domain knowledge can measure high level conceptual difficulties of a text.

From the field of information theory, some statistical properties can capture the complexity of a text. Entropy is constant in any communication through a noisy channel. A text is a communication medium between an author and its readers. That is why the entropy of a easier text will be different than a difficult text. Apart from these, there are more quantitative features that have been used to determine text readability.

3.3 Applications of text readability

The basic application of text readability analysis is the measurement of the readability of given texts. People from a variety of professions use such a tool. Professionals, such as teachers, journalists, or editors, produce texts for their specific audiences. A readability measurement tool could be very useful for them. Apart from this basic use, there are more places where such a tool could be very useful. The following sub-sections describe a few applications of text readability analysis.

3.3.1 Customized content delivery

Now a days people access information from the internet more than ever. However, the reading ability differs among people. To make the best use of the content information, it is important to provide content that is suitable for the target readers. A system with customized content delivery will have different version of the same web page. The different version will be distinguished based on readability difficulty levels.

3.3.2 Web page recommendation

The *Google* search engine ranks first by the *Alexa Internet*[1] among all of the websites in the world. People use the site for searching with their specific *keyword* or *queries*. The amount of information in the Internet is enormous and grows exponentially. It is difficult for a search engine to recommend web pages that are suitable for a specific group of users. Addressing search queries from children, for instance, need to be handled specially by solving problems such as interface design, content filtering and result presentation. The main factor is to provide relevant results with appropriate reading level (Collins-Thompson, Bennett, White, Chica, and Sontag 2011).

As an example the *Google* search engine mostly shows Wikipedia pages as the first recommendation, if users search for an *entity* where Wikipedia has an article on it. However, the article about the same entity from the *Simple Wikipedia* would be more suitable for children.

[1] Alexa Internet is a company that provides commercial web traffic analysis.

3.3.3 Readability investigation of web pages

During the design of a new website, content is commonly overlooked to increase accessibility. Even if a web site follows all W3C recommendations it might be inaccessible to a certain reader group due to reading difficulty. There are some tools that use traditional formulas to measure reading difficulty. However, the formulas do not work well on web pages. The reason for this failure is that the content of web pages are non-traditional (i.e., not always well formed sentences) and contents vary greatly. Web pages also contain other components such as images, video, tables and other layout elements. The ability of users to understand a web page would be a critical aspect of the page's value. That is why readability of a website has to be considered during designing of the site. Collins-Thompson, Bennett, White, Chica, and Sontag (2011) and Vajjala and Meurers (2013) addressed several issues related with readability investigation of web pages.

3.3.4 Readability for second language learners

The second-language learners (L2) have different reading skills compared to first-language (L1) learners. The main difference between these learners is the time of life when they learn the language and the process they use to learn. According to Bates (1999), an L1 learner starts language acquisition when the learner is in *infancy* and starts acquiring grammar from the age of four. L2 learners often start acquiring the language later when they are college-age or older. During that time they already have a conceptual lexicon, complex ideas and contextual environment that are related with lexicon and ideas. L1 learners are not conscious of the syntax during learning. However, L2 learners spend a significant amount of time to learn the grammar of the target language. Most of the traditional formulas and also recent readability research are focused on L1 learners. A readability assessment tool for L2 learners will be very different than a tool focused on L1 learners. Syntax and grammatical features might play an important role in readability analysis for L2 learners.

Also, second language teaching centers utilize different reading materials along with textbooks. Textbooks alone do not provide enough material in order to learn to read fluently. Here the challenge is to match the reading materials with learners's reading ability. To circumvent the challenge, a readability classifier can be used.

3.3.5 Supporting readers with disabilities

It is always important to understand how typical students learn reading. However, it is also necessary to understand how reading development occurs for student with disabilities. The reading development of students with learning disabilities and dyslexia may differ in may ways from typical student. Abedi, Kao, Leon, Mastergeorge, Sullivan, Herman, and Pope (2010) examined the impact of segmentation for both types of students. They found that segmentation has no impact on typical students or on students with learning disabilities. Abedi, Bayley, Ewers, Mundhenk, Leon, Kao, and Herman (2012) performed another examination to explore traditional readability formulas for reading test items for identifying grammatical and cognitive features that influence readability for students with disabilities and have a negative impact on performance. According to them *long words* (i.e., words with eight or more letters), *font*, *spacing* and *distracting images* are important features for texts that are targeted for readers with disabilities. Rello, Baeza-Yates, Dempere-Marco, and Saggion (2013) found similar results for student with dyslexia. A readability measurement tool is necessary in order to identify or prepare suitable reading materials for student with disabilities.

3.3.6 News filtering for children

Newspapers try to target certain audiences through different topics of stories. Children are also in their target audience. This target group are their future readers.

Nowadays children also read news online. One third of children in developed countries such as the *Netherlands*, the *United Kingdoms* and *Belgium* browse the internet for news (De Cock 2012; De Cock and Hautekiet 2012). Livingstone, Haddon, Görzig, and Ólafsson (2010) showed that, one fourth of the British children between the age of *nine* and *nineteen* look for news on the internet. The ratio could be similar in other developed countries where most of the citizens have access to the internet.

The news for children will vary linguistically and cognitively from news for adults. This is similar for websites that are dedicated to children. De Cock and Hautekiet (2012) observed difficulties for children in navigating these websites. Readability of the texts is one of the reasons. So, a readability measurement tool would be useful to measure readability of these news sites.

3.3.7 Other applications

Text simplification is a NLP application that removes grammatical or lexical difficulties of text by modifying a text while the underlying meaning and information remain the same. A readability classifier can be used to identify difficult sentences as a module of a text simplification system.

A readability classifier can be useful for other NLP applications. Automatic essay grading can benefit from readability classification as a guide to how good an essay actually is. Automatically generated documents, for example documents generated by text summarization systems or machine translation systems, tend to be error-prone and not very readable. In this case, a readability classification system can be used to filter out documents that are less readable. Similarly, a readability classifier can also be used to evaluate machine translation output.

3.4 Cognitive dimension

Text readability is influenced by readers's cognitive model. While we read a text, we understand something from the text by building a mental model. The construction of the mental model is sequential. It is psychologically impossible to construct and integrate a text representation of a text or a whole book chapter (Kintsch 1998). The text or the chapter has to be processed word by word, sentence by sentence.

Readers start building their reading skills from an early age. Children use their cognitive skills to perform different tasks in different environments. Kali (2009) stated that children refine their motor skills and start to be involved in different social games when they are 5 to 6 years of age. From age of 6 to 8, children start to expand their vision beyond their immediate surroundings. Children from 8 to 12 years of age acquire the ability to present different entities of the world using concepts and abstract representations. Children become more interested in social interactions in their teenage years. As their age increases, their reading and cognitive skills improve.

Reading skills require two processes: *word identification* and *comprehension.* Readers have to identify a pattern of letters in the*word identification* process. As a prerequisite, readers must have some knowledge about these letters and their patterns. For example: it is impossible to recognize any word from any language without knowledge of the letters of that language.

Comprehension is a process of extracting meaning from a sequence of words. The sequence of words follow an order. It could be impossible for children to understand a sentence where the order of the words is random. A sentence can be broken into text segments, if it is long. A new mental model starts to be built when a reader starts reading a text. When a text segment is processed, it will be immediately integrated with the existing mental model. The model keeps growing as the reader moves forward.

The process of reading is guided by a schema. The schema regulates the *comprehension* process in a top-down fashion (Schank and Abelson 1977). However, flexibility and context-sensitivity of human communication make this schema questionable. Kintsch (1998) argues that *comprehension* is a loosely structured, bottom up process that is highly sensitive to the context. He proposed a model called *construction integration* (CI). In this model, at first readers construct *propositions* from text segments. A new *proposition* is be added to the network as a node after each text element is processed. Nodes in the network are connected with other nodes. These connections are basically four types. *Direct*, where two *propositions* are directly related; *indirect*, where two *propositions* are indirectly related; *subordinate*, where one *proposition* is subordinated by another one and *negative*, where two *propositions* are negativly related. Nodes with negative or fewer connections will be suppressed. A network will be strengthened when nodes are positively connected with many other nodes (Kintsch 1998). It is necessary to satisfy constraints between nodes in order to be connected. This connectivity is an indication of *coherence* (Mehler 2006). A network becomes stronger when nodes in that network satisfy multiple constraints (Kintsch 1998). At the end it shows the role of each node in the network and the strength of nodes in the mental model.

The output of the *comprehension* process is a mental representation of a text, the episodic memory (Kintsch 1998). There are two components of this episodic memory: the *textbase* and the *situation model*. The *textbase* contains different elements and their relations derived from the text. Readers read a text and translate it to a propositional network then expand it cycle by cycle. In order to make the structure complete and coherent, a reader has to add nodes and establish links between nodes from his or her knowledge and experience. This knowledge and these experiences are considered as external sources that are necessary to build the situation model. Example of these sources are: knowledge about the world, knowledge about the language and other specific communications. This knowledge grows as a reader gets more experience in reading.

The *comprehension* process is also influenced by the memory system. The cognitive system of humans contains three different memories: *sensory memory*, *short-term memory* and *long-term memory* (Rayner, Pollatsek, Ashby, and Clifton Jr 2012). Sometimes, *sensory memory* is referred as *sensory store*. There are two types of sensors in the sen-

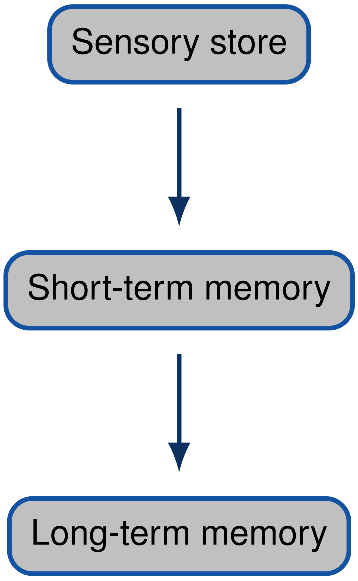


Figure 3.1: Human information-processing system (Rayner, Pollatsek, Ashby, and Clifton Jr 2012)

sory store: *visual* and *auditory*. The auditory sensor is referred to as echoic memory while the visual is called iconic memory. Iconic memory is more transient than echoic memory. Through these sensors information enters the short term memory. The *sensory store* contains raw, un-analyzed information very briefly. Most of the information is lost before it is stored into the *short-term memory*. The *short-term memory* is referred to as *working memory*. The cognitive process takes place in *short-term memory*. Older children sometimes are better where they simply retrieve a word from their memory while reading. A younger children might have to ”sound out” a novel word spelling. However they are also able to retrieve some familiar words. Some children might struggle to recognize words which make them unable to establish links between words. Children without this problem are able to recognize words and derive meaning from a whole sentence. Generally, older children are better readers due to their working memory capacity where they can store more of a sentence in their memory as they are able to identify propositions in the sentence (De Beni and Palladino 2000). Finally information comes into the *long-term memory* from the *short term memory*. This process is time consuming due to the information loss. *Long-term memory* is the permanent storehouse of knowledge about the world (Kali 2009).

Experienced readers are able to comprehend more than younger readers due to their word recognizing skills and their extended working memory (Kali 2009). They also know more about the world and use appropriate reading strategies. In summary, a reader become more skilled as their working memory develops over time. The combination of these memories called *information-processing* system. Figure 13.1 shows the *information-processing* system.

The *information-processing* system is actively involved from the beginning of the cognitive process until the end. Information has to be perceived in order to be processed. *Eyes* perceive information from a text when they stay still and fix on one point in the text. This period is called *fixation.* The *eyes* jump from one point to another point. These continuous movements of the eyes is called *saccade* (Rayner, Pollatsek, Ashby, and Clifton Jr 2012). Between two *saccades* the eyes are relative still, and this *fixation* lasts for about $200-300$ milliseconds (Rayner 1998). This *fixation* phase for *silent reading* is less than the *fixation* phase when *reading aloud.* The *saccade* shows how the eyes move during reading. Most *saccades* go forward about eight to nine character positions while reading texts in *English.* Generally *saccade* in a typical reading session takes 30 milliseconds (Rayner 1998). Reichle, Pollatsek, Fisher, and Rayner (1998) defined another term called *return sweeps* where readers move their eyes from the end of a line to the beginning of the line. During this time, the eyes can capture around $18 - 19$ characters at time with $3 - 4$ to the left of the focus point and up to 15 to the right (Harly 2008; Temnikova 2012).

Eye movements in reading are largely driven by lexical access (Reichle, Pollatsek, Fisher, and Rayner 1998). In a *lexical access* interval, the mind requires $100 - 300$ milliseconds in order to recognize a word. The *fixation* time is short for most frequent words (Rayner 1998). Letters in common short words appear to be processed in parallel (e.g., at the same time) (Temnikova 2012).

According to Reichle, Pollatsek, Fisher, and Rayner (1998) and Reichle, Rayner, and Pollatsek (2003), readers' eyes are fixated at the first point of text while they start reading. The eyes keep moving constantly. The reason the move relatively constantly is that there are severe limits to how much visual information can be processed during a *fixation* (Reichle, Rayner, and Pollatsek 2003). The eyes keep moving forward until the visual system can not recognize any more due to the acuity limitations. The eyes shift to a previous point until information is processed. It is necessary to move the eyes backward if a textual component (e.g., words or phrase) is unclear. About 10% readers move the eyes backward in case something is not clear; the term is called *regression* (Reichle, Pollatsek, Fisher, and Rayner 1998). The *fixation* time is longer if comprehension difficulties arise. Readers need to move their eyes to the previous *fixation* point in order to have a better comprehension of the difficult point. At the beginning of a reading session, a reader's task involves *word identification*, *phrase* and *sentence comprehension.* The ultimate goal is to extract meaning from the basic component units of the texts. Word identification does not consist only of identification of a given string of letters, but includes relating all linguistic information of the word that these letters signify. This information includes the meaning of the word, the part-of-speech, how the word is pronounced and also the underlying meaning.

Word identification also influenced by different factors. Temnikova (2012) listed some of the factors that aid *word identification. Semantic priming* is one of them, where a word can facilitate understanding of the immediate context because of related meaning. The *syntactic structure* of a sentence provides information about the part of speech of the next word. Other factors include *high frequency* words and vocabulary. *Frequent* or *familiar* words are faster and easier to identify (Whaley 1978). The vocabulary that are learned in early age of acquisition that are also easier to process (Barry, Morrison, and Ellis 1997). The *word identification* process is also influenced by *experience.* Martin (2004) shows that experienced readers can process a larger span of characters than ordinary readers. *Longer words* could be difficult for particular types of readers. For example: identification of *biodegradable* could be hard for children at lower grade levels. Orthographically similar words (words that can be written by changing one letter) carry extra visual difficulty for readers with surface dyslexia. Most orthographically similar words have a length of four letters (Andrews 1997). Temnikova (2012) provides examples such as: *mine/dine/pine/fine/nine/sine.* A word with multiple parts-of-speech (POS) or with semantic ambiguity can also be difficult even for experienced readers.

3.5 Challenges

Although, it is one of oldest research field in NLP, the field still faces some challenges. These challenges differ with different languages. For high-resource languages, researchers explore different features at the linguistic level. However, the readability measurement result is influenced by the performance of different linguistic tools. State of the art syntactic parsers can achieve an *F-score* value of around 85% for morphologically rich languages (Seddah, Tsarfaty, Kübler, Candito, Choi, Farkas, Foster, Goenaga, Gojenola, Goldberg, et al. 2013). The performance of other deep linguistic processing tools is not as good as the syntactic parser. This performance effects the readability analyses that use linguistic features.

Languages that are identified as low-resourced have different problems. There are few or no linguistic resources available for these languages. That is why selection of features is limited. Traditional formulas only work for a specific language. Lexical and surface features could be used for these languages. However, these features skip different aspects of texts that are important in measuring the readability of a text. Therefore, the challenge is to explore features for these low-resourced languages that do not require any degree of linguistic processing but are still able to achieve reasonable accuracy.

4 Text Readability Related Work

The research on text readability started in the early twentieth century (DuBay 2004). By the 1980's, over 200 readability formulas had been developed (DuBay 2004). Nowadays, these formulas are called *traditional readability formulas*. Research in the field is carried out in two disciplines, *computer science* and *computational linguistics*. However,the problem is not solved yet, and research in this area is still active. Most of the researches in the field is focused on texts from the English language. However, texts from other well-resourced languages such as *German*, *French*, *Italian* are now receiving similar analysis. The following sub-sections describe some related works for texts in English and other languages.

4.1 Related work for English texts

English is the dominant language in the field of *text readability* research. All of the traditional formulas listed in DuBay (2004) are proposed for *English* texts. These features are fairly simple and were explored due to a lack of linguistic resources and computational power. The *traditional formulas* rely on two main factors: semantic units such as the *word* or *phrase* and the syntactical complexity of a sentence (Collins-Thompson 2014). In order to make these formulas easy to apply, researchers made two simple assumptions where these syntactical and semantic factors can be obtained using proxy variables. In most of traditional formulas, the *number of syllables* in a word is used as a proxy variable to project *semantic difficulties*, and the average WL or average SL are used to measure *syntactical complexity* (Collins-Thompson 2014). These proxy variables are averaged over words and sentences without considering the ordering of words. That is why text difficulty related to higher-level linguistic structures such as discourse flow and topical dependencies are overlooked by these traditional formulas.

We can summarize that traditional formulas use simple linear functions with some shallow surface level features to model the difficulty of a text. These features focus on the lexical and syntax level difficulty of a text. Lexical level features are the number of syllables a word contains, average WL and word frequency in a text. The average SL can be used as a measure of grammatical complexity assuming that a longer sentence has a more complex grammatical structure than a shorter sentence. All of the traditional formulas use these shallow features. Dale and Chall (1948) and Dale and Chall (1995) proposed a vocabulary based traditional formula where they estimate semantic difficulty using a *reference word list.* They identified a *difficult* or *unfamiliar* word if the word was not in the *reference word list.* The *reference word list* consists of 3,000 commonly known words for the 4^{th} grade (i.e., children approximately 10 years old). Their assumption was that a text containing familiar words will be easier to read and understand. The *difficult* or *unfamiliar* variable that measures semantic difficulty is combined with a syntactic variable, such as average SL in order estimate sentence difficulty. Lennon and Burdick (2004) also used vocabulary-based traditional formulas in their *Lexile Measure* tool.

Gunning (1952) also considered the numbers of sentences and complex words in order to measure text readability. The formula uses similar lexical features to (Dale and Chall 1948; Dale and Chall 1995), but with different constants. This research focused on word complexity that is directly related to the number of syllables in a word: the smaller number of syllables, the easier the word. In the time when these formulas were proposed, it was not easy to count the number of syllables in a word. Any word that contained three or more syllables was counted as a complex word. Later average word length was used to capture word complexity.

The Flesch-Kincaid readability index (Kincaid, Fishburne, Rodegers, and Chissom 1975) considers the average number of words per sentence and the average number of syllables per word. This is the most widely-used traditional formula. The formula has been implemented as a feature to measure text readability in word processing software such as *Microsoft Word* (Friedman and Hoffman-Goetz 2006). They proposed two different formulas, one for measuring how easy a text is to read and the other one for measuring grade 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. The formula gives a number which approximates the grade level required to comprehend a text. A number 3 means any student who is in the $3rd$ grade (ages 8 - 9 years old) should be able to comprehend the text.

McLaughlin (1969) also proposed a traditional formula that calculated the years of education needed to understand a text. The formula is called SMOG, that is derived from *Simple Measure of Gobbledygook.* The formula requires 30 sentences from three different positions in a text (e.g., *beginning*, *middle* and *end*). The formula assume that *polysyllabic* words in a sentence influence readability. Any word that contains three or more *syllables* will be considered as a *polysyllabic* word. Fitzsimmons, Michael, Hulley, and Scott (2010) stated that the "*SMOG* should be the preferred measure of readability when evaluating consumer-orientated healthcare material". They also found that the readability formula proposed by Kincaid, Fishburne, Rodegers, and Chissom (1975) significantly under-estimated reading difficulty compared with the gold standard *SMOG* readability formula.

All of these traditional formulas are still being used to build text readability classification tools because they are simple and easy to compute. However these traditional formulas have significant drawbacks. They assume that texts do not contain noise and the sentences in the texts are well-formed. However this is not always the case. Traditional formulas require significant sample sizes of text; they become unreliable for a text that contains less than 300 words (Kidwell, Lebanon, and Collins-Thompson 2011). Bruce, Rubin, and Starr (1981), Si and Callan (2001), Petersen and Ostendorf (2009), and Feng, Elhadad, and Huenerfauth (2009) show that these traditional formulas are not reliable. They showed that the formulas rely on a small number of summary text features, which is both a strength and weakness. However, while the formulas are easy to implement, they have a basic inability to model the semantic of vocabulary usage in context. The most important limitation of these formulas is that these measures are based only on surface characteristics of text and ignore deeper properties of texts. They ignore other important factors such as readability comprehension, syntactical complexity, discourse coherence, syntactic ambiguity, rhetorical organization and propositional density of texts. Longer sentences are not always syntactically complex, and counting the number of syllables of a single word does not show word difficulty. Sometimes, shorter sentences can increase reading difficulty (Davison and Kantor 1982). These formulas also ignored the reader's cognitive aptitudes, and prior knowledge and language skills that readers use during interaction with texts. That is why the validity of these traditional formulas for text comprehensibility is often suspect.

Due to the limitation of traditional formulas, along with the opportunity to exploit the growing resources, with recent advancements in NLP and machine learning techniques, researchers inspired to explore richer linguistic properties of texts and combine them with machine learning techniques. This exploration is labeled as an "artificial approach " to readability by François (2009). In the new approach researchers include a variety of linguistically motivated features with sophisticated prediction models based on machine learning. These features and models lead to robust and flexible readability assessment algorithms that have been proposed more recently.

Researchers experimented with the idea of language models from the field of *speech processing* and *machine translation*. The model can be considered as a *word histogram* that gives relative probability of seeing any given vocabulary word in a text (Collins-Thompson 2014). Respective training examples are required in order to build language models. The models consist of *frequency* and *order* of words. Generally, this is a *vocabulary-based* approach where we need to build multiple language models automatically from training example, typically one for each difficulty level. These models capture fine-grained information about vocabulary usage of each word in corresponding training samples. This method provides probability distribution of the candidate document across all language models. The candidate document will get the label of the language model that gives the highest probability. Si and Callan (2001), Collins-Thompson and Callan (2004), Pitler and Nenkova (2008), Schwarm and Ostendorf (2005), Aluisio, Specia, Gasperin, and Scarton (2010), Kate, Luo, Patwardhan, Franz, Florian, Mooney, Roukos, and Welty (2010), and Eickhoff, Serdyukov, and Vries (2011) use statistical language models for text readability classification. Si and Callan (2001) have used *unigram* language models for each readability class to capture content information from scientific web pages. *Sentence length* was combined with these language models to build a linear model. Their experimental results show that the average SL is a good feature but the average number of syllables per word does not perform as well as they expected. Collins-Thompson and Callan (2004) and Pitler and Nenkova (2008) have found that the *unigram* language model is a strong predictor of readability. Schwarm and Ostendorf (2005), Aluisio, Specia, Gasperin, and Scarton (2010), Kate, Luo, Patwardhan, Franz, Florian, Mooney, Roukos, and Welty (2010), and Eickhoff, Serdyukov, and Vries (2011) use higher order language models. The motivation for using a higher order language model is that a probabilistic model provides a prediction of how likely a given sentence can be generated by the same underlying process that generated a corpus of different readability classes. They show that *trigrams* are more informative than *bigram* and *unigram* models. Combining information from statistical language models with other features using the SVM outperforms traditional readability formulas.

Kireyev and Landauer (2011) and Landauer, Kireyev, and Panaccione (2011) proposed a new approach to measure readability of texts that is similar to statistical language models. The approach is called *Word Maturity* (WM). This is a computational model of the development of the meaning of individual words and paragraph with increased exposure to text (Landauer, Kireyev, and Panaccione 2011). The model measures how knowledge of word meaning evolves toward that of literate adults despite the spelled "word forms" remaining unchanged (Landauer and Way 2011). Additionally, it also accounts for how the word's *usage in context* changes over time. The WM model uses *Latent Semantic Analysis* (LSA) in order to model the richness of context in which a word appears over time kireyev:landauer:2011.

With the recent development of NLP tools many linguistics features have been explored to measure text difficulty. These features are extracted from the result of different linguistic processing such as *POS-tagging*, *syntactical parsing* and *semantic parsing*. Schwarm and Ostendorf (2005) found that the feature selection based on the word-POS model performs better than word-based models. POS-based features were also shown to be useful in readability classification (Feng, Elhadad, and Huenerfauth 2009; Aluisio, Specia, Gasperin, and Scarton 2010; Feng, Janche, Huenerfauth, and Elhadad 2010). Primarily, they have focused on five major classes of words: *nouns*, *verbs*, *adjectives*, *adverbs* and *prepositions*. Generally, the count of different POS-tags in the sentence level or document level are considered as an input of different machine learning algorithms. POS based features outperform language model based and syntactic features in (Feng, Janche, Huenerfauth, and Elhadad 2010). Among different POS-based features, *nouns* have the most predictive power (Feng, Janche, Huenerfauth, and Elhadad 2010). The reason could be that the number of common nouns gives an approximation of the number of *entities* in a text. Readers have to keep these *entities* in memory to understand the text.

Gibson (1998) stated that the syntactic complexity leads to longer processing times in comprehension. This longer processing time might impose difficulty in reading. The state-of-the-art text readability analyzer use a linguistically rich set of features to capture syntactical complexity. These tools use natural language parsers for shallow or deep syntactic analysis of texts. Syntactic readability features are then computed from the parse result of the parser. Schwarm and Ostendorf (2005), Pitler and Nenkova (2008), Heilman, Collins-Thompson, and Eskenazi (2007), Heilman, Collins-Thompson, and Eskenazi (2008), and Ma, Singh, Fosler-Lussier, and Lofthus (2012) provide some empirical evidence where readability of texts were affected by syntactic constructions. Heilman, Collins-Thompson, and Eskenazi (2007) show the effect of syntactical complexity in first language (L1) and second language (L2) learning. Their finding is that syntactical features play a more important role in L2 texts than L1 texts. This is because L1 learners learn language syntax through natural interaction where L2 learners depend on grammar books to learn the syntax of the language. 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. Adding verb phrases in each sentence raises the text complexity. Adult readers prefer to have related clauses grouped together whereas children do not. More than one verb phrase in a sentence increases the sentence complexity. Further, multiple verb phrases in a sentence may indicate the presence of explicit discourse relations in the sentence. Pitler and Nenkova (2008) showed the strongest correlation between text readability and the number of verb phrases. They also demonstrated that average parse tree height is a useful feature for readability classification.

Nowadays, many high-level features have been used in order to measure readability. These features capture discourse and cohesiveness of texts. A cohesive text is less difficult to read (Pitler and Nenkova 2008). Any text that contains well-structured and thematically arranged content, clear chapter and section names should be more readable than a text without that kind of content information (Collins-Thompson 2014). Pitler and Nenkova (2008) showed that discourse relations are strongly associated with perceived text readability. The *Coh-Metrix* tools (Graesser, McNamara, Louwerse, and Cai 2004) uses a variety of linguistic and discourse features for text representation. The latest version of this tool uses 108 features. Some of the features are able to capture *degree of referential cohesion*, *deep cohesion* and *temporality*.

The ultimate goal of reading a text is to understand the (semantic) meaning of the text. Semantic based features become especially important for readers with *intellectual disabilities* (ID) (Feng 2010). Linguistic factors (e.g, discourse) and cognitive factors become vital for this group of readers. On the semantic level, a paragraph that refers to many entities at once burdens the reader since the reader 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 ID (Feng, Elhadad, and Huenerfauth 2009). Their hypothesis is that the complexity of a text for readers with ID is related to the number of entities referred to in the text. If a text contains more entities it will be harder due to the mapping of each entity with a semantic representation and how these entities are related. They have also shown how working memory limits the semantic encoding of new information by readers. (Barzilay and Lapata 2008) show that a text written for adults generally will have more entities than a text written for children.

Researchers also experimented with semantic features like *lexical chains* and *entity grids* (Feng, Janche, Huenerfauth, and Elhadad 2010; Barzilay and Lapata 2008). (Barzilay and Lapata 2008) proposed an entity-grid model to capture entity distribution patterns from sentence to sentence. Each sentence in a text was abstracted by four possible grammatical functions. The task was not motivated by text readability but the model can capture the local coherence of the text. It has been shown that these features are useful for readability classification.

4.2 Related work on other languages

As we have said, English is the dominant language in the field of text readability. Recently, some work has been done for other languages like German, French, Italian, Portuguese and Chinese. However, there are many tools freely available that measure text readability of texts from few well-resourced languages. These tools use traditional formulas in order to calculate the measures. The use of these traditional formulas is questionable due to the fact that languages such as German, French and Italian have different linguistic properties than English. Traditional formulas may not work for texts in a language other than English. Also, some of the traditional formulas are modified in order to cope with linguistic difference between two languages. Douma (1960) modified the *Flesch reading ease* formula (Flesch 1948) to measure readability of Dutch texts (Oosten, Tanghe, and Hoste 2010).

Recently, Brück and Hartrumpf (2007), Brück, Hartrumpf, and Helbig (2008a), and Brück, Hartrumpf, and Helbig (2008b) proposed the *DeLite* readability checker for German texts. A corpus of 510 texts is used that has been human annotated into ten different readability levels. The corpus contains mostly legal texts from municipalities. They have used several features that are able to capture the lexical, syntactic and semantic properties of a text. Hancke, Vajjala, and Meurers (2012) also proposed 155 features for German text readability classification. These features are able to capture the lexical, syntactic and morphological properties of a text. They have used a two class (*easy* vs. *difficult*) corpus with texts collected from the web. Their assumption is that a text that is written for children is easier than a text that is written for adults. It has been shown that these 155 features are able to achieve 89.7% accuracy. We also consider this corpus in this thesis. Our proposed 18 features can achieve accuracy that is comparable with their accuracy using 155 features.

Dell'Orletta, Montemagni, and Venturi (2011) performed a study on Italian text readability classification with a specific view of text simplification. A two level *easy* vs. *difficult* corpus in the news domain is used for this study. The study used many features ranging from traditional to morpho-syntactic, lexical and syntactic features. Syntactic features were the best performing features.

François and Fairon (2012) shows a readability classification of French text as a second language. The corpus is harvested from textbooks for second language learners. A total of 46 textual features were used for this study. They considered *verb tense mood* based features along with lexical, syntactical and semantic features.

Researchers proposed readability measurement formulas for different languages. These formulas are influenced by traditional readability formulas for English. Yuka, Yoshihiko, and Hisao (1988) proposed two readability formulas for Japanese texts. These two formulas considered different factors that might influence readability of texts. The factors include *average number of characters per sentence*, *average number of Roman letters and symbols* and *average number of different Japanese characters*. In a more recent study, Sato, Matsuyoshi, and Kondoh (2008) have built a tool to measure the readability level of Japanese texts using language model-based approach. They used a textbook corpus that consists of 1,478 passages extracted from 127 textbooks of elementary school, junior school, high school and university.

Ghani, Noh, and Yusoff (2014) performed a very recent study on readability analysis of *Arabic* texts. Linguistic features are very good predictors of text difficulty of *Arabic* texts.

Only a few approaches consider the readability of Bangla texts. Das and Roychoudhury (2004) and Das and Roychoudhury (2006) show that traditional formulas proposed by Kincaid, Fishburne, Rodegers, and Chissom (1975) and Gunning (Gunning 1952) work well for Bangla. However, the measures were tested for only seven documents, mostly novels.

Sinha, Sharma, Dasgupta, and Basu (2012) proposed two readability measures that are similar to classical readability measures for English. They conducted a user experiment to identify important structural parameters of Bangla texts. The measures are based on the *average word length*, the *number of poly-syllabic words* and the *number of consonant–conjuncts*.

All of the previous works listed above used various types of features to measure readability of texts. However, all of them used texts from one language.

In this thesis, we considered traditional, lexical, linguistic and information-theoretic features. The baseline system we built here using five traditional formulas described above. We build a classifer using 40 linguistic features in order to compare our proposed features with state-of-the-art features. For this comparison we use the English Wikipedia corpus and the English textbook corpus. Some of the linguistic features (e.g., semantic-based features) are used for the first time for text readability analysis. Finally, we propose a list of lexical and information-theoretic features. The experimental results show that the proposed features are not only useful for English corpora but also useful for the Bangla textbook corpus and the German GEO/GEOLino corpus.

5 Readability Corpora

There is no standard corpus for text readability analysis. A human-annotated data set is the prerequisite for machine learning based text readability classification. One of the goals of this thesis is to provide features that are language independent and suitable for text readability classification. Along with these features, this thesis also provides three corpora for text readability classification.

In this thesis, we use four corpora from three different languages: English, Bangla and German. The English Wikipedia corpus is compiled from the English Wikipedia. The second one, the Bangla textbook corpus, is compiled from textbooks that are being used to teach in public schools in Bangladesh. The third one, the English textbook corpus, is compiled from the same source as the Bangla textbook corpus. The *German readability corpus* was collected by Hancke, Vajjala, and Meurers (2012). The following subsections describe the corpora in detail.

5.1 English Wikipedia readability corpus

For English text readability analysis, the *Weekly Reader*[1] is used as a gold standard corpus (Hancke, Vajjala, and Meurers 2012). However, we could not use that corpus because a license is required in order to use the resource for the purpose of research. Our goal was to compile the corpus from a freely available source so that we can provide the resource freely. Wikipedia is a product of collaborative crowd-sourcing, and is freely available.

The *Wikimedia Foundation* used a tool called the *article feedback tool* [2] from September 2010 to early 2013 (Flekova, Ferschke, and Gurevych 2014). The *article feedback tool* is an extension of Wikipedia's interface to engage readers of an article in order to submit their feedback about the article to editors. There are several versions of the *Wikipedia Article Feedback Tool.* But, the data used in our study was collected when

[1] `www.weeklyreader.com`

[2] Wikipedia article about the tool: `http://en.wikipedia.org/wiki/Wikipedia:Article\_Feedback\_Tool`

version 4 of the tool was active. Using the tool, a reader can rate an article in terms of *trustworthy*, *objective*, *complete* and *well-written*. The readability measure is reflected by the *Well-written* feature of the feedback tool, which places a document in a class (one to five stars: incomprehensible, difficult to understand, adequate quality, good clarity, exceptional clarity). Figure 5.1[3] shows a screen-shot of the feedback tool (version 4).

Articles in Wikipedia differ with respect to their quality. Many of them are well written, while for others, the authors have not even reached an agreement among themselves regarding the topic. According to Wikipedia[4], around 40,000 ratings were submitted everyday, 97% of them by anonymous users. 90% of users claim that the page ratings are useful. It should be noted that the *well-written* score might be affected by contributors' perception of the non-linguistic aspects of the article, such as the objectivity or bias of the article. But, as a corrective, the article ratings are evaluated by a group of experts of the corresponding subject areas. A recent study by Giles (2005) showed that many of Wikipedia articles' accuracy are comparable to Wikipedia's rival, the Encyclopedia Britannica. Halfaker, Keyes, and Taraborelli (2013) did a very recent study on readers and editors engagement in Wikipedia. In line with the collaborative nature of Wikipedia, this provides both primary data of a readability assessment together with a gold standard. In any event, the users' quality statements have to be pre-processed appropriately.

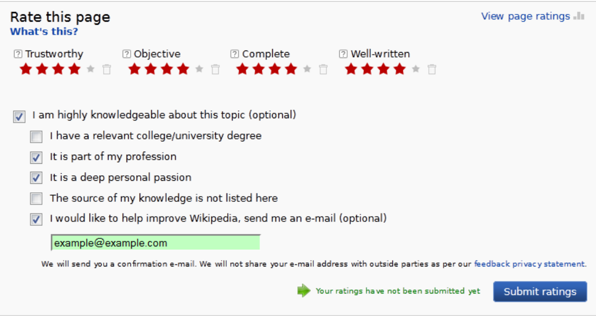
The *Wikimedia foundation* provides feedback data as *raw rating data* and *article summary data*. The *raw rating data* shows the raw ratings of an article. These ratings include raw information about the article, the reader, the reader's background, the rating and the time the rating was submitted. The ratings were submitted on a scale of $1 - 5$ with 0 meaning that no rating was submitted. An article with a 1 rating is the least readable and an article with a rating of 5 is the most readable one. The *article summary data* summarizes this information. It provides *the sum of the rating submitted* and *the total number of users who rated the article*. The *article summary data* is taken into account in the text extraction process, because it helps us to put a threshold on the process of text extraction. From this article summary only *aap_total_readable* and *aap_count_readable* were considered to extract an article.

The *Wikimedia foundation* offers free copies of each Wikipedia as a database dump. The English Wikipedia dump is available in the form of *SQL* and *XML* dumps. To build a readability corpus, we used the *article summary data* of September 19, 2011[5] and an

[3] Source: `http://en.wikipedia.org/wiki/File:Aft-version4.png`

[4] `http://en.wikipedia.org/wiki/Wikipedia:Article\_Feedback\_Tool/Version\_5`

[5] `http://dumps.wikimedia.org/other/articlefeedback/aap\_combined-20110919.csv.gz`



Rate this page

What's this?

View page ratings

Trustworthy	Objective	Complete	Well-written
★★★★☆	★★★★☆	★★★★☆	★★★★☆

☑ I am highly knowledgeable about this topic (optional)

- □ I have a relevant college/university degree
- ☑ It is part of my profession
- ☑ It is a deep personal passion
- □ The source of my knowledge is not listed here
- ☑ I would like to help improve Wikipedia, send me an e-mail (optional)

example@example.com

We will send you a confirmation e-mail. We will not share your e-mail address with outside parties as per our feedback privacy statement.

Your ratings have not been submitted yet

Figure 5.1: Wikipedia article feedback tool (version 4).

XML dump of the English Wikipedia from August 2011.

Each article contains a great deal of extra information such as *infobox*, *figure*, *tables*, *hyperlinks* and more. We used only the texts from each article in order to build a corpus for readability analysis. To extract Wikipedia articles, we utilized a freely available extraction tool[6]. Article length is a factor when extracting text features for classification, as well as the number of evaluations. The rating of an article that is evaluated by many readers is more reliable than the rating of an article that is evaluated by a small number of readers. These two factors were taken into account in the selection of articles for text extraction from Wikipedia: all extracted articles were rated by at least 10 readers and contain at least 10 sentences.

From the *article summary data*, five classes of Wikipedia articles were extracted. Each class corresponds to a readability level (e.g., one, two, three, four and five) in ascending order of readability. The class of an article is determined by a score that is computed by dividing its *total rating score* by the *number of ratings submitted*. We consider four different classes: *very easy*, *easy*, *medium* and *difficult*.

If the score is less than 1, the article falls under the class *Difficult*; if it is greater than 1 and less than 2, the article falls into the next readability class. After the article extraction process, only 12 documents got a rating of less than 1. That is why any article with a rating of less than 2 was assigned to the class *Difficult*. Any article with a rating from 2 to 2.9 was assigned to the class *Medium* and so on. An article with a rating of more than 4 was assigned to the *Very easy* class. Table 5.1 shows the corpus statistics. The Avg. document length (DL) shows the average number of sentences in a document. The average number of words in a sentence is represented by avg. SL. The avg. WL shows the average number of characters in a word.

[6] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

Classes	Documents	Avg. DL	Avg. SL	Avg. WL
Very easy	205	175.27	76.15	5.14
Easy	208	164.51	72.96	5.31
Medium	168	81.98	64.41	5.31
Difficult	60	36.40	54.25	5.17

Table 5.1: Statistics of the English Wikipedia corpus.

5.2 Bangla readability corpus

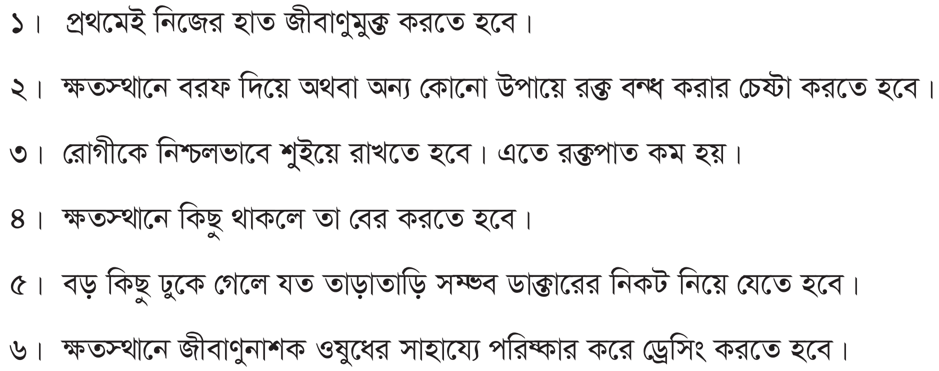
The education system of Bangladesh has three major stages-primary, secondary and higher education. The duration of primary education is a 5 years and secondary education is a 7 years with three sub-stages: 3 years of junior secondary, 2 years of secondary and 2 years of higher secondary. Generally speaking, all textbooks are flawless and written in an easy and lucid language, directed towards creating an interest in the students. The government agency *National Curriculum and Textbook Board* (NCTB)[7], Bangladesh makes textbooks available for primary and secondary education. It is mandatory to use these textbooks for teaching in all of the public schools in Bangladesh.

The textbooks cover many different subjects, including Bangla literature, social science, general science and religious studies. The government agency provides books 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, Habib, and Khan 2007) available but it is unable to extract text from these scanned PDF books. The tool requires training that is very time consuming. Therefore, we only considered textbooks that were converted from typed text. The *Apache PDFBox*[8] was used to extract text from PDFs. Note that 51 textbooks were selected from class *two* to class *ten*. Textbooks such as English, Math, Chemistry, Physics were not considered because these texts might not contain descriptive Bangla texts. After text extraction, it was observed that the text was not written in Unicode Bangla. Figure 5.2 shows a clipped portion from an original pdf book. The extracted text is shown in Figure 5.3. The texts are not readable. A non-standard Bangla input method called *Bijoy* (Islam 2008) is used to type these textbooks. This is an ASCII based method for writing Bangla that was widely used in the 1990s. It is a challenging task to convert the non-standard ASCII texts to Bangla Unicode texts.

The *Bijoy* includes a variety of fonts. The *code point* of some Bangla characters vary

[7] `http://nctb.gov.bd/book.php`

[8] `http://pdfbox.apache.org/`



১। প্রথমেই নিজের হাত জীবাণুমুক্ত করতে হবে।

২। ক্ষতস্থানে বরফ দিয়ে অথবা অন্য কোনো উপায়ে রক্ত বন্ধ করার চেষ্টা করতে হবে।

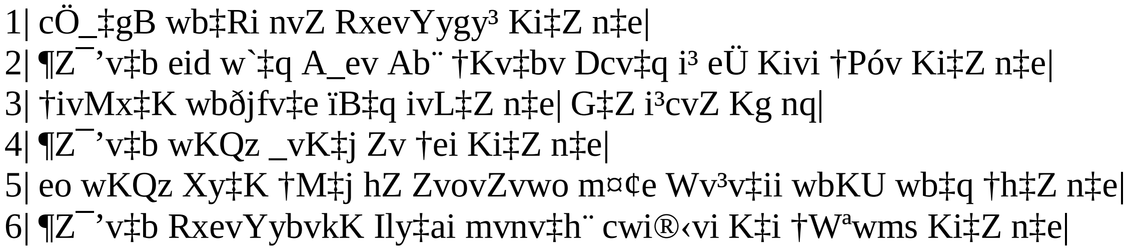
৩। রোগীকে নিশ্চলভাবে শুইয়ে রাখতে হবে। এতে রক্তপাত কম হয়।

৪। ক্ষতস্থানে কিছু থাকলে তা বের করতে হবে।

৫। বড় কিছু ঢুকে গেলে যত তাড়াতাড়ি সম্ভব ডাক্তারের নিকট নিয়ে যেতে হবে।

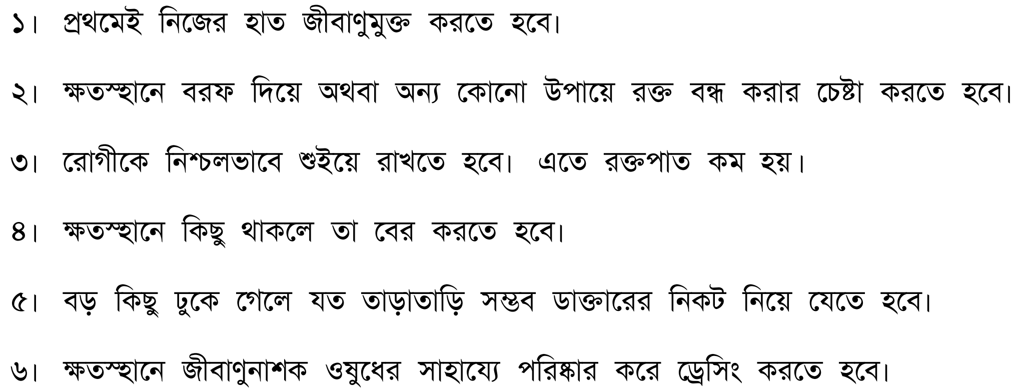
৬। ক্ষতস্থানে জীবাণুনাশক ওষুধের সাহায্যে পরিষ্কার করে ড্রেসিং করতে হবে।

Figure 5.2: Screen shot of an original Bangla pdf book



1| cÖ_‡gB wb‡Ri nvZ RxevYygy³ Ki‡Z n‡e|
2| ¶Z¯’v‡b eid w`‡q A_ev Ab¨ †Kv‡bv Dcv‡q i³ eÜ Kivi †Póv Ki‡Z n‡e|
3| †ivMx‡K wbðjfv‡e ïB‡q ivL‡Z n‡e| G‡Z i³cvZ Kg nq|
4| ¶Z¯’v‡b wKQz _vK‡j Zv †ei Ki‡Z n‡e|
5| eo wKQz Xy‡K †M‡j hZ ZvovZvwo m¤¢e Wv³v‡ii wbKU wb‡q †h‡Z n‡e|
6| ¶Z¯’v‡b RxevYybvkK Ily‡ai mvnv‡h¨ cwi®‹vi K‡i †Wªwms Ki‡Z n‡e|

Figure 5.3: Bangla texts (as Figure 5.2) after extraction



১। প্রথমেই নিজের হাত জীবাণুমুক্ত করতে হবে।

২। ক্ষতস্হানে বরফ দিয়ে অথবা অন্য কোনো উপায়ে রক্ত বন্ধ করার চেষ্টা করতে হবে।

৩। রোগীকে নিশ্চলভাবে শুইয়ে রাখতে হবে। এতে রক্তপাত কম হয়।

৪। ক্ষতস্হানে কিছু থাকলে তা বের করতে হবে।

৫। বড় কিছু ঢুকে গেলে যত তাড়াতাড়ি সম্ভব ডাক্তারের নিকট নিয়ে যেতে হবে।

৬। ক্ষতস্হানে জীবাণুনাশক ওষুধের সাহায্যে পরিষ্কার করে ড্রেসিং করতে হবে।

Figure 5.4: Bangla Unicode texts (as Figure 5.2) after conversion

Misspelled word	Correct spelling
◌ুদই	দুই
◌্লিস্ভ	স্লিভ
ত্রস্◌ীর	স্ত্রীর
প্রস্◌াব	প্রস্তাব

Table 5.2: Examples of conversion problems.

in different fonts. A font called *SutonnyMJ* is used to write these books. The conversion task became more complicated when we discovered that there are several versions of *SutonnyMJ* available and within these versions the code point of some *consonant conjuncts* are different. We used a freely available open source CRBLPConverter[9] to convert these non-standard Bangla texts to Unicode. To cope with the font and its different versions, we needed to modify the CRBLPConverter. The Bangla Unicode text of the original pdf text shown in Figure 5.2 is shown in Figure 5.4.

The conversion was not straightforward; it produced several kinds of noise in the converted texts. The original textbooks not only contain descriptive texts but also poems, religious hymns, texts from other languages (e.g., Arabic, Pali) and transcriptions of Arabic texts (e.g., Surah). These non-descriptive texts were manually removed. The conversion process left many sentences with unexpected spaces between letters, which we also removed manually. Further, we detected many spelling errors possibly produced by the conversion. The converter mostly affected dependent vowels or consonant conjuncts. In Bangla, a dependent vowel can only occur with a consonant. A consonant conjunct becomes invalid due to the misplacement of the Bangla *virama* sign *hoshonto*. Table 5.2 shows some spelling errors generated by the converter. All these mistakes have been manually corrected. There were many English terms in the original Bangla texts that were converted to meaningless strings. For example: the English word *hepatitis* was converted to ঈৎঁফবনরৎঃযৎধঃব. The converter cannot identify the English alphabet because the code points of the Bangla alphabet in *Bijoy* are similar to the code points of the English alphabet. These kinds of errors require manual efforts to correct them.

The readers of the compiled corpus are not adults. That is why it would necessary to consider age group in order to categorize the corpus. Generally, in Bangladesh start children going to school when they are 6 years old and finish the grade *ten* when they are

[9] I am one of the co-developers of this tool. The tool is available at: `http://sourceforge.net/projects/blp/files/CRBLPConverter/`

Classes	Docs	Avg. DL	Avg. SL	Avg. WL
Very easy	234	88.28	7.46	5.27
Easy	113	150.46	9.09	5.27
Medium	201	197.08	10.35	5.47
Difficult	113	251.30	12.19	5.66

Table 5.3: Statistics of the Bangla textbook corpus.

fifteen (Arends-Kuenning and Amin 2004). Duarte Torres and Weber (2011) proposed different categories of children based on their age. The categorized list is related to our study. However, our categorized list is different from their. The corpus is categorized as the following age ranges:

1. early elementary: $7-9$ years old
2. readers: $10-11$ years old
3. old children: $12-13$ years old
4. teenagers: $14-15$ years old
5. old teenagers: $16-18$ years old
6. adults: above 18 years old

The corpus is categorized as the top four categories. Our classification of the corpus distinguishes similar classes to the English Wikipedia corpus: *very easy*, *easy*, *medium* and *difficult*. Documents of (school) grade *two*, *three* and *four* are included into the class *very easy*. Class *easy* covers texts of grade *five* and *six*. Texts of grade *seven* and *eight* were subsumed under the class *medium*. Finally, all texts of grade *nine* and *ten* were mapped onto the class *difficult*. Table 5.3 shows the classes and their statistics. Note that the original grades could not be used as target classes due to problems of data sparseness.

5.3 English textbooks corpus

The government agency NCTB[10], Bangladesh also provides textbooks in English. These textbooks are translations of Bangla textbooks. Most of these books come as PDF. The

[10] http://nctb.gov.bd/book.php

Classes	Documents	Avg. DL	Avg. SL	Avg. WL
Very easy	103	109.67	9.14	4.61
Easy	120	169.10	11.25	4.69
Medium	179	189.47	12.75	4.76
Difficult	117	256.12	14.59	4.90

Table 5.4: Statistics of the English textbook Corpus.

PDF production process is similar to the production of the Bangla texts: some of the books are just scanned where others are typed first then converted to the PDF format. Again we only selected books that were typed. The *PDFBox* toolbox is used to extract text from these PDFs. We did not use scanned books due to the font used to write these books. A proprietary font was used to type these books which creates a lot of garbage in the extracted text. Although we did not consider the scanned books, the extracted text required human efforts in order to make the texts usable for this study.

Only 48 of the books were found to be usable for the extraction of texts. To classify texts into different readability levels, we followed the same idea as we followed for Bangla. Table 5.4 shows the English textbooks corpus for Bangla speakers. The readers of this corpus learn English as their second language. That is why this corpus could be useful for exploring different features for L2 learners.

5.4 German GEO/GEOLino corpus

There is no freely available resource from which we could compile a corpus for German text readability classification. The only German readability corpus available is the *DeLite* (Brück, Hartrumpf, and Helbig 2008a). However, the corpus contains legal texts from the municipal domain, such as city ordinances (Brück, Hartrumpf, and Helbig 2008b). In this thesis, we use a German readability corpus that is collected by Hancke, Vajjala, and Meurers (2012). The idea of using the corpus is that they use similar kinds of techniques for readability classification. The corpus contains texts that belong to two difficulty levels: *difficult* and *easy*. The corpus is compiled from the two websites of German magazines GEO[11] and GEOlino[12] (Hancke, Vajjala, and Meurers 2012). Both of the magazines publish news in similar topics ranging from *nature* and *culture* to *science*. However, the latter magazine targets children from age eight to fourteen (Hancke, Vajjala, and Meurers 2012). Their assumption was that texts that are written for chil-

[11] Web page of the GEO magazine `www.geo.de`

[12] Web page of the GEOline magazine `www.geolino.de`

Classes	Documents	Avg. DL	Avg. SL	Avg. WL
Easy	1627	34.09	15.30	5.63
Difficult	2977	39.26	18.72	6.13

Table 5.5: Statistics of the German GEO/GEOLino corpus

dren are easier to read than texts that are written for adults. The corpus contains 4604 documents of two difficulty classes. The characteristics of the different average values in the German readability corpus are similar to the Bangla corpus. Table 5.5 shows the statistics of the German readability corpus.

6 Representation of Readability Measurement

In this chapter, we describe features that are used in this thesis for *text readability* and *source and translation* classification. Some of the *lexical* and *linguistic* features have been proposed in previous studies. However, we use these features in order to compare with our proposed features. Five traditional readability formulas are used to build a baseline system. These formulas use lexical features that are also representative of some of the translation properties. Although we describe a variety of features, finally we propose just 18 features that are language independent and do not require any kind of linguistic pre-processing. The following sections describe different subsets of features.

6.1 Traditional readability formulas

Sherman (1893) noticed (See (DuBay 2004)) that the average SL of English texts changes over time. During the *Elizabethan* time the average SL of English texts was 45, while in the *Victorian* period the length was reduced to 29. In Sherman's time, the average slipped further to 23. Currently, the average sentence length of English texts is 20 (DuBay 2004).

There are many readability formulas available starting in the early twentieth century. Gunning (1952) proposed one of the earliest readability formulas. His aim was to identify the grade level of texts from the *newspaper* and *business writing* domains. The reading difficulty of a text is mostly influenced by average SL and the amount of DW in the text (Gunning 1952). Any word that has two or more syllables is considered to be a *difficult word*. The proposed index estimates the years of formal education required to understand a text on a first reading. A text with an index below 8 will be understood by most of the readers.

Dale and Chall (1995) proposed a similar formula that also uses average SL and percentage of DW. However the definition of a DW is different than the definition used by Gunning (1952). Dale and Chall proposed a list of 3,000 words that are mostly understood by a fourth grade student. Any word that is not in the list will be considered to be a difficult word. In order to apply this formula, it is advised to select several chunks with 100 words. The formula produces a score that shows the difficulty of a text. A text with a score below 5 will be understood by an average 4_{th} grade student and a text with a score of 10 and above will be understood by a collage graduate.

Kincaid, Fishburne Jr, Rogers, and Chissom (1975) proposed a readability formula that was used by the United States (US) Army to assess the difficulty of its technical manuals. They proposed two different formulas to measure *reading ease* and *grade level.* The *reading ease* formula returns a score between 0 and 100. A higher score indicates the text is easier to read and understand. The *grade level* formula translates the *reading ease* score to a *US* grade level. The aim is to make the score easier to understand for different stake holders of the readability measurement tools. Both formulas use average *number of words* in a sentence and average *number of syllables* in a word.

McLaughlin (1969) proposed another formula that also uses *syllable* counts. The formula only returns a grade level of the target text. Instead of counting syllables, it counts polysyllabic words. Any word that has three or more syllables is considered as a polysyllabic word. The formula yields 0.985 correlation with human judgment (McLaughlin 1969). Fitzsimmons, Michael, Hulley, and Scott (2010) show that the *SMOG* readability formula should be preferred to measure the readability of health care materials.

Senter and Smith (1967) proposed another readability formula that uses average WL and average SL. This formula was designed for real time monitoring of readability of electric typewriters (Senter and Smith 1967). The formula also gives a score that shows the number of years of education needed to understand a text. All of these formulas are shown below:

- Dale-Chall readability formula (Dale and Chall 1948; Dale and Chall 1995)

$$0.1579\left(\frac{difficult\ \ words}{words}*100\right)+0.0496\frac{words}{sentences}+3.6365 \tag{6.1}$$

- Gunning fog index (Gunning 1952)

$$0.4\left(\left(\frac{words}{sentences}\right)+100\left(\frac{complex\ \ words}{words}\right)\right) \tag{6.2}$$

- Automated readability index (Senter and Smith 1967)

$$4.71\left(\frac{characters}{words}\right) + 0.5\left(\frac{words}{sentences}\right) - 21.43 \tag{6.3}$$

- Flesch reading ease (Kincaid, Fishburne Jr, Rogers, and Chissom 1975)

$$206.835 - 1.015\left(\frac{total\ words}{total\ sentences}\right) - 84.6\left(\frac{total\ syllables}{total\ words}\right) \tag{6.4}$$

- Flesch–Kincaid grade level (Kincaid, Fishburne Jr, Rogers, and Chissom 1975)

$$0.39\left(\frac{total\ words}{total\ sentences}\right) + 11.8\left(\frac{total\ syllables}{total\ words}\right) - 15.59 \tag{6.5}$$

- SMOG readability formula (McLaughlin 1969)

$$1.0430\sqrt{number\ of\ polysyllables\frac{30}{number\ of\ sentences}} + 3.1291 \tag{6.6}$$

DuBay (2004) showed more than 200 traditional formulas. But, the traditional formulas shown above are most widely used for text readability classification. However, these formulas use some lexical features (e.g., average SL, average WL) that are indicators of some translation properties as described in Section 9.1 of Chapter 9. Also note that translators are conscious about the readability of their translated texts. That is why these features could be useful for *source and translation* classification.

6.2 Lexical Features

Most of the traditional formulas use average SL and average WL. Recently, Learning (2001) showed that these are the two most reliable measures that influence readability of a text. In general, there is a correlation between sentence length and syntactic complexity. Longer sentences might have complex syntactic constructions that cause a text to be difficult. Children are not aware of syntax. However, a longer sentence might contain more entities that a child has to keep in mind in order to understand the sentence. A longer sentence that has more than one clause could lead to difficulty while reading. On the other hand, a shorter sentence might have fewer entities and fewer connections between them which lead to easier reading. SL also plays a significant role in translation.

A translator tries to make a translation *explicit* and also *simple*. Translated texts become longer due to the *explicitation*. However, the opposite can happen when a translator tries to make a translation *simple*.

The average WL is another lexical feature that is useful for readability classification. Zipf (1935) showed that the *word length* and *frequency* are inversely proportional. That is, most frequent words are shorter. The reason for this is to maximize efficiency by using short words that take less effort to produce than longer words (Zipf 1935; Zipf 1949). Mahowald, Fedorenko, Piantadosi, and Gibson (2013) conducted a study that behavioral showed that users choose the short form more often in predictive contexts, but use longer forms in a neutral context. However, the short form the same word (e.g., *exam* and *examination*) conveys a lower information content than long forms (Mahowald, Fedorenko, Piantadosi, and Gibson 2013). A long word that contains many syllables is morphologically complex and leads to comprehension problems (Harly 2008). Longer words impose complexity for some readers like children and readers with intellectual disabilities. A word with two or more syllables can be problematic for less experienced readers or people with dyslexia (Harly 2008). For example: the word *biodegradable* will be harder to pronounce, spell and understand for a child. There are two ways to calculate average word length. One is counting characters per word. Another is to count the number of syllables per word. In this study, we measure WL by counting characters, because most of the related works follow this approach. The average word length varies from language to language. The average WL in English is 5.5 (Nádas 1984). The average WL is longer for inflectional and agglutinative languages such as *Turkish*. The average length of Turkish root words is 6.60 (Güngör 2003), which is higher than English. However, the average WL in Turkish will be higher when affixes are added to root words. The average WL is also useful for source and translation classification. The average WL in translated texts is longer than a source text (Frankenberg-Garcia 2009). The average number of difficult words (DW) is related to the average WL. A translated text that contains more difficult words will be harder to comprehend. Crossley, Dufty, McCarthy, and McNamara (2007) showed that a simplified text contains a lower number of sophisticated words (e.g., low frequent words) compared to authentic[1] texts. The syntax of simplified texts is simpler than that of authentic texts.

A text with rich vocabulary (lexical density) could be problematic for some readers. The vocabulary size is a measure of vocabulary richness. There are four different ways to measure vocabulary richness (Wimmer and Altmann 1999). But the most common

[1] Text that are written for native speakers

one is by counting the number of different words in a text and applying the type/token ratio (TTR), which indicates the lexical density (vocabulary) of a text. Apart from representing vocabulary richness, TTR also represents a model of information flow in a text (Wimmer 2008). Low lexical densities involve a great deal of repetition with the same words occurring again and again. Conversely, high lexical density shows the diverseness of a text. A diverse text is supposed to be difficult for readers. In a diverse text different synonyms might be used to represent similar concepts. Temnikova (2012) shows two different problems caused by rich vocabulary that affect text readability. Non-native speakers or non-specialists in the domain could face problems finding the relationship between the synonym and the main term (Temnikova 2012). There are many different versions of TTR formulas available. Carroll (1964) proposed a variation of TTR in order to reduce the sample size effect. This variation of TTR is called corrected TTR. Another version is called Bi-logarithmic TTR (Herdan 1964), in which logarithmic values of types and tokens are considered to calculate the Bi-logarithmic TTR. Guiraud (1960) in (Lu 2012) defines another variation of TTR called root TTR. Vajjala and Meurers (2012) used these three TTR formulas for text readability classification. Köhler and Galle (1993) also defined a version of TTR (see Equation: 6.7) that considers position within a text. In the equation 6.7 x refers to position in the text, t_x = number of types up to position x, T = number of types in the text and N refers to the number of tokens in the whole text. Figure 6.1 shows analysis of Köhler–Gale TTR of the first 100 tokens in four randomly selected documents from four difficulty classes of the English textbook corpus. The outcome is similar to the shown in Wimmer (2008). From the Figure 6.1, it is hard to measure the usefulness of the Köhler–Gale TTR in the readability of texts. Figure 6.2 also shows an analysis of Köhler–Gale TTR in the same documents, but this time whole documents are considered. This time values are flat and discriminating in terms of readability. The Figure 6.2 displays TTR values of the first 100 tokens for a better comparison.

We also propose a version of TTR that focuses on document level TTR as well as sentence level TTR. Generally, TTR at the document level will always be lower than TTR at the sentence level. The formula takes the summation of the deviation of document level TTR from the sentence level TTR. Our hypothesis is that easier texts will have lower deviation TTR than a difficult text.

The TTR has been considered as a useful feature by Pastor, Mitkov, Afzal, and Pekar (2008), Ilisei, Inkpen, Pastor, and Mitkov (2009), and Ilisei, Inkpen, Pastor, and Mitkov (2010) for source and translation classification. TTR is a quantitative measure

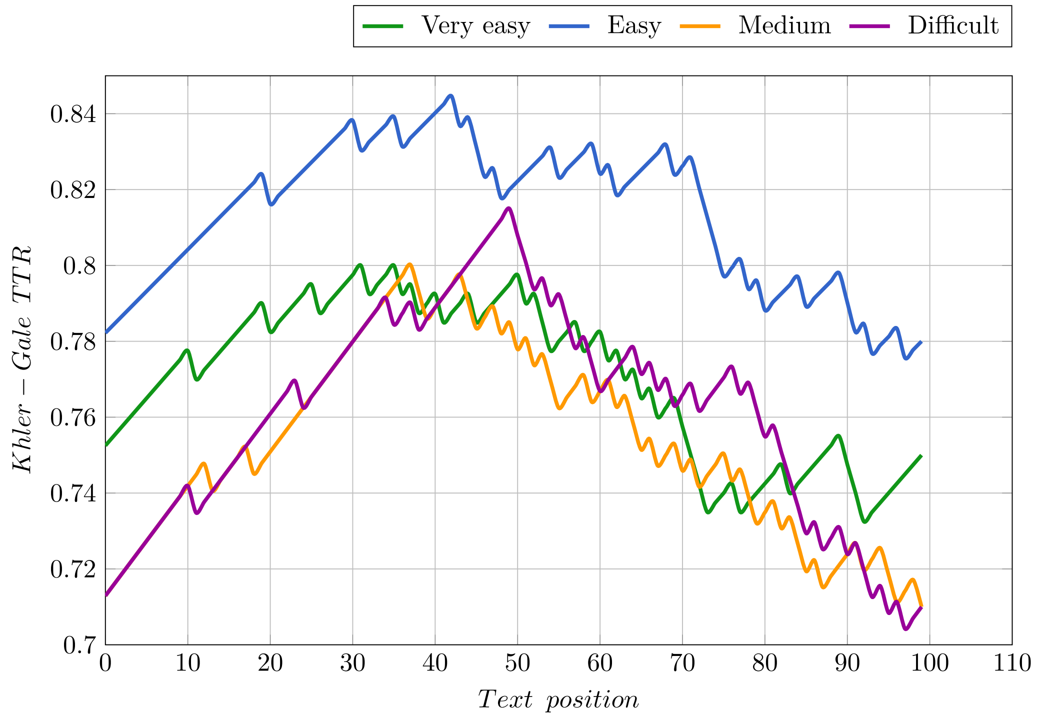


Figure 6.1: Köhler–Gale TTR in four randomly selected documents from the English textbook corpus (only 100 tokens)

for *simplification* translation properties (See Section 9.1 of Chapter 9).

- Köhler–Gale TTR

$$TTR_x = \frac{t_x + T - \frac{xT}{N}}{N} \tag{6.7}$$

- Root TTR

$$\frac{T}{\sqrt{N}} \tag{6.8}$$

- Corrected TTR

$$\frac{T}{\sqrt{2N}} \tag{6.9}$$

- Bi-logarithmic TTR

$$\frac{\log T}{\log N} \tag{6.10}$$

- Deviation TTR

$$\sum_{i=0}^{n} \left(\frac{T}{N} - \frac{t_i}{n_i} \right) \tag{6.11}$$

The TTR does not capture morphological variants, but rather counts morphological

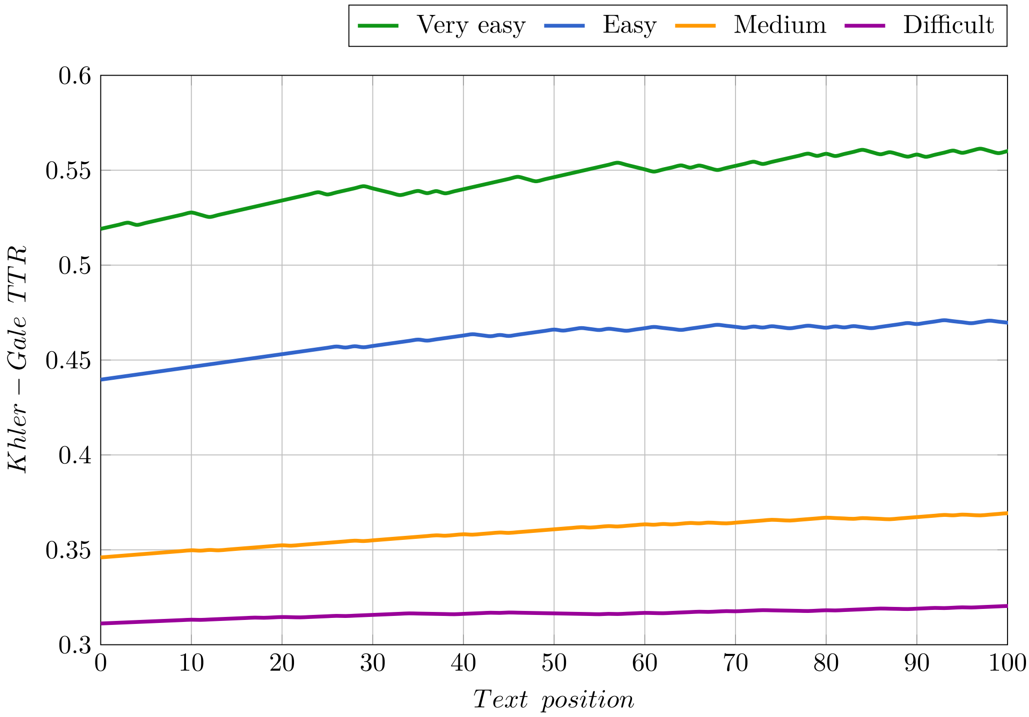


Figure 6.2: Köhler–Gale TTR in four randomly selected documents from the English textbook corpus (whole document)

variants as different word types. For instance, *Book* and *books* will be considered different words and word types even though they are lexically the same. To circumvent this deficit, we compute the TTR on the level of types as lemma for texts from the English corpora. The feature is not as useful as we expected, that is why the feature was discarded from the feature list. All of these features are computed at the sentence level and on the level of the text as a whole. Lexical features are listed in Table 6.1.

6.3 Linguistic features

Readability is influenced by different linguistic properties of a text. The linguistic complexity of a text can be problematic for many readers. While these features have already been used in previous studies, seven of the semantic features are proposed for the first time for text readability classification. We use these features in order to compare our proposed features. We do not consider this feature set for source and translation classification. The reason is that the corpus we use for source and translation classification contains text from 21 European languages. We could not find the required linguistic tools for all of these languages. It should be noted that the performance of this feature

	Features
1	Average SL
2	TTR per document
3	TTR per sentence
4	TTR (lemma level) per document
5	TTR (lemma level) per sentence
6	Average DW per document
7	Average DW per sentence
8	Average WL
9	Köhler TTR
10	Root TTR
11	Corrected TTR
12	Bi-logarithmic TTR
13	Deviation TTR

Table 6.1: Lexical feature set.

set also depends on which tools are used for linguistic processing. Note also that these tools can generate erroneous output.

6.3.1 POS-based features

Heilman, Collins-Thompson, and Eskenazi (2007) and Falkenjack, Mühlenbock, and Jönsson (2013) showed that POS-based features are useful in readability prediction for *English* and *Swedish* texts. For example, the number of named entities in a text can be approximated by the number of words tagged as *noun*. Readers need to co-refer pronouns with appropriate nouns which might impose an extra burden for less-skilled readers. Here the hypothesis is that the more name entities, the higher probabilities that need to be resolved. The number of definite articles provide a measurement of how abstract a text is since an abstract text will have fewer definite articles. In this thesis, we group words based on their POS level.

We focus on seven POS categories (noun, verb, pronoun, adjective, adverb, preposition and determiner) where for each category X, we implement two features such as the *avg. number of* X *in a sentence* and the *number of* X *in a document.* It should be noted that scaling is very different for these features. However, this will be managed by the tool we use (WEKA), in which there is an option available to normalize all feature values. The 14 POS-based features used in this study are listed in Table 6.2.

Features	
1	Average nouns per sentence
2	Number of nouns per document
3	Average verbs per sentence
4	Number of verbs per document
5	Average adjectives per sentence
6	Number of adjectives per document
7	Average adverbs per sentence
8	Number of adverbs per document
9	Average pronouns per sentence
10	Number of pronouns per document
11	Average prepositions per sentence
12	Number of prepositions per document
13	Average determiners per sentence
14	Number of determiners per document

Table 6.2: POS-based feature set.

6.3.2 Syntactic features

A text can be less readable due to unusual linguistic constructions or ungrammatical language that tend to manifest themselves in the syntactic properties of the text. For instance, a sentence with more than one noun phrase imposes an extra burden on readers. That is, readers need to remember multiple noun phrases in order to understand the sentence. As an example, Barzilay and Lapata (2008) found that an article that is written for adult readers contains more noun phrases than an article that is written for children. Multiple verb phrases in a sentence may also play a role in making a sentence difficult to read. More than one verb in a sentence increases the syntactical complexity of the sentence. Also, these kinds of sentences contain explicit discourse relations.

Subordinate clauses, according to (Pitler and Nenkova 2008), correlated positively with text readability. Therefore, a sentence with a subordinate clause will be difficult for children or a less-skilled reader. In our experiment, we focus on *VP*s, *NP*s, *PP*s, *subordinate clauses* and *embedded clauses* (e.g., clauses in argument position). For each of these syntactic classes *XP*, we compute two features: the *avg. number of XP in a sentence* and the *number of XP in a document*. This results in 10 syntactic features. Note that we used the Stanford PCFG parser (Klein and Manning 2003) for parsing Wikipedia articles. Table 6.3 listed syntax-based features.

6.3.3 Semantic features

The semantics of a text plays an important role in text, and influences its readability. A number of semantic indicators of readability, which are not accounted for in related studies, are *co-reference* (Holler and Irmen 2007), *frame semantics* (Fillmore 1982; Alan

Features	
1	Average noun phrases per sentence
2	Number of phrases per document
3	Average verb phrases per sentence
4	Number of verb phrases per document
5	Average prepositional phrases per sentence
6	Number of prepositional phrases per document
7	Average length of subordinate clauses per sentence
8	Number of subordinate clauses per document
9	Average embedded clauses per sentence
10	Number of embedded clauses per document

Table 6.3: Syntax-based features.

2001) and *semantic roles* (Gildea and Jurafsky 2002). The readability of antecedents of anaphoric expressions is subject to the *right frontier constraint* (Polanyi 1988). That is, anaphoric expressions can only be attached to elements that lie on the right hand side of the text tree or graph spanned by rhetorical relations (Mann and Thompson 1988). From this perspective, it should be easier to resolve anaphora that are close to their antecedents in terms of their distances in the text structure graph as spanned, for example, by rhetorical relations (Mann and Thompson 1988). The underlying hypothesis is that the longer the distance between an anaphoric expression and its antecedent, the less readable the text. In order to explore this feature, we use the tool named *Reconcile* (Stoyanov, Cardie, Gilbert, Riloff, Buttler, and Hysom 2010) to annotate Wikipedia articles with co-reference information.

A *semantic frame* is a coherent structure of related concepts (Alan 2001). More specifically, the meaning of a single word cannot be understood without considering its context in the text. A *semantic frame* contains many facts that represent characteristic features, attributes and functions of a referent. As the context has to be considered during the reading of such sentences, a text with more semantic frames will be harder to read for less-skilled readers.

Further, semantic role labeling is a task of shallow semantic parsing where each predicate is mapped onto its semantic roles. Our hypothesis is that the more *semantic frames* that are manifested by a text, the less readable the text is. Semantic roles represent the underlying relationship of a participant with the main verb in a system. It is important to understand the semantic roles in order to understand the text. That is why the number of semantic roles presents difficulties for readers. Named entities are also useful features for readability classification, as shown by Feng, Elhadad, and Huenerfauth (2009).

We have derived 9 semantic indicators of text readability. Note that the *Semafor*

Features	
1	Average named entities per sentence
2	Number of named entities per document
3	Number of co-reference chains in a document
4	Average co-reference chain length
5	Average distance between antecedent and anaphora
6	Average number of semantic frames per sentence
7	Number of semantic frames per document
8	Average number of semantic roles per sentence
9	Number of semantic roles per document

Table 6.4: Semantic-based features.

semantic parser (Das and Smith 2011) is used in order to annotate semantic information in Wikipedia articles. The parser uses FrameNet (Fillmore, Johnson, and Petruck 2003) based annotations. The Stanford named entity recognizer (Finkel, Grenager, and Manning 2005) is used for named entity parsing. All semantic-based features are listed in Table 6.4.

6.3.4 Other features

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, this will raise the difficulty level. Frequent content words in a corpus are considered to be familiar words. A text with more familiar words is easier to read. We extracted a list of frequent words (that occur more than 100 times) from the English readability corpus.

The Simple English Wikipedia[2] uses simple vocabulary and simple syntactic constructions. The vocabulary overlaps between a Wikipedia article and the simple Wikipedia will show the simpleness of an article. Our hypothesis is that a simple document will be more readable than a complex article. The probability of an article derived from an *unigram* model based on the simple Wikipedia indicates the simplicity of this article. This sort of probability is calculated in a fashion similar to the approach presented in (Pitler and Nenkova 2008). An article with a higher probability will be more readable than an article with lower probability. In this class of features, we considered 6 features that are listed in Table 6.5.

[2] Simple Wikipedia: `http://simple.wikipedia.org`

Features	
1	Average Hapax-legomena per sentence
2	Number of Hapax-legomena per document
3	Average number of familiar words per sentence
4	Number of familiar words per document
5	Average number of simple words per sentence
6	Number of simple words per document

Table 6.5: Other linguistic feature set

6.4 Information-theoretic features

Information theory involves in quantification of information. It is a branch of *mathematical theory*, *electrical engineering* and *computer science*. Information theory was introduced in the 1940s by *Claude E. Shannon* in his book *The Mathematical Theory of Communication* (Shannon 1948). The main criterion of the classical information theory is the problem of the transmission of information over a noisy channel. Shannon (1948) explored the problem and showed that on average, the number of bits needed to represent the result of an uncertain event is given by its *entropy*, and this becomes one of the key measures of information.

This key measure is based on *probability* and *statistics*. In this field, there are also other measures that estimate statistical significance of how documents vary with different types of probability distributions. That is, these determine how much information can be encoded from a document using a certain type of probability distribution. The following subsections describe different types of information-theoretic features.

6.4.1 Entropy based features

Brainerd (1976) described variability of a text at the micro and macro level. According to him, readers like to predict the next when we already have read n numbers of word in a text. This predictability makes a text *interesting* or *uninteresting*. In macro level, this variability depends upon the flow of the story while in micro level it depends on the author's linguistic style (Brainerd 1976). A text become *uninteresting* at both levels when the predictability is always high. This predictability is related with the readability of the text. A highly predictable text will be easier to read and follow. The text becomes very hard to follow at the macro level when the predictability is very low. At the micro level the text requires the reader's attention when the predictability is low. It will be easier for a reader to build the cognitive network when a text is more predictable. The degree of predictability can be captured by *entropy* and other similar measures.

The term *entropy* was introduce more than a century and a half ago to study thermodynamics and statistical mechanics (Balasubrahmanyan and Naranan 2005). In the field of thermodynamics *entropy* measure the number of specific ways in which a thermodynamics system can be arranged. This measure is commonly understood as a measure of disorder. This concept was introduced into the field of information theory by Shannon (1948). *Entropy* is a key measure of information that is usually expressed by the average number of bits needed to store or communicate with messages.

Recently, researchers have found that the most efficient way to send information through a noisy channel is at a constant rate (Genzel and Charniak 2002; Genzel and Charniak 2003). That is, the local measure of *entropy* will increase with the number of sentences. The reason for this increase is lexical and non lexical. Plotkin and Nowak (2000) have shown that this principle also correlates with biological evidence of how human language processing evolved. This rule must be retained in any communication in order to make the communication efficient. The texts we consider in this thesis are also media of direct communication between writers and readers. For a translated document, a translator is also involved in the process of communication. Based on the finding of Genzel and Charniak (2002) and Genzel and Charniak (2003), the increasing rate of *entropy* of a good readable document will be different than a less readable document. That is the amount of information will vary among texts that belong to different reading difficulty classes. In order to utilize this information-theoretic notion we start from random variables that represent different properties of texts. Table 6.6 lists all of the different random variables we explored in this thesis. The entropy of a random variable X is defined as

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \tag{6.12}$$

In the Equation 6.12, the more the outcomes of X tend to 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. The validity of our hypothesis is shown in Section 2.2 of Chapter 2. Genzel and Charniak (2002) also suggest that this measurement may be able to capture semantic contextual influence that could be useful to measure reading difficulty.

During the translation process, translators always try to make the translation more readable for the target readers. If there is any complexity in the source texts, translators try to make the translation simple in order to avoid complex structure in the translation. Our hypothesis for source and translation classification is that entropy of a translated

	Features
1	Word probability
2	Character probability
3	POS probability
4	Word length probability
5	Word frequency probability
6	Lemma length probability
7	Lemma frequency probability
8	Character frequency probability
9	POS frequency probability

Table 6.6: Entropy-based feature set

text will be less than the corresponding original text.

6.4.2 Information transmission based features

There is a relation among text difficulty, sentence length and word length. The usefulness of similar lexical features such as SL or number of DW in a sentence is shown in section 6.2. Generally a longer sentence contains more entities that influence the difficulty level. Similar things happen with longer words. But, a sentence becomes even more difficult if it is longer and contains more long words. These kinds of properties can be defined by *joint* and *conditional* probabilities.

In the field of information theory, joint probability measures the likelihood of two or more events occurring together. This probability is the intersection of two more events. Given two random variables X and Y that are defined in the probability space, the joint probability distribution for X and Y will give the probability where each instance of X and Y falls in any particular range or discrete set of values specified for that variable.

The conditional probability gives the probability that the event will occur given the knowledge that another event has already occurred. By considering the joint probability and two random variables X and Y, Shannon's joint entropy can be defined as:

$$H(X,Y) = -\sum_{<x,y>\in XxY} p(x_i,y_i)\log p(x_i,y_i) \tag{6.13}$$

Two conditional entropies can be defined as:

$$H(X|Y) = -\sum_{y\in Y} P(y_i) \sum_{x\in X} p(x_i|y_i)\log p(x_i|y_i) \tag{6.14}$$

	Features
1	SL and WL probability
2	SL and DW probability

Table 6.7: Information transmission-based feature set

$$H(Y|X) = -\sum_{x \in X} P(x_i) \sum_{y \in Y} p(y_i|x_i) \log p(y_i|x_i) \tag{6.15}$$

From the equation 6.12, 6.13, 6.14 and 6.15, it can be shown that:

$$T_s(X,Y) = H(X) + H(Y) - H(X,Y) \tag{6.16}$$

The function is called *information transmission*, and it measures the strength of the relationship between elements of random variables X and Y. Details about this theorem can be found in (Klir 2005). Borst and Theunissen (1999) used this feature to measure the amount of information about the stimulus carried in a neural response. Additionally they have shown how to use this feature to validate simple stimulus-response models of neural coding of dynamic stimuli. In this work, two information transmission based features are used. They are listed in Table 6.7. The *SL and WL* probability shows the relation between SL and WL and *SL and DW* probability shows the relation between sentence length and the number of complex words. Our hypothesis is that a longer sentence with, on average, longer words or many difficult words will be more difficult.

6.4.3 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:

$$D(p||q) = \sum_{i=1}^{n} p(x_i) \log \frac{p(x_i)}{q(x_i)} \tag{6.17}$$

$D(p||q)$ 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, this is a measure of information lost when p is approximated by q. The relative entropy measures the expected number of extra bits needed to code samples from p when using a code based on p. Generally p represents the *true* distribution of candidate data while q represents the model, another distribution or approximation of p.

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, this gives for any random variable X four features of *relative entropy*. It has to be noted that the training and test set have to be separated before calculating this probabilities.

6.5 Proposed features

In this thesis, our goal is to propose a set of features that are language independent and do not require any kind of linguistic pre-processing. And, a classifier using the proposed feature set should achieve reasonable classification accuracy for both applications. We have experimented with more than one hundred features ranging from lexical and linguistic to information theoretic. Some of them have already been explored in previous studies.

Many recent studies are proposing different linguistically motivated features that achieve state-of-the-art classification accuracy based on the readability of a document. However, we can not rely on this type of feature due to the fact that these features require linguistic tools. After observing the performance of different features in both applications, a features list was selected. These are presented in Table 6.8. The feature list contains eighteen features that belong to *information-theoretic* and *lexical* feature classes. Seven of them are *information-theoretic* features. We experimented with several *Kullback-Leibler* divergence-based features. However, these features have not proven useful. It should be noted that most of the *lexical* features that are in the proposed list have been used for similar text categorization tasks.

Features	
1	Average SL
2	TTR per document
3	TTR per sentence
4	Avg. DW per document
5	Avg. DW per sentence
6	Average WL
7	Köhler–Gale TTR
8	Root TTR
9	Corrected TTR
10	Bi-logarithmic TTR
11	Deviation TTR
12	Word entropy
13	Word frequency entropy
14	Word length entropy
15	Character entropy
16	Character frequency entropy
17	Information transmission of SL and WL probability
18	Information transmission of SL and DW probability

Table 6.8: Proposed features for both applications.

7 Experiment of Text Readability

In this section we describe the experiments we performed to show the usefulness of our proposed features for readability classification. In order to compare with different features, we used traditional readability formulas to build a baseline system. We use five traditional formulas. We use these formulas such that the outputs will be used as inputs for the classifier.

We start with the English Wikipedia corpus. Several linguistic features together achieve a satisfactory classification accuracy. However, we can not use these features as they are dependent on language and linguistic tools. Instead of them, we experimented with several *information-theoretic* and *lexical* features. Finally, we selected seven *information-theoretic* and eleven *lexical* features, such that a classier using them achieves a similar accuracy to one using 40 linguistic features. The proposed features are language independent and linguistic tools independent. Later we show that these features not only achieve satisfactory classification accuracy for the English corpora but also perform as expected with the Bangla textbook corpus and the German GEO/GEOLino corpus. We also compare our results with recent works for each language respectively. The following subsections describe the experiments and results of the different features in the four corpora.

7.1 English Wikipedia corpus

7.1.1 Traditional features

Many traditional readability formulas are being used in commercial readability classification tools. The *Flesh reading ease* score is one that is being used in *Microsoft Word* document processing tools (Stockmeyer 2009). The *SMOG* formula is one widely-used formula for measuring readability of texts in the medical domain (Fitzsimmons, Michael, Hulley, and Scott 2010). In total five traditional formulas are used to build the baseline system for the English corpus. Sub-section 6.1 describes these formulas. Figure

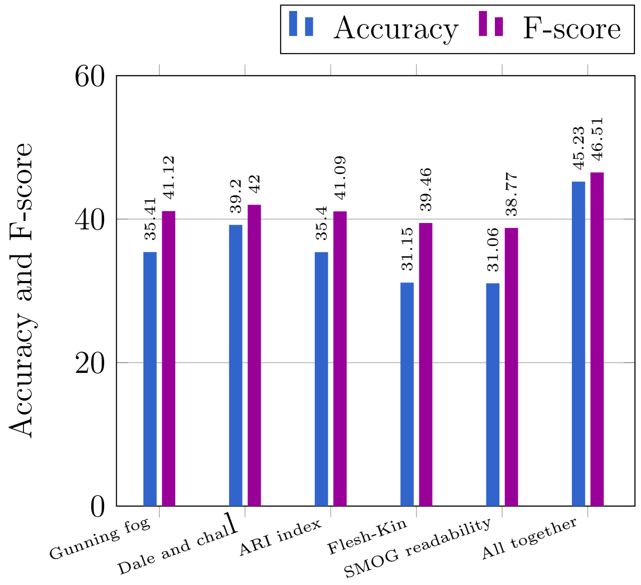


Figure 7.1: Evaluation of traditional readability formulas in the English Wikipedia corpus.

7.1 shows the evaluation of these formulas in the English readability corpus. The table describes individual and group evaluation.

Individually the *Dale and Chall* formula outperforms others. However, all of these formulas together produce 45.23% *F-score*. This is a similar finding to Collins-Thompson and Callan (2004), Feng, Elhadad, and Huenerfauth (2009), and Petersen and Ostendorf (2009) who found that these formulas are not reliable enough.

7.1.2 Lexical features

Traditional formulas use lexical features such as: average SL and average WL. Lexical features indicate some degree of linguistic complexity of texts (See subsection: 6.2). Figure 7.2 shows the evaluation. The results show that lexical features are a strong predictor of the readability of *English* texts. Document level TTR related features perform better than other features. However, these same features at the sentence level are among the worst performers. One noticeable observation is that document level features are better predictors than sentence level features. That means that text readability is influenced by document length. It is important to normalize features that are calculated at document level. We relied on the tool we use to build the classifier. The tool uses a filter to normalize the training data. Although average SL and average WL are used in most of the traditional features, these two do not perform as expected. Table 5.1

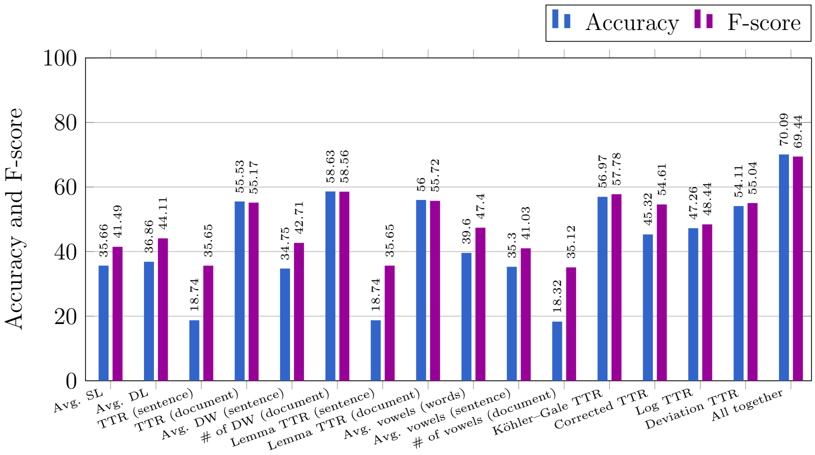


Figure 7.2: Evaluation of lexical features in the English Wikipedia corpus.

shows that good readable articles are longer than less readable article. However, this characteristic is not reflected in the evaluation. All together 15 lexical features achieve around 70% accuracy in classification.

7.1.3 POS-based features

The evaluation of POS-based features in the *English* corpus is presented in Figure 7.3. The figure shows that the combination of POS-based features also achieves good classification results. *Nouns*, *adverbs* and *pronouns* perform better than other POS classes investigated in this work. It has to be noted that the class, *nouns* is the best performing feature among all POS-based features. This is similar to the findings of Heilman, Collins-Thompson, and Eskenazi (2007) and Falkenjack, Mühlenbock, and Jönsson (2013). Feng, Janche, Huenerfauth, and Elhadad (2010) showed that content words (e.g., nouns, verbs, adjectives and adverbs) have higher predictive power than function words (e.g., pronouns, prepositions, grammatical articles). However, the experimental results show that function words like *pronouns prepositions*, and *definite articles* also have good predictive power for measuring text readability. All together the 14 POS-based features perform better than 15 lexical features.

We use an in-house POS tagger called the TTLab pre-processor[1] (Waltinger 2010; Brück, Mehler, and Islam 2014) that is based on the TnT tagger (Brants 2000) to tag the English readability corpus.

[1] The link of the pre-processor: `http://api.hucompute.org/preprocessor/`

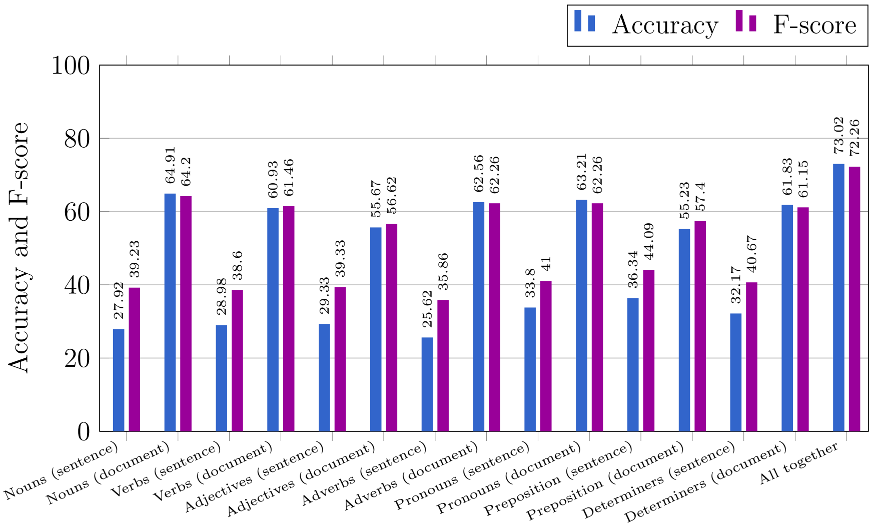


Figure 7.3: Evaluation of POS-based features in the English Wikipedia corpus.

7.1.4 Syntax-based features

There is a correlation between readability difficulty and syntactical complexity. Readability of a document is also influenced by structural complexity. The number of *verb phrases* show the number of clauses in a sentence. A sentence with more clauses will be more difficult than a sentence with one clause. However, the *noun phrases* are more influential than verb phrases. That confirms similar statements made by Barzilay and Lapata (2008), i.e, that an article aimed for children contains fewer *noun phrases* than an article for an adult. We have gotten similar results, that *noun phrases* at the document level are the best performing syntax-based features. Figure 7.4 shows the evaluation of syntactical features. A sentence with subordinate clauses is difficult for *less skilled* readers (Pitler and Nenkova 2008). However, Figure 7.4 shows that other syntactical features are more important than subordinate clauses. We use the Stanford PCFG parser (Klein and Manning 2003) to parse Wikipedia articles. We assume that a parsing error generated by the parser is less likely.

7.1.5 Semantic-based features

A text is said to be easier for readers when the underlying meaning of the text is easier to understand. The complexity of the meaning is measured by different semantic features. That is why reading difficulty could be influenced by different semantic entities of an article. Figure 7.5 shows the evaluation of semantic based features. Feng, Elhadad, and

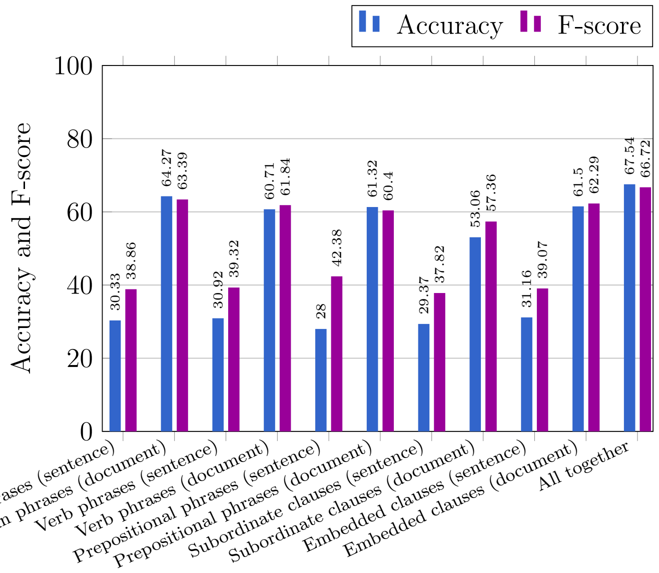


Figure 7.4: Evaluation of syntax-based features in the English Wikipedia corpus.

Huenerfauth (2009) noted that a named entity is a cognitively motivated feature. An article with many named entities is difficult for readers with intellectual disabilities. But that is not reflected by our findings. The *Average co-reference chain length* represents the average number of noun phrases that refer to the same entity. Our hypothesis was that a document with a longer co-reference chain would be more difficult. Readers have to keep these entities in memory in order to understand the relation between them. However, the results in Figure 7.5 do not support this hypothesis. As we stated earlier, 7 of the semantic features we propose here are being used for readability classification for the first time. The *number of semantic roles per document* is one of the best performing features. All of the semantic related features have good predictive capabilities regarding reading difficulty. Note that the *Semafor* semantic parser of (Das and Smith 2011) is used in order to annotate semantic information in Wikipedia articles. The parser uses FrameNet (Fillmore, Johnson, and Petruck 2003) based annotations. The Stanford named entity recognizer (Finkel, Grenager, and Manning 2005) is used for named entity parsing.

7.1.6 Other linguistic features

Figure 7.6 shows the evaluation of some other features. The *number of familiar words per document* is one of the best performing individual features. That is, an article with more known (frequent) words is more readable. Also, the field of cognitive science shows

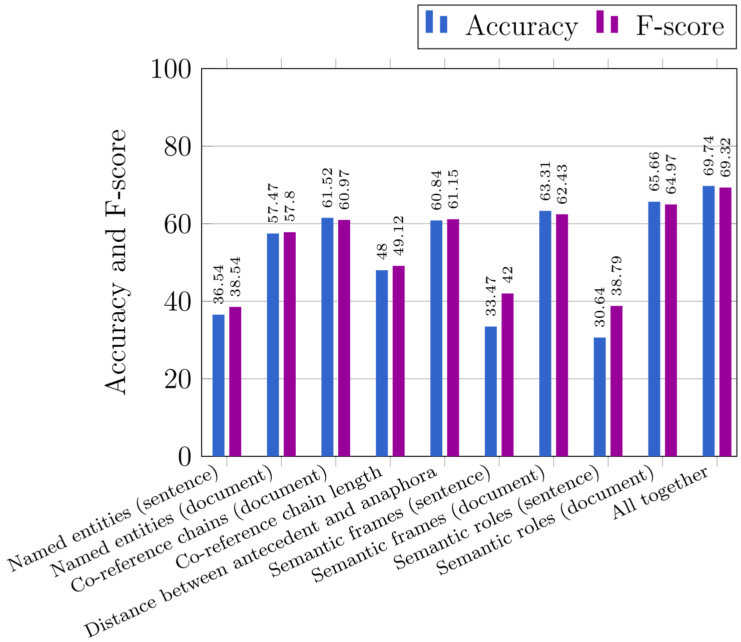


Figure 7.5: Evaluation of semantic-based features in the English Wikipedia corpus.

that eye movements are faster in a text with more familiar words. Vocabulary overlap between a Wikipedia article and the *Simple Wikipedia* unigram model is also a good indicator of readability. The simple Wikipedia can be used to build a language model for a similar kind of task. Figure 7.7 shows the combined result of all linguistic features, and demonstrates that the 40 linguistic features we used give better results than those reported by many previous research projects (as noted in the related work section, See Chapter 4). A classifier with lexical and linguistic features can achieve around 75% classification accuracy.

7.1.7 Entropy-based features

As noted earlier, entropy measures the amount of information in an article. The entropy rate is constant in human communication (Genzel and Charniak 2002; Genzel and Charniak 2003). More specifically, the local measure of entropy will increase with the number of sentences. Wikipedia's articles are assumed to be a medium of communication (i.e., an information transfer) between Wikipedians and readers. We assume that the increasing rate of entropy will be different in more readable text than a less readable one. As a single feature, these entropy-based features perform similarly to some of the linguistic features and produce an F-score around 50%. However, collectively the performance

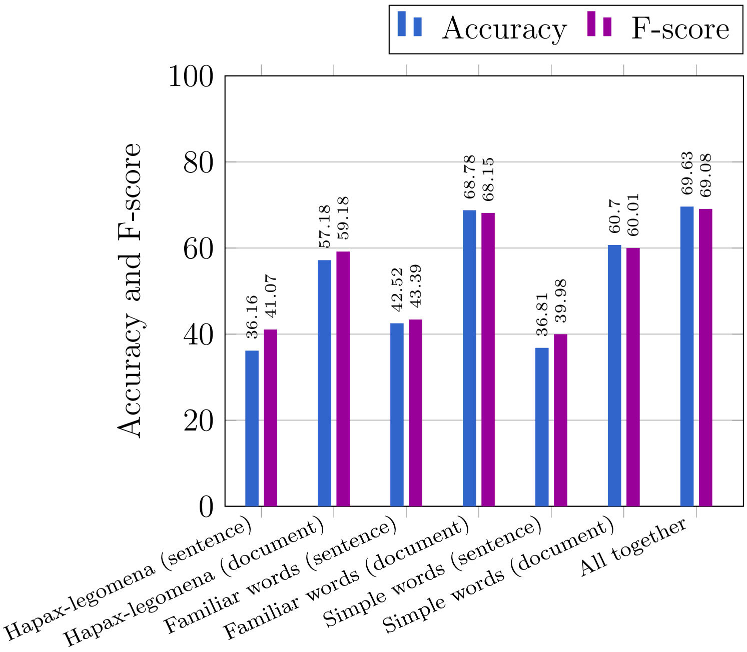


Figure 7.6: Evaluation of other features in the English Wikipedia corpus.

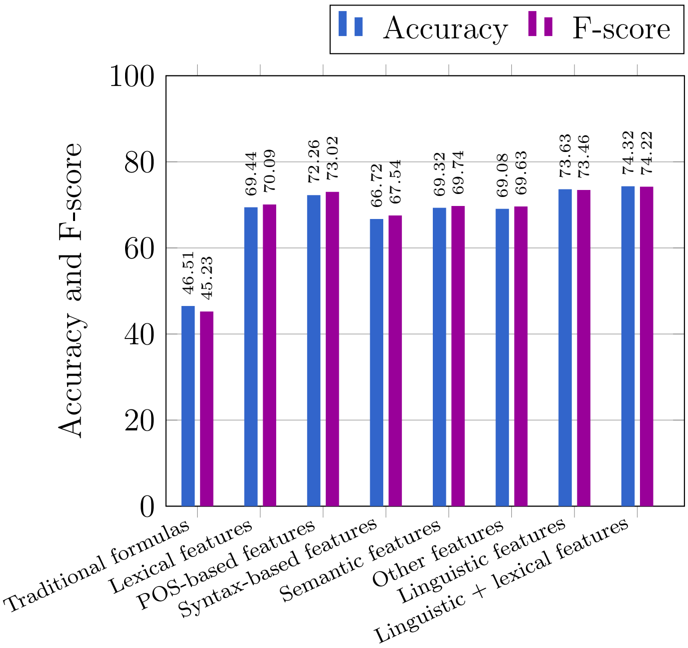


Figure 7.7: Evaluation of different feature sets in the English Wikipedia corpus.

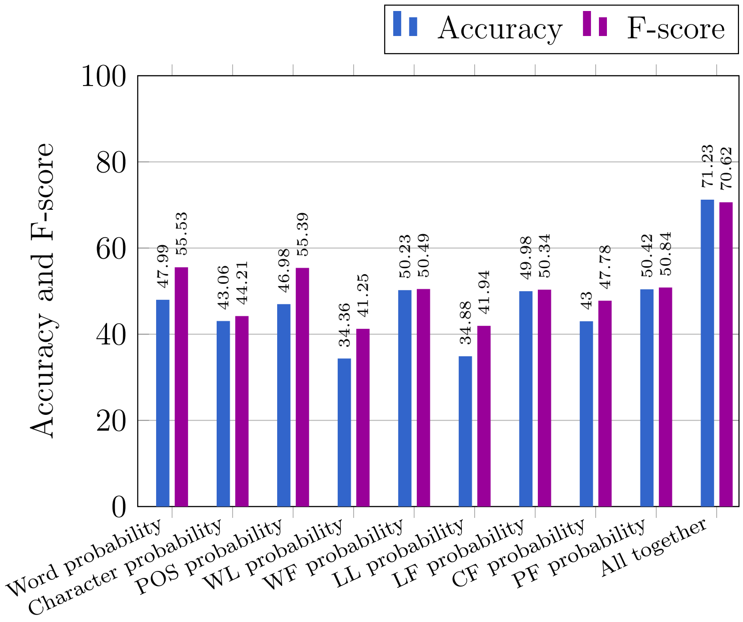


Figure 7.8: Evaluation of entropy based features in the English Wikipedia corpus.

of these features is comparable with other set of linguistic features. Among all similar features the random variable with *word probability* works better than others. It should be noted that some of the random variables such as *POS probability*, *POS frequency probability*, *Lemma length probability* and *Lemma frequency probability* are linguistically motivated. That is, probabilities are calculated from the output of some of the linguistic tools. These three features are not considered in our proposed feature list. Figure 7.8 shows the evaluation.

7.1.8 Information transmission-based features

Figure 7.9 shows the evaluation of information transmission-based features. These features show the relationship of WL and number of DW with the SL. Individually these features perform better than many other individual features. The result shows that information transmission with SL and WL are the best performing *information-theoretic* features. A classifier with 11 information-theoretic features achieves a classification accuracy that is comparable with its linguistic counterpart. The classification accuracy remains similar when the lexical features set is added to these features. These evaluations show that *information-theoretic* features are capable of capturing textual properties that influence the readability of a text. However, it is important to use these features with other corpora in order to check their usefulness for text readability classification. We

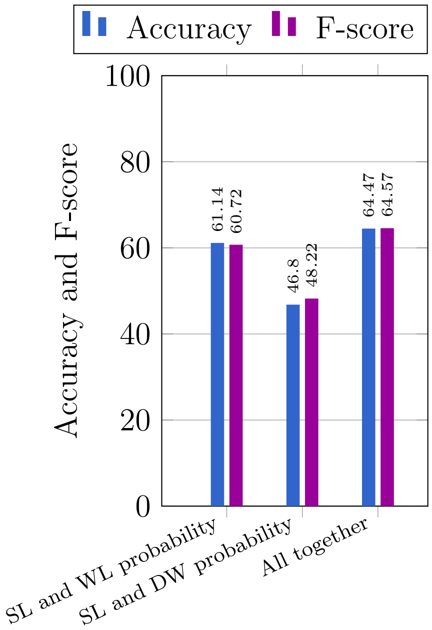


Figure 7.9: Evaluation of information-transmission based features in the English Wikipedia corpus.

get 74.21% *F-score* when these 11 *information-theoretic* features are added with *lexical* features (See Figure 7.10).

7.1.9 English textbook corpus

In this thesis, we use another English corpus that is compiled from textbooks. This corpus is slightly different than the Wikipedia corpus. The following subsections describe the performance of different feature sets.

7.1.10 Traditional features

We observe similar performance of English traditional formulas in the English Wikipedia corpus as observed by Collins-Thompson and Callan (2004), Feng, Elhadad, and Huenerfauth (2009), and Petersen and Ostendorf (2009). A classifier using five traditional formulas achieves a classification accuracy of less than 50%. However, these formulas perform better in the English textbook corpus. The reason could be that Wikipedia texts are mostly for adult readers. The average SL of documents belonging to the *very easy* readability class is 76.15 in the English Wikipedia corpus. But, in the English textbook corpus, the average SL of documents belonging to the same readability class

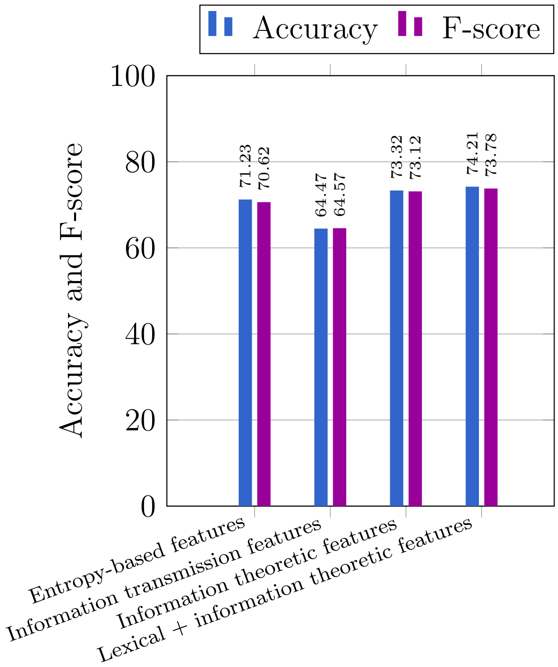


Figure 7.10: Evaluation of information theoretic and lexical features in the English Wikipedia corpus.

is 9.14. Also, articles in the Wikipedia are domain dependent. Among all of the traditional formulas, the *Dale and chall* formula performs the best. The texts are targeted for young readers where reading difficulty is influenced by the number of difficult words in texts. But, the *Gunning fog* formula uses similar difficult words but achieved the lowest *F-score* compared to the other formulas. Figure 7.11 shows the evaluation result of these formulas.

7.1.11 Lexical features

Figure 7.12 shows the evaluation of lexical features. The results show that the lexical feature set is a strong predictor of the readability. Among all of the lexical features, the average SL is the best performing feature. Table 5.4 shows that difficult texts have more sentences and the average SL is longer. We observed similar behavior of average WL where it increases over difficulty levels. However, the average WL is one of the worst performing features in the English textbook corpus. The *Dale-Chall* formula performs among traditional formulas that uses DW as a feature. However, the average DW here performs average as a lexical feature. All together 15 lexical features achieve around 67% accuracy in classification.

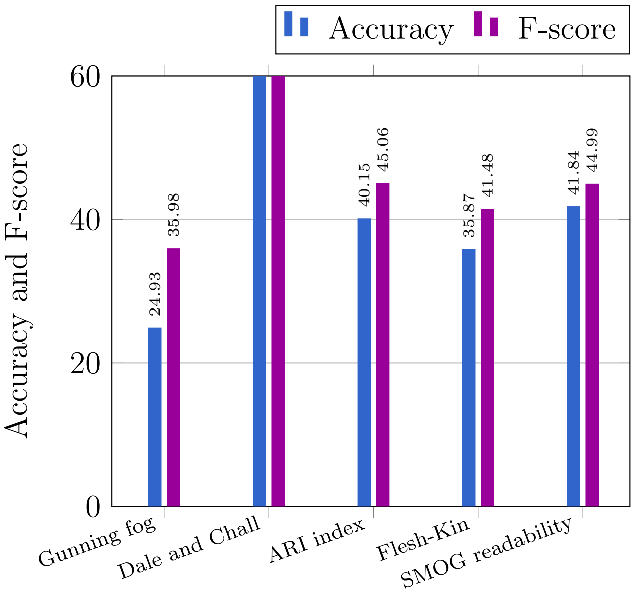


Figure 7.11: Evaluation of traditional readability formulas for the English textbook corpus.

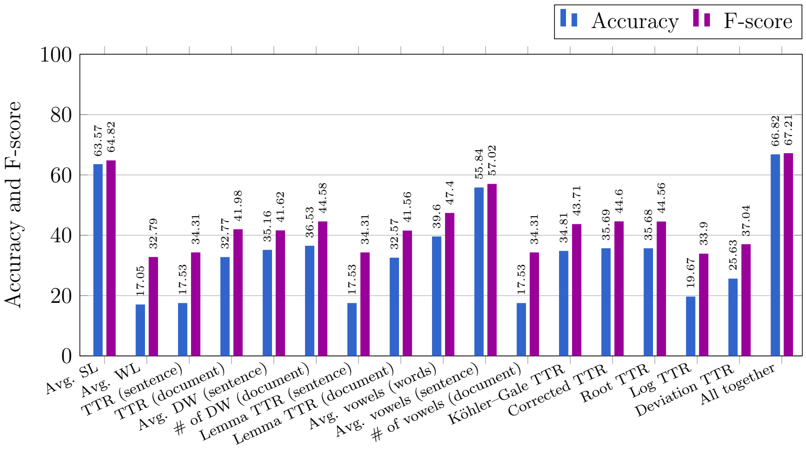


Figure 7.12: Evaluation of lexical features in the English textbook corpus.

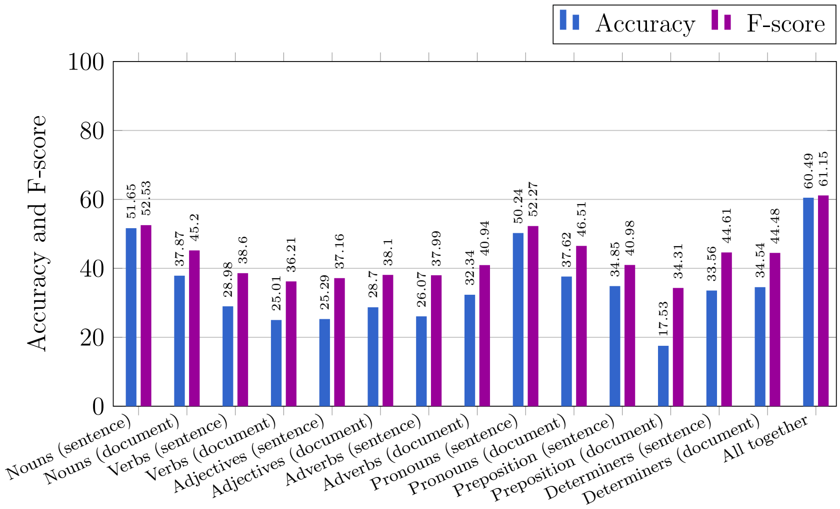


Figure 7.13: Evaluation of POS-based features in the English textbook corpus.

7.1.12 POS-based features

POS-based features were the best performing feature set in the English Wikipedia corpus. In this corpus, this feature set performed similar to traditional features. We observe a similar finding with English Wikipedia corpus, where *nouns* and *pronouns* perform better than other POS classes. By considering the *F-score*, the class, *nouns* is the best performing feature. Heilman, Collins-Thompson, and Eskenazi (2007) and Falkenjack, Mühlenbock, and Jönsson (2013) observed similar findings. However, Figure 7.13 shows that *pronouns prepositions*, and *definite articles* also have good predictive power. A classifier using all 14 POS-based features achieves classification accuracy of more than 61%. We use the same tool for tagging that we used for the English Wikipedia corpus.

7.1.13 Syntax-based features

We already have seen that the readability of a text is influenced by its syntactical complexity. Barzilay and Lapata (2008) and Pitler and Nenkova (2008) observed similar findings. Figure 7.14 shows the evaluation of different syntax-based features in the English textbook corpus. Sentence level *noun phrases* and *prepositional phrases* are the two best performing features. But, both of the features perform better at the sentence level rather than the document level. Other features perform almost the same for this corpus. All together, a classifier using this feature set achieves a classification accuracy of around 64%.

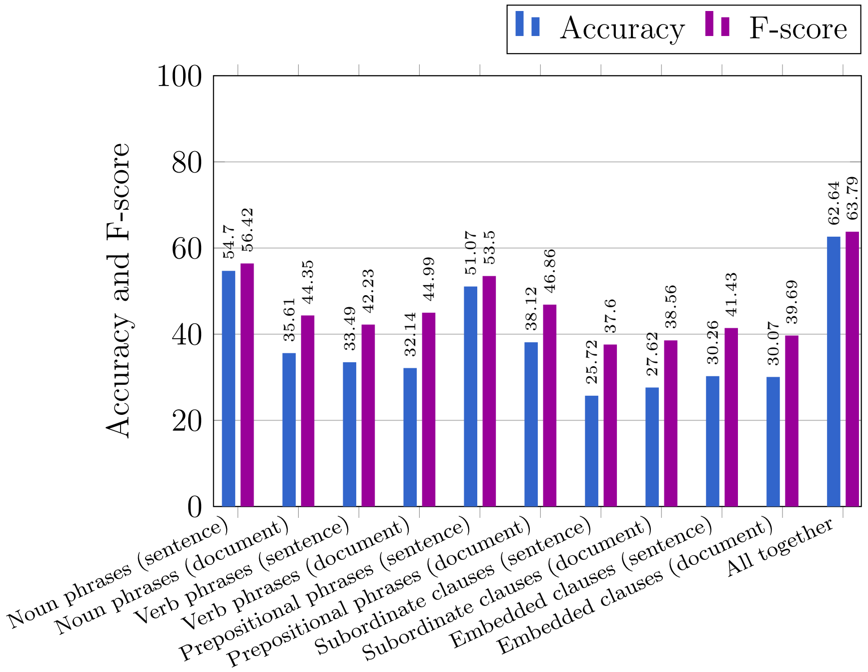


Figure 7.14: Evaluation of syntax-based features in the English textbook corpus.

7.1.14 Semantic-based features

Figure 7.15 shows evaluation of semantic features. This is the worst performed feature set among all linguistic based features. *Name entity* related features perform the worst. We observed the NER tagged texts and found that there are not many entities in these documents compared to the Wikipedia corpus. Among all of these, the *semantic frame* performs better.

7.1.15 Other linguistic features

Figure 7.16 shows the evaluation of other features. Again, the *number of familiar words per document* is one of the best performing individual features. That is, most of the documents in the textbook corpus have more frequent words. A text with more frequent words tends to be easier to read. Vocabulary overlap between a textbook corpus and the *Simple Wikipedia* unigram model is not a good indicator of readability. *Simple words* perform better in the Wikipedia corpus. This is due to domain overlap between simple Wikipedia and English Wipedia. A classifier with other features can achieve around 66% classification accuracy. All together, a classifier with 40 linguistic features achieves 67.19% classification accuracy. Figure 7.17 evaluation of different linguistic feature sets.

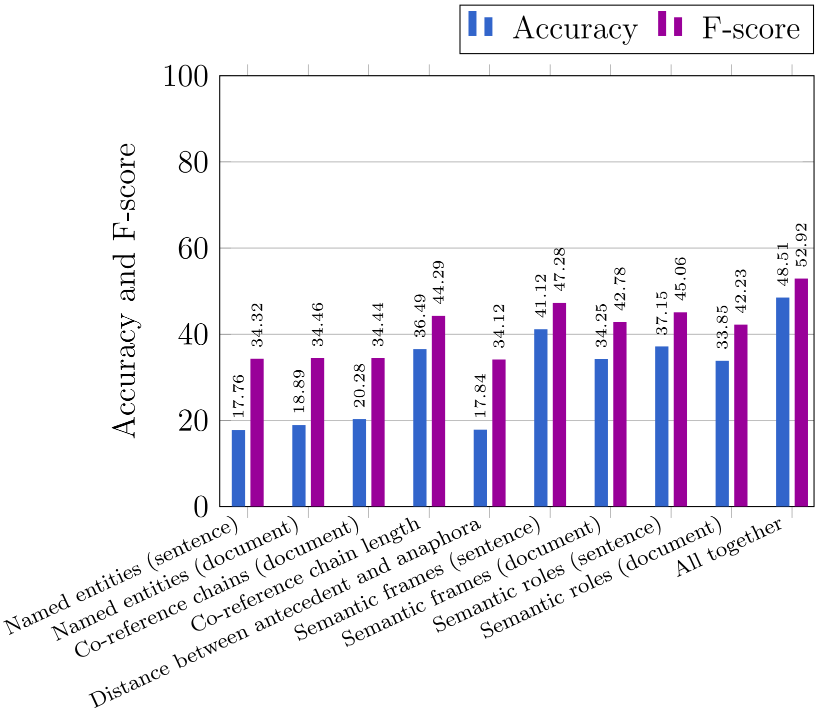


Figure 7.15: Evaluation of semantic-based features in the English textbook corpus.

7.1.16 Entropy-based features

In this thesis, we propose this feature set for text readability classification. These features are generally used to measure information density. The readability of a text can be influenced by the information density. We have already seen that this feature set performs well for the English Wikipedia corpus. However, this feature set does not perform individually as well as it performed in the previous corpus. However, collectively these features perform similar as the other feature set. A classifier with these features achieves classification accuracy of more than 63%. *word probability* is the best performing individual feature.

7.1.17 Information transmission-based features

Figure 7.19 shows the evaluation of information transmission-based features in the English textbook corpus. Both of the features individually perform better than entropy based features. It shows that a text becomes difficult when it is longer and contains

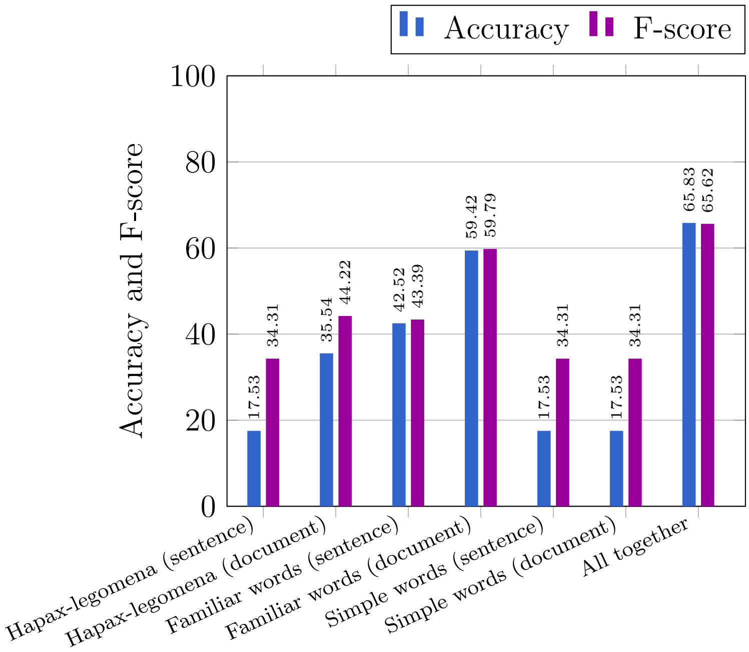


Figure 7.16: Evaluation of other linguistic features in the English textbook corpus.

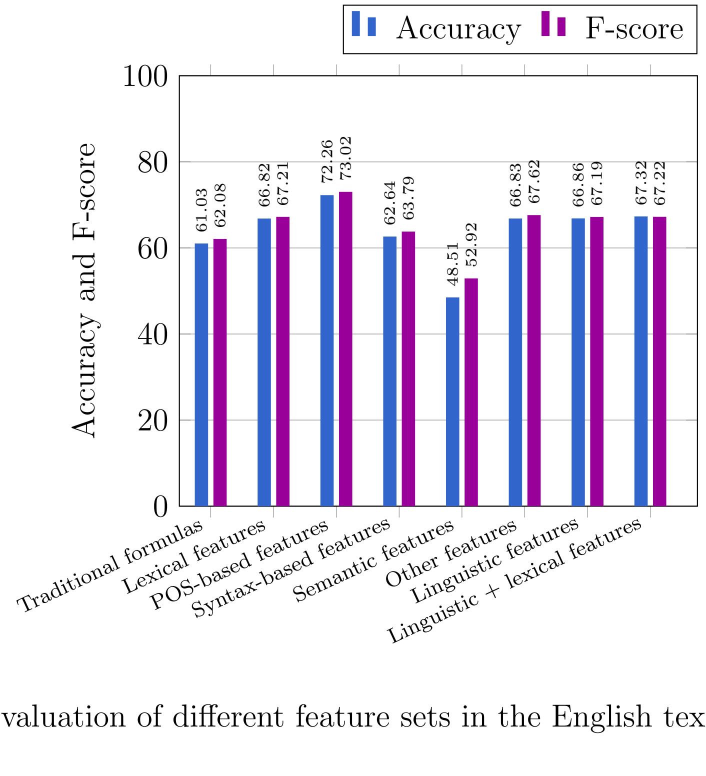


Figure 7.17: Evaluation of different feature sets in the English textbook corpus.

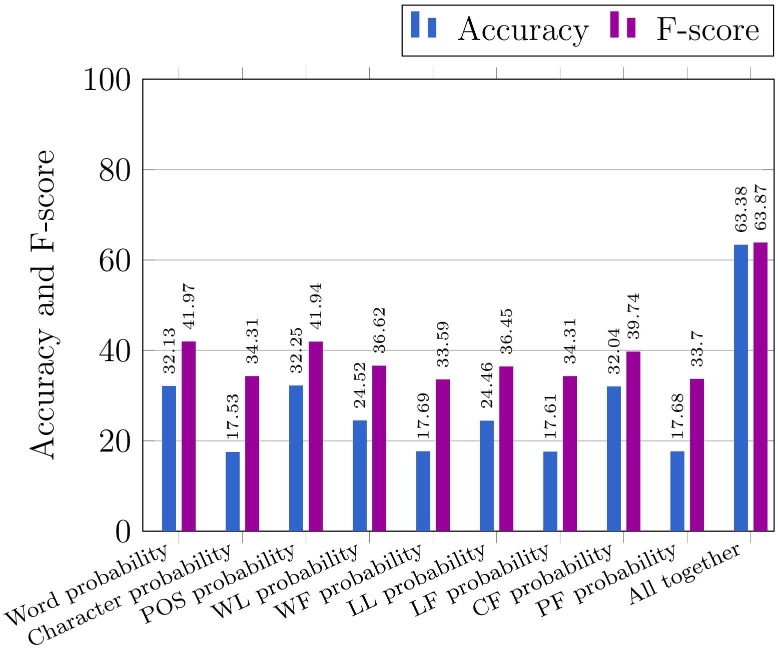


Figure 7.18: Evaluation of entropy based features in the English textbook corpus.

more long and difficult words. The classification *F-score* moves to 63.73% when these two features are added to entropy based features. All together a classifier with our proposed lexical and information-theoretic features achieves a classification accuracy of more than 75%. That is more than 12% improvement over linguistic features. Figure 7.20 shows the evaluation.

7.2 Bangla textbook corpus

The preceding subsections have shown that a readability classifier with *information-theoretic* features is able to achieve a classification accuracy that is similar to a readability classifier with 40 linguistic features. These *information-theoretic* features are language and linguistic tools independent. In this subsection we will describe experimental results of our proposed features on the *Bangla readability corpus*. However, we will build a baseline system with the same traditional readability formulas as we used for the English Wikipedia corpus. The domain of this corpus is similar to the English textbooks corpus. Bangla is considered to be a low-resource language; there are not many linguistic tools and resources available for it. That is why we propose *lexical* and *information-theoretic* features to build a readability classifier for Bangla texts.

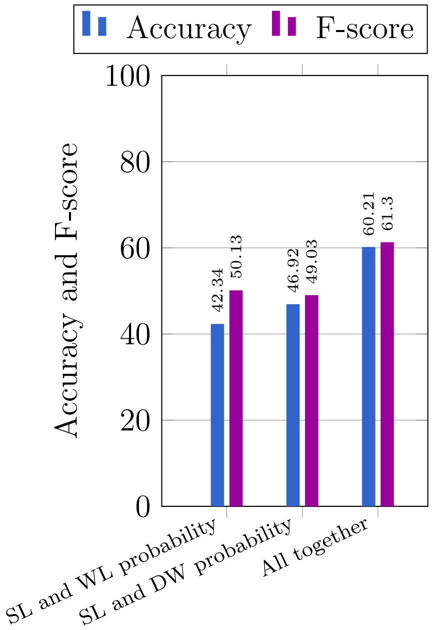


Figure 7.19: Evaluation of information-transmission based features in the English textbook corpus.

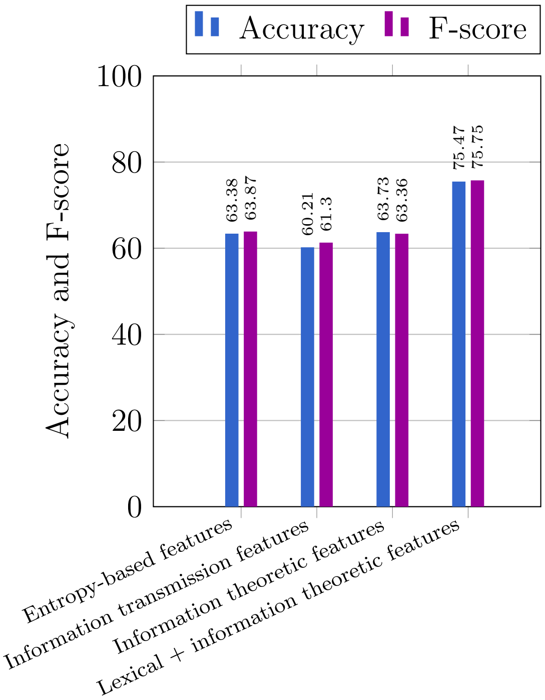


Figure 7.20: Evaluation of information theoretic features in the English textbook corpus.

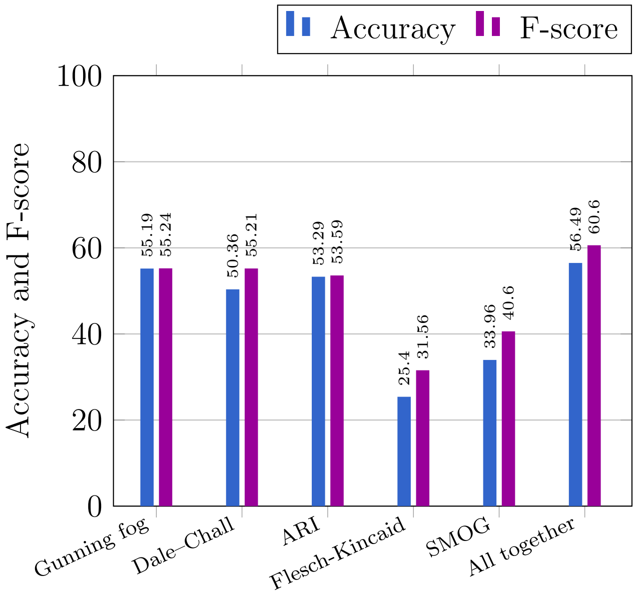


Figure 7.21: Evaluation of traditional formulas in the Bangla textbook corpus.

7.2.1 Traditional readability formulas

Traditional readability formulas for English texts have been used for readability classification of texts in different languages. Although these other languages have properties that are very different from English, researchers rely on the same formulas, anyway. Lack of resources for these languages could be a reason.

In order to make our proposed features comparable, we implemented a baseline system similar to the English corpora. The *Gunning fog readability index* and the *Dale-Chall readability measure* both analyze complex or difficult words, while differing in the definition of these words. In this thesis, we define a difficult word as any word with at least 10 letters. Figure 7.21 shows the evaluation of the baseline. The evaluation shows that the baseline features do not perform well for Bangla texts. Sinha, Sharma, Dasgupta, and Basu (2012) found a similar evaluation of these traditional formulas. However, Das and Roychoudhury (2004) and Das and Roychoudhury (2006) showed that the *Flesch-Kincaid* formula from, from Kincaid, Fishburne, Rodegers, and Chissom (1975) and the *Gunning fog* formula, from Gunning (1952), work well for Bangla texts. They used a small set of data that contains only seven documents. Among the formulas, the *Gunning fog* readability index performs best.

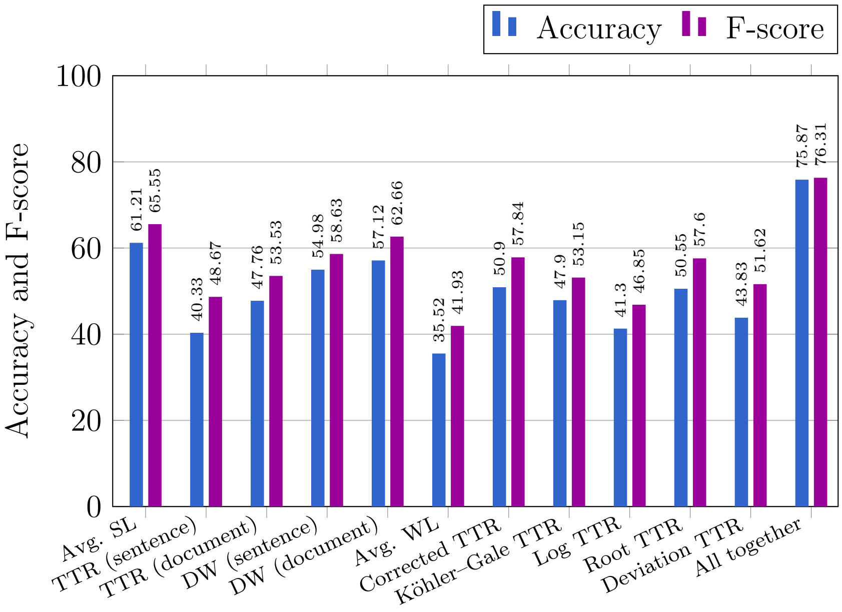


Figure 7.22: Evaluation of lexical features in the Bangla textbook corpus.

7.2.2 Lexical features

The lexical features used in our study perform better than the classical readability measures (which also include but are not limited to lexical features). Table 5.3 shows that the average SL and the difficulty levels correlate: sentence length increases for higher readability classes. This characteristic is reflected in our experiment. Although Table 5.3 shows the same for the average WL, our experimental results show that this is not a good indicator of readability. Figure 7.22 shows the evaluation of the system using only lexical features. Now, DW are a good indicator of readability. Although the individual accuracy of some of the lexical features is similar to the classical measures, the combination of all lexical features performs far better than the baseline.

7.2.3 Entropy-based features

As noted earlier, entropy measures the amount of information in a document. Genzel and Charniak (2002) and Genzel and Charniak (2003) provide evidence for the hypothesis that for certain random variables, the entropy rate is constant in a text. Textbooks are also a medium of communication between book authors and students. Conversely, information flow of a very readable document will differ from that of a less readable document. Information flow in a book of Grade 2 will be different than information flow in a book of grade nine. Thus, we expect an impact of entropy on measuring readability.

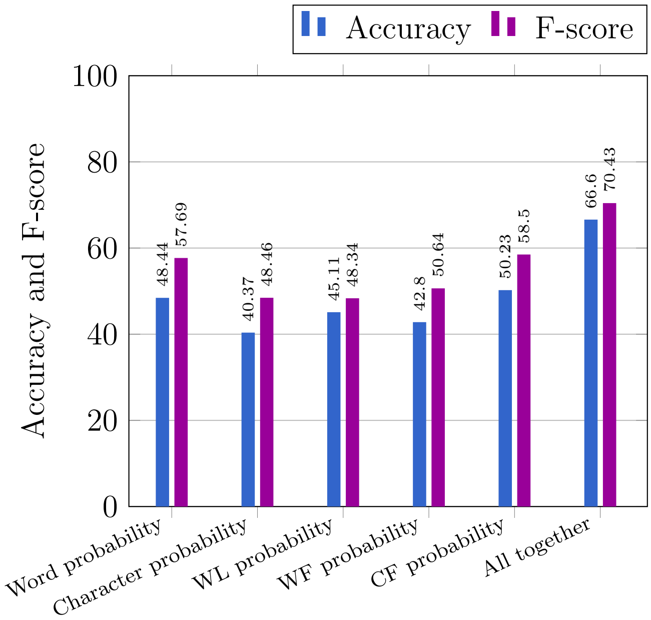


Figure 7.23: Evaluation of entropy based features in the Bangla textbook corpus.

Analyzed in isolation, entropy-related features perform on an equal footing with lexical features or classical models of readability. However, as a collective, entropy-related features outperform their classical counterpart. Among all entropy-related features, the *character frequency probability* performs best. This could be due to *dependent vowels* in Bangla texts. The frequency of these characters might influence readability. Figure 7.23 shows the results of the respective experiment.

7.2.4 Information transmission based features

Figure 7.24 gives results regarding the evaluation of two information transmission-related features. The transmission of *sentence length and word difficulty* is the best performing feature among all features considered here. Also as a group, transmission-related features perform better than the baseline and the entropy-based classifier. However, it is outperformed by the classifier based on lexical features. The average SL and the number of DW both perform well as individual features. Their combined joint and conditional probabilities give a good indicator of readability. It should be noted that these two features perform better than five entropy-based features. We get an F-score of 86.46% (See Figure 7.25) when combining 11 lexical and 7 information-theoretic features. Our experiment indicates that these features are very useful for measuring text readability.

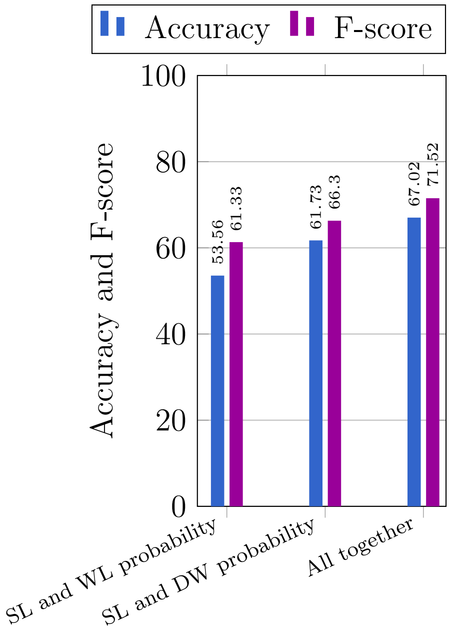


Figure 7.24: Evaluation of information transmission-based features in the Bangla textbook corpus.

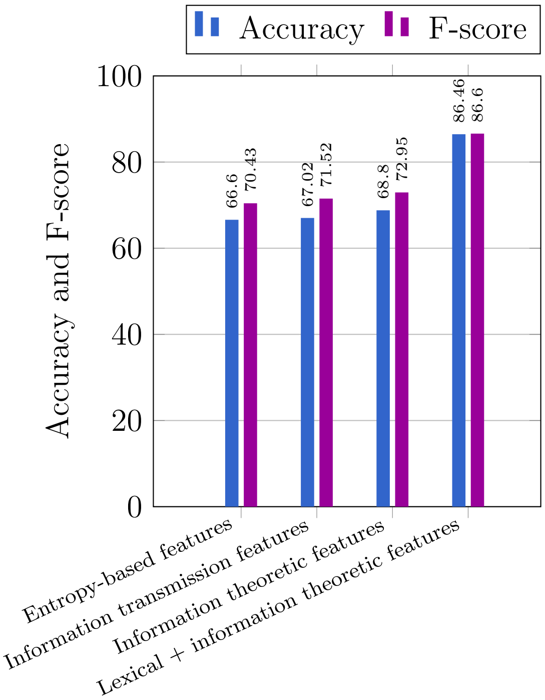


Figure 7.25: Evaluation of information theoretic features in the Bangla textbook corpus.

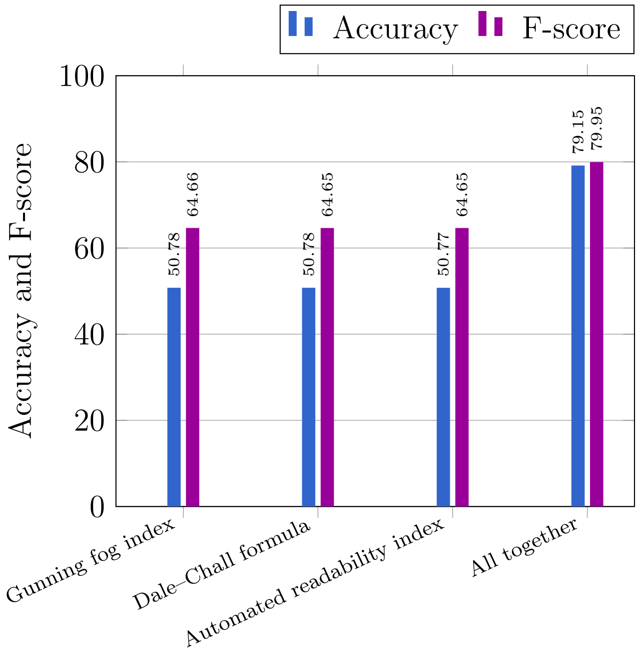


Figure 7.26: Evaluation of baseline system in the German GEO/GEOLino corpus.

7.3 German GEO/GEOlino corpus

The above subsection described performance of different traditional, lexical and information-theoretic features in the *English readability* corpus and the *Bangla readability* corpus. We describe the performance of these features in texts from another *Indo-European* language: German. The corpus is different than the preceding corpora. The corpus contains texts in two difficulty classes: *easy* and *difficult*. The subsections below show the detailed evaluations of different features in the German readability corpus.

7.3.1 Traditional formulas

Traditional readability formulas work well for readability classification of German texts. Existing readability tools that use these formulas can therefore measure reading difficulty of German texts (See 1.1). We use three traditional formulas to build the baseline system. The reason is that we did not find a freely available syllable identification tool for German. And with the system we achieve a reasonable classification accuracy. The individual performance of each feature is similar. Figure 7.26 shows that these traditional formulas are useful for readability classification of German texts.

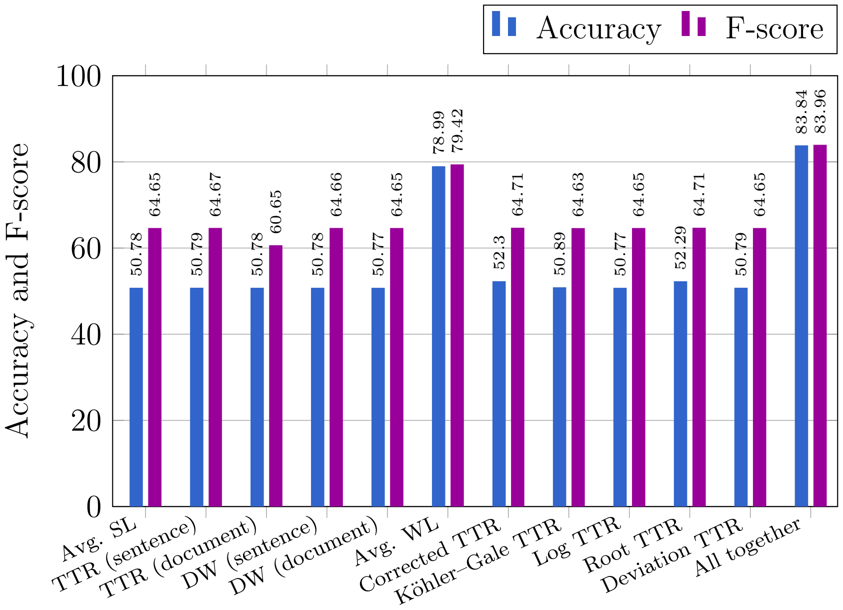


Figure 7.27: Evaluation of lexical features in the German GEO/GEOLino corpus.

7.3.2 Lexical features

Figure 7.27 shows the evaluation of lexical features in the German GEO/GEOlino corpus. Performance of most of the individual features is similar. However, average WL is the best performing feature among lexical features. This feature out-performs the readability classifier using three traditional formulas. The average WL is also selected as one of the best performing features by Hancke, Vajjala, and Meurers (2012). The reason is that German is an inflectional language and words can be very long due to compound word formation. The concatenation of words hampers reading (Inhoff, Radach, and Heller 2000). These compounds become harder to read as word boundaries are removed. TTRs performed well in Bangla corpus. However, we do not see a similar situation here. Altogether lexical features achieve a classification accuracy of 83.96%.

7.3.3 Entropy based features

The preceding subsection shows that average WL is the best performing lexical feature. That is why we can assume that a random variable related with WL could be useful among entropy-based features. The probability of WL is the best performing entropy-based feature. This evaluation shows that the readability of German text is influenced by the number of longer words. Figure 7.28 shows the evaluation of entropy-based features in the German readability corpus. The performance of a classifier using 5 entropy-based

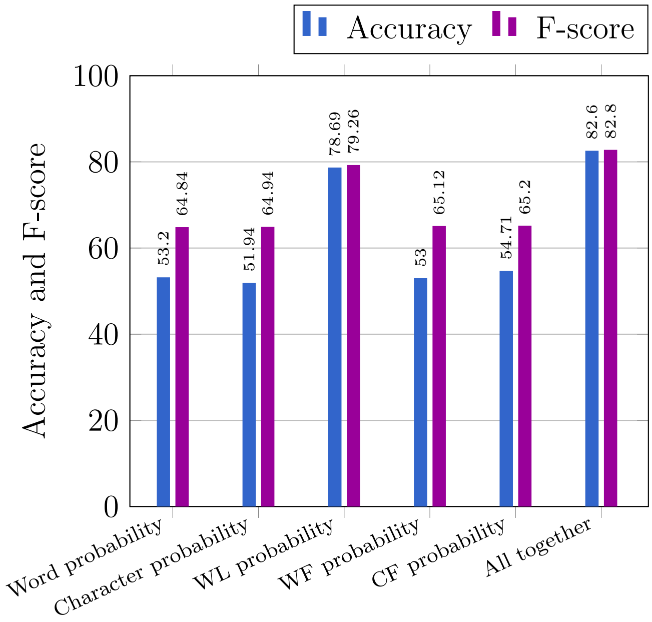


Figure 7.28: Evaluation of entropy based features in the German GEO/GEOLino corpus.

features is almost similar to the classifier using lexical features. A classifier with entropy based features achieves a *F-score* of more than 82% .

7.3.4 Information transmission based features

We already have seen that WL plays an important role in German texts. However, the information transmission of SL and WL does not perform as well as expected. Figure 7.29 shows the evaluation of information transmission-based features separately and combined. According to the table, these two features are not as useful as other measures. However, if these features are added with other feature sets then the performance of the classifier improves. Altogether *lexical*, *entropy* and *information-transmission* based features achieve a classification accuracy that is similar to other text readability research on German text. A classifier with our proposed features achieves a classification accuracy more than 86% (See Figure 7.30). It should be noted that the features we used here are language and linguistic tools independent.

7.4 Best performing features

We have experimented with a variety of features. In the end, we selected a feature list that contains eighteen lexical and information theoretic features that are able to achieve

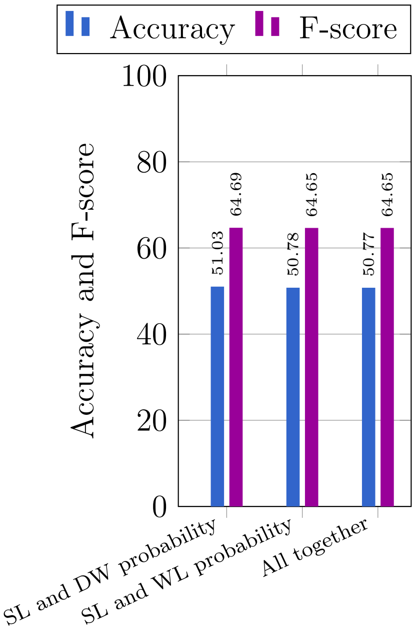


Figure 7.29: Evaluation of information-transmission based features in the German GEO/GEOLino corpus.

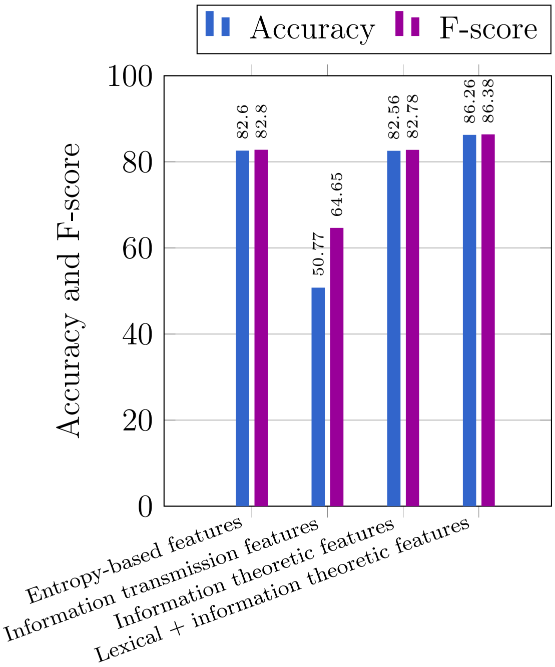


Figure 7.30: Evaluation of information theoretic features in the German GEO/GEOLino corpus.

a reasonable classification accuracy for all three readability corpora. It is important to find the best performing subset among these features; which means identifying and removing irrelevant and redundant information as much as possible (Hall and Holmes 2003). Feature selection consists of the generation of the subsets, evaluation of each of the subsets, and the stopping criteria for search and validation procedures (Dash and Liu 1997; Pitt and Nayak 2007).

We used different machine learning algorithms in the WEKA toolkit (Hall, Frank, Holmes, Pfahringer, Reutemann, and Witten 2009) for classification. The toolkit provides feature selection algorithms and search methods that can be used to select a subset of the best performing features. The feature selection algorithms are: *CfsSubsetEval* (Hall 1998), *ConsistencySubsetEval* (Liu and Setiono 1996), *ChiSquaredAttributeEval* (Bouckaert, Frank, Hall, Kirkby, Reutemann, Seewald, and Scuse 2013), *InfoGainAttributeEval* (Hall and Holmes 2003), and *OneRAttributeEval* (Bouckaert, Frank, Hall, Kirkby, Reutemann, Seewald, and Scuse 2013). We also explored eight search methods such as: *BestFirst* (Bouckaert, Frank, Hall, Kirkby, Reutemann, Seewald, and Scuse 2013), *ExhaustiveSearch* (Slaney, Fujita, and Stickel 1995), *GeneticSearch* (Goldberg et al. 1989), *GreedyStepwise* (Bouckaert, Frank, Hall, Kirkby, Reutemann, Seewald, and Scuse 2013), *LinearForwardSelection* (Gutlein, Frank, Hall, and Karwath 2009), *ScatterSearchV1* (Garciá López, Garciá Torres, Melián Batista, Moreno Pérez, and Moreno-Vega 2006), *SubsetSizeForwardSelection* (Gutlein, Frank, Hall, and Karwath 2009), and *Ranker* (Bouckaert, Frank, Hall, Kirkby, Reutemann, Seewald, and Scuse 2013).

Table 7.1 shows the best performing feature subset for the four corpora we use in this thesis for text readability analysis. One important observation is that all of these features are selected at least once. That means, all of these selected features are useful for measuring reading difficulty. Only two of these features are selected for all corpora and both of them are information-theoretic features. In total, seven features were selected for three corpora and four of them are information-theoretic. However, these information-theoretic features are explored for the first time for text readability classification. The *CfsSubsetEval* is the best performing feature selection algorithm with *genetic search* as a search algorithm for the English Wikipedia corpus. The combination is able to achieve an *accuracy* value of 82.60% and an *F-score* value of 82.93%. This setting also performed best for the English textbook corpus. The combination achieves a classification *accuracy* value of 83.33% and an *F-score* value of 83.23%. However, *ConsistencySubsetEval* performed better than other search algorithms for the Bangla textbook corpus with the *BestFirst* search algorithm. An *accuracy* value of 96.29% and an *F-score* value of 96.88% are achieved by this combination. For the German GEO/GEOLino corpus, *ConsistencySubsetEval* is the best performing feature selection algorithm with the *ScatterSearchV1* search algorithm. The combination achieved an *accuracy* value of 88.13% and an *F-score* value of 88.06%. Three different search algorithms perform better than others in the three corpora we use in this thesis.

However, Aha and Bankert (1996) found that the *BestFirst* search algorithm is better than *GreedyStepWise* search algorithm for subset evaluation. The *BestFirst* search algorithm differs from the *GreedyStepwise* in the inclusion of a backtracking component in which the number of non-improving nodes is controlled (Pitt and Nayak 2007).

English Wikipedia corpus	Bangla textbook corpus	German GEO/GEOLino corpus	English textbook corpus
	Avg. SL	Avg. SL	Avg. SL
	TTR per sentence	TTR per sentence	TTR per sentence
TTR per document			
	Av. DW per sentence	Avg. DW per sentence	Av. DW per sentence
Number of DW per document	Number of DW per document	Number of DW per document	
	Avg. WL	Avg. WL	
Corrected TTR			
Köhler–Gale TTR		Köhler–Gale TTR	
	Log TTR	Log TTR	
	Root TTR	Root TTR	
	Deviation TTR	Deviation TTR	
Word probability	Word probability	Word probability	Word probability
Character probability	Character probability	Character probability	
	WL probability	WL probability	
		WF probability	
CF probability	CF probability	CF probability	
	SL and DW probability		SL and DW probability
SL and WL probability	SL and WL probability	SL and WL probability	SL and WL probability

Table 7.1: Best performing feature sub-sets

8 Discussion of Text Readability

There are many readability tools and readability formulas available for many high-resourced languages. However, a tool that is designed for one language does not always work for another (See Section 1.1), although the traditional readability formulas that are proposed for English have been used for some other languages. However, the effectiveness of these formulas is questionable even for English texts. Si and Callan (2001), Petersen and Ostendorf (2009), and Feng, Elhadad, and Huenerfauth (2009) show that traditional formulas are not reliable enough even for English texts. Due to the recent development of linguistic tools and available computation power, different types of linguistically motivated features have been proposed for readability classification mostly for English texts. These features range from counting the number of different POS in a text to exploring sophisticated features that related to semantic and discourse structure.

In this thesis, we propose eighteen information-theoretic and lexical features for readability classification of texts from English, German and Bangla. Some of the features we use here were previously studied. It is important to compare these proposed features with state-of-the-art linguistic features. That is why we compiled the English Wikipedia corpus and the English textbook corpus. Text from the *Weekly Reader*[1] is often used as a *gold standard* for readability analysis of English texts (Hancke, Vajjala, and Meurers 2012). However, using this resource requires authorization from the owner of the *Weekly Reader*. We compiled a readability corpus from the English Wikipedia, because it is free to use and distribute. Articles in Wikipedia are rated by Wikipedia readers and these ratings are regulated by experts. Traditional readability formulas have been used in many commercial readability analysis tools for English texts. Figure 7.1 shows the average values of traditional formulas according different difficulty levels. The *Gunning fog index* of the popular family magazine *Reader's Digest*[2] is 8 (DuBay 2004). A Wikipedia article should get a higher *Gunning fog index* than *Reader's Digest* as most of the articles contain some specialized information (e.g., articles from a specialized do-

[1] Web address of Weekly Reader: `www.weeklyreader.com`

[2] Web address of Reader's Digest: `http://www.rd.com/`

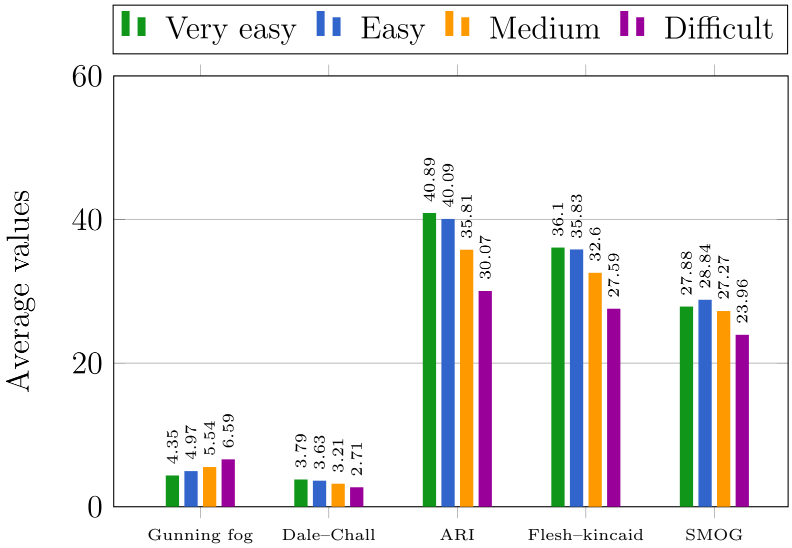


Figure 8.1: Observation of traditional formulas in the English Wikipedia corpus.

main). However, we get, for the Wikipedia, an average *Gunning fog index* of documents far below the average for the *Reader's Digest*.

Similarly, any document with a *Dale-Chall* score of 4.9 or lower will be easily understood by average *4th-grade* students (Dale and Chall 1948; Dale and Chall 1995). According to Figure 8.1, all English Wikipedia articles should be understood by average *4th-grade* students. Senter and Smith (1967) show that the ARI will vary from 41 to 57 for readers from grade level 1 to 7. However, we have got ARI scores lower than 40 for all classes, while Jatowt and Tanaka (2012) found a high score of ARI 80 for English Wikipedia articles. The grade level of *Flesch–Kincaid* is also very high for all difficulty classes. The SMOG readability formula estimates the year of education needed to understand a text. A text should contain 30 or more sentences in order to be suitable for the SMOG readability formula. Figure 8.1 shows that many years of education are needed in order to understand English Wikipedia articles. Although the average values of traditional formulas are different for each class, these traditional formulas clearly do not work well to measure readability of English Wikipedia articles. The reasons could be the same as those found by Si and Callan (2001), Petersen and Ostendorf (2009), and Feng, Elhadad, and Huenerfauth (2009). However, we have to acknowledge that a very small number of Wikipedia articles contain less than 30 sentences.

Some of the traditional formulas use lexical features such as average SL, average WL or *average number of syllables* in a word. In the English Wikipedia corpus, average SL negatively correlates, that is SL decreases when the difficulty level rises. However,

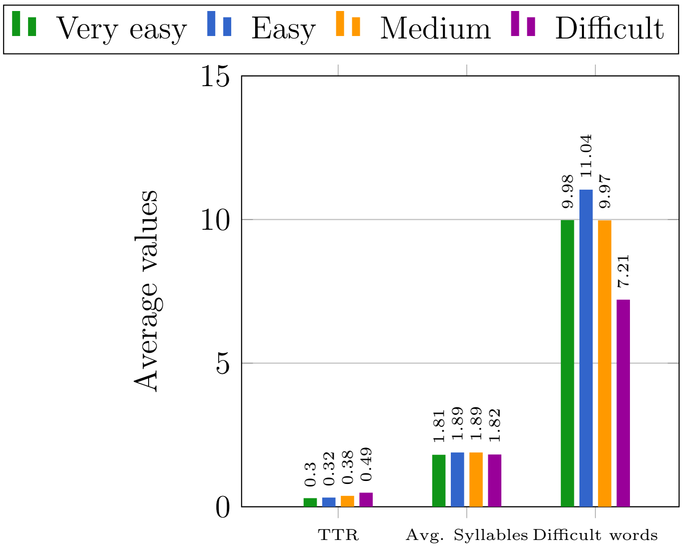


Figure 8.2: Lexical features in the English Wikipedia corpus.

an easier text has lower average WL than a difficult text (See Table 5.1). Figure 8.2 shows that the TTR plays a strong role in distinguishing documents based on reading difficulty. Our assumption was that an easier document will have a lower TTR than a difficult document. This was the motivation for using different TTR formulas. However, the observation of TTR is the opposite of our hypothesis. Here, TTR positively correlates with reading difficulty. Among these TTR formulas, *Köhler–Gale TTR*, *log TTR* and *deviation TTR* are good indicators of reading difficulty. Figure 8.3 shows average value of these TTR formulas. Although *number of DW in a document* is selected as one of the best performing features for all corpora (See Table 7.1), this behavior is not visible in Figure 8.2.

Figure 7.3, Figure 7.4 and Figure 7.5 show the evaluation of different linguistic features. All of the linguistic features perform better when being considered at the document level as opposed to the sentence level. This behavior of features shows that it is more challenging to measure the reading difficulty of a sentence than of a document. In other words, it is very hard predict the reading difficulty of a document only by considering features at the sentence level. Figure 8.4 shows the average score of the three best performing (individual evaluation) linguistic features. The difficulty level rises as the number of *pronouns*, *adverbs* and *determiners* is reduced. These numbers are influenced by the document length. The article quality in Wikipedia is therefore biased by the document length (Flekova, Ferschke, and Gurevych 2014). Another reason could be that the *well-written* feature of the Wikipedia article feedback tool is influenced by other

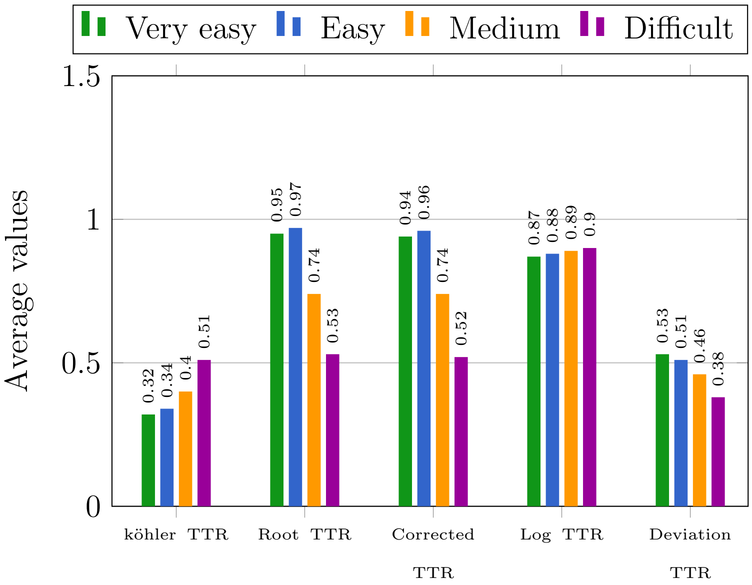


Figure 8.3: Different TTRs in the English Wikipedia corpus.

features such as: *complete*. Generally, the *document length* of a *complete* article will be longer than an *incomplete* article. Wikipedia readers are biased to give a better score regarding *well-written* features when a document is *complete*. Also, Flekova, Ferschke, and Gurevych (2014) show that the correlation between *trust-worthy* and *well-written* is 0.89. Our findings about the English Wikipedia corpus show that, sometimes crowed sourcing is problematic. For annotation, it is better when annotators have some specific expertise about the annotation subject and parameters. It would be better if each of the Wikipedia articles were annotated by considering one parameter at a time in order to avoid being influenced by other parameters.

Figure 8.5 shows our observations of entropy based features in the Wikipedia corpus. Our hypothesis is that a document with higher entropy will be harder to read and understand than a document with lower entropy. However, we observe similar behavior as TTR. The average word entropy of the *very easy* difficulty level is the highest. Again, entropy values could be influenced by document length. Features related to frequency are more reliable than other random variables. Information transmission based features are also good indicators of reading difficulty. Figure 8.6 shows observation of information transmission based features.

The readers feedback on the *well-written* criteria of the English Wikipedia articles was influenced by other criteria. That is why it was necessary to compare our proposed features with linguistic features using another corpus that is more suitable for this study.

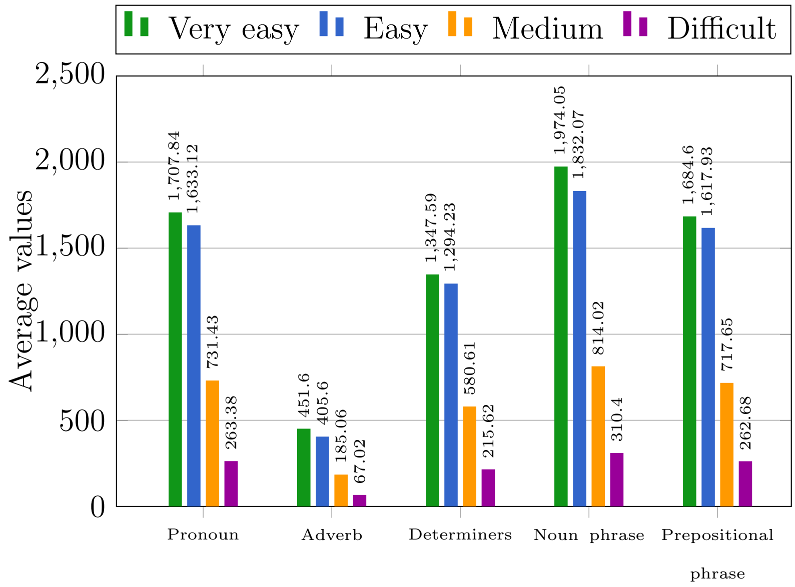


Figure 8.4: Best performing linguistic features in the English Wikipedia corpus.

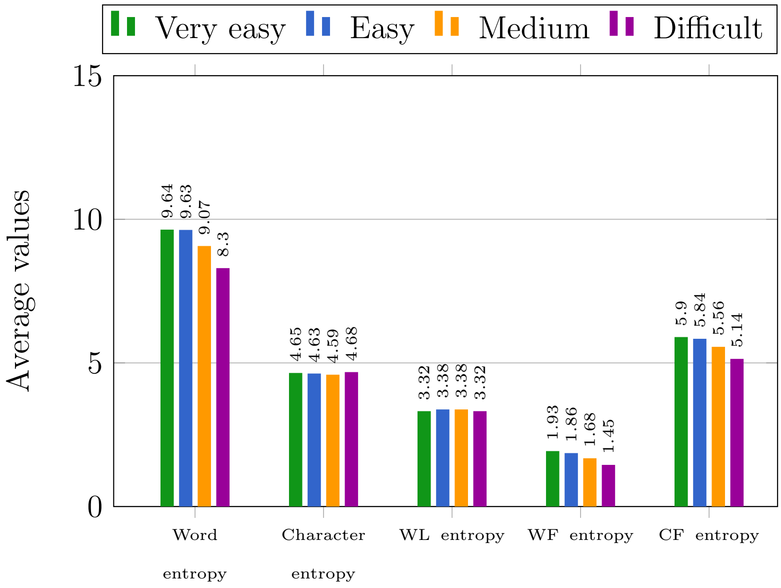


Figure 8.5: Entropy based features in the English Wikipedia corpus.

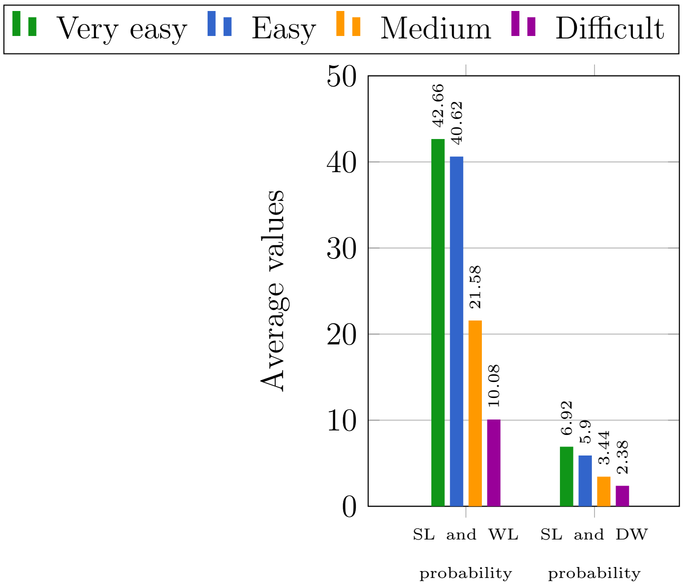


Figure 8.6: Information transmission based features in the English Wikipedia corpus.

Figure 8.7 shows our observation of traditional formulas in the English textbook corpus. A classifier only with the *Dale-Chall* formula achieves more than 60% classification accuracy. However, the *Dale-Chall* score of each of the documents is below 1. That is, all of the documents will be understood by an average 4th grade student. But this is not the case here. By observing the score of the *Gunning fog* formula, all of the documents are very easy to understand. The *ARI* score correctly evaluated the corpus. The average score of the documents belong *difficult* readability class is 8.98. The score corresponds to the typical reading level of a 14 year old child. The traditional formulas perform better than the English Wikipedia corpus.

The Lexical feature set is the best performing feature set. A classifier using this set achieves a classification accuracy more than 67%. Among all of the lexical features, Figure 8.8 and Figure 8.9 show the most discriminating features. The *difficult words* performs the best. A difficult document contains more *difficult words* than an *easy* or *medium* document.

The behavior of all TTR formulas are opposite in this corpus compared to the English Wikipedia corpus. The average value of different TTR formulas increases with the difficulty level. However, the average value of *Köhler–Gale TTR* decreases. All of the TTR formulas are good indicators of reading difficulty.

A total of 40 linguistic features were considered in this study. Among all of these, *noun*, *pronoun*, *adverb*, *determiner*, *noun phrase* and *prepositional phrase* were found

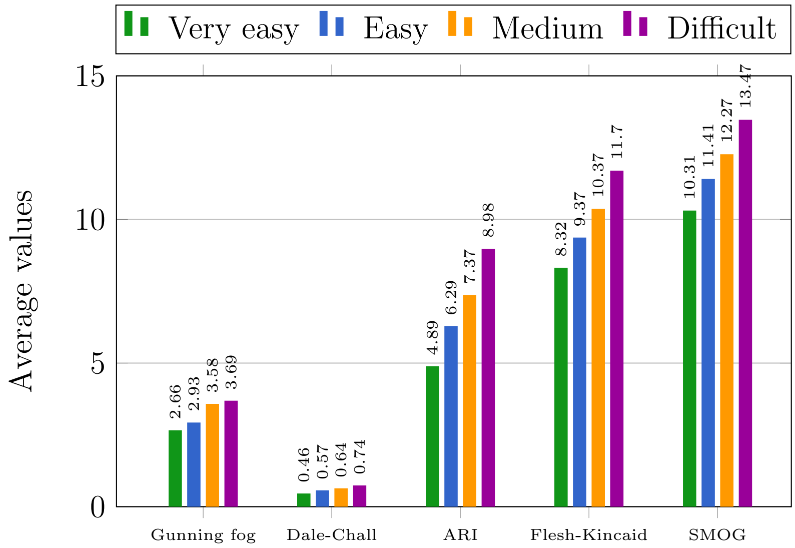


Figure 8.7: Observation of traditional formulas in the English textbook corpus.

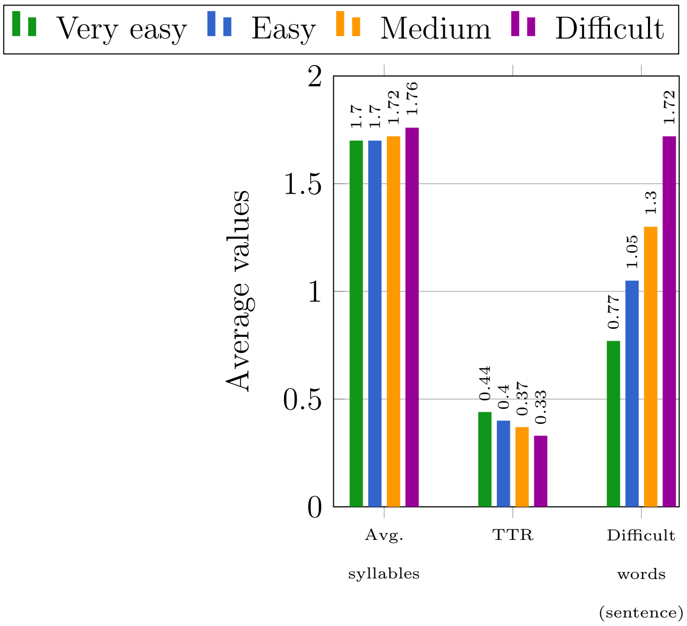


Figure 8.8: Observation of traditional formulas in the English textbook corpus.

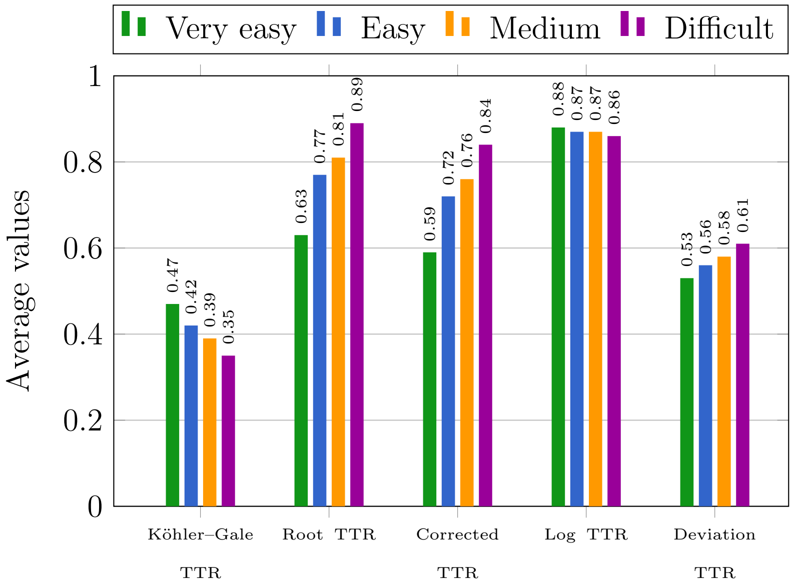


Figure 8.9: Observation of different TTR formulas in the English textbook corpus.

to be good predictors of reading difficulty. Figure 8.10 shows our observations of four best performing linguistic features in the English textbook corpus. None of the *semantic* features are good indicators of reading difficulty. The reason could be the performance of tools for semantic processing are still not satisfactory, or the domain of the corpus we use here is different than the tools were trained on.

Figure 8.11 shows our observations of *entropy* based features. The figure shows that all of the proposed features are good indicators of reading difficulty. The average value of each of the features increases with the reading difficulty. This satisfies our hypothesis. From Figure 8.11, we can infer that a difficult document will have higher entropy than an easy document.

Similar to the *entropy* based feature, two information transmission based features are also good indicators of readability. Figure 8.12 shows the observation of these features. The average value of these features increase with the difficulty. According to the average values of our proposed features, these features have some predicting power to measure the readability of text. We have to also consider their behavior in the other two corpora we use in this thesis.

The Bangla readability corpus is also a suitable resource for readability classification. Texts produced for different grade levels are controlled by the National Curriculum and Textbook Board (NCTB). The observation of different features in this corpus will be more acceptable. Our experimental results show that classical readability models as

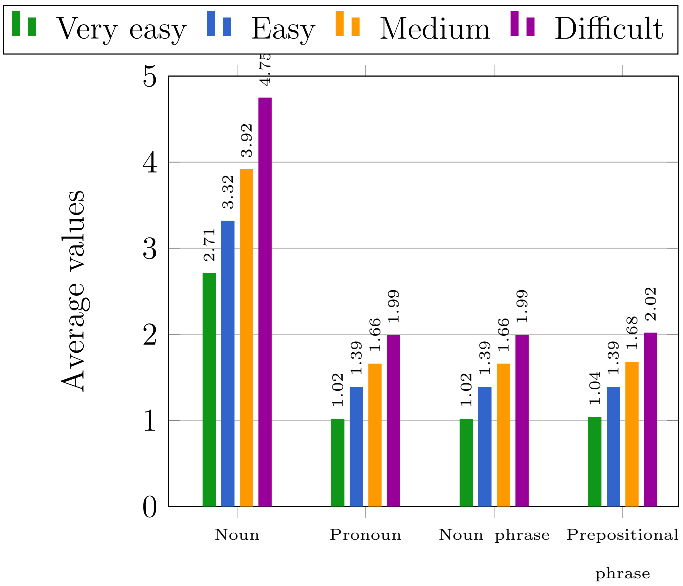


Figure 8.10: Best performing linguistic features in the English textbook corpus.

proposed for English texts are not useful for Bangla texts. Sinha, Sharma, Dasgupta, and Basu (2012) produced similar results. Das and Roychoudhury (2004) and Das and Roychoudhury (2006) found that classical readability measures are useful for Bangla readability classification. However, the size of the data they used was very small. The reason behind the poor performance of these classical measures on the Bangla corpus could be that Bangla script contains glyphs which represent clusters and ligatures.

The value of some of the lexical features correlates positively. That is, the value of these features rises when the difficulty level rises. The DW is one of the oldest lexical features used for readability classification. Figure 8.2 shows that this feature does not influence the readability of the English Wikipedia corpus. However, Figure 8.13 shows that a document from the *very easy* class contains more or less one difficult word. The behavior of TTR in the Bangla corpus is different than the Wikipedia corpus. Here TTR negatively correlates with difficulty levels. However, our hypothesis was that a difficult text will be more dense lexically than an easy text.

Figure 8.14 shows the behavior of different TTR formulas in the Bangla corpus. The average value of each TTR formula varies with the difficulty levels. All TTR formulas positively correlate with difficulty level. However, TTR formulas negatively correlated in the Wikipedia corpus. In the Bangla corpus, the average values of Köhler TTR and Log TTR decrease as the difficulty level rises.

Although entropy-based features are useful for readability classification of the English

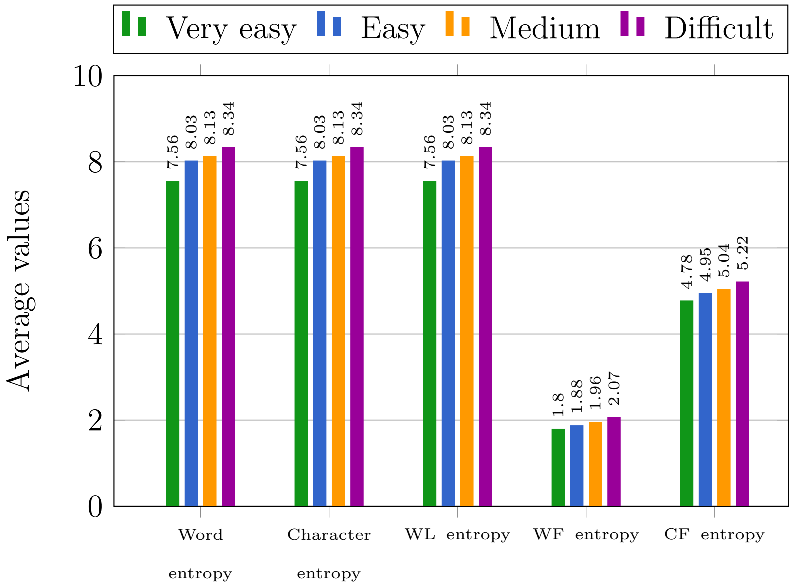


Figure 8.11: Entropy based features in the English textbook corpus.

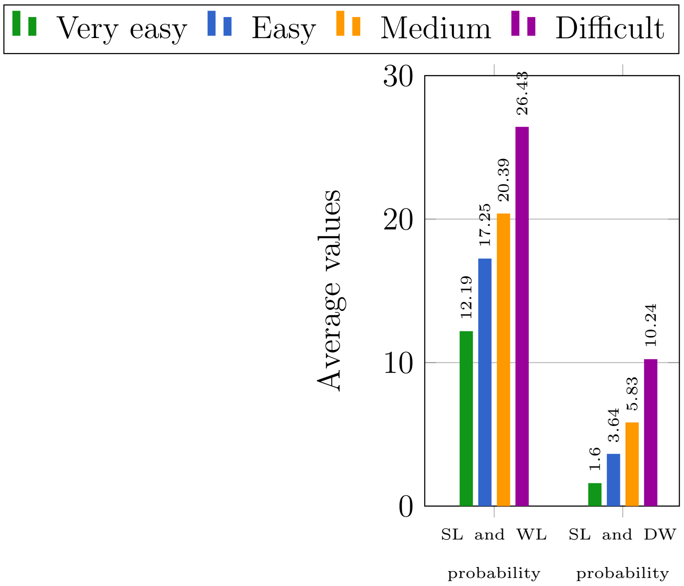


Figure 8.12: Information transmission based features in the English textbook corpus.

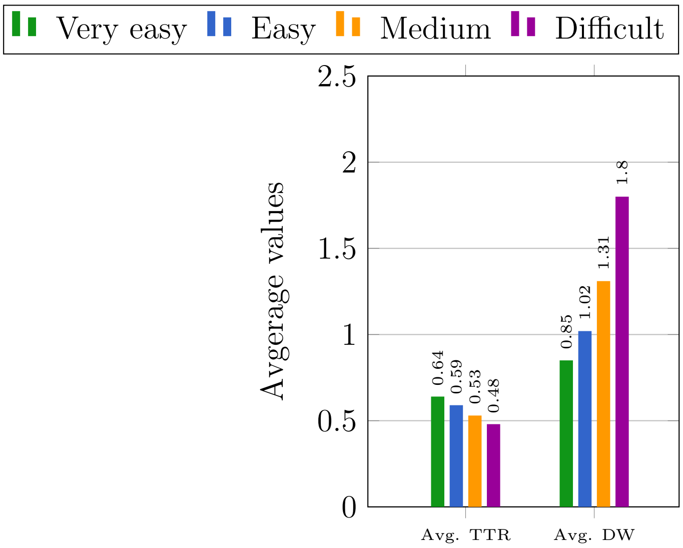


Figure 8.13: Lexical features in the Bangla textbook corpus.

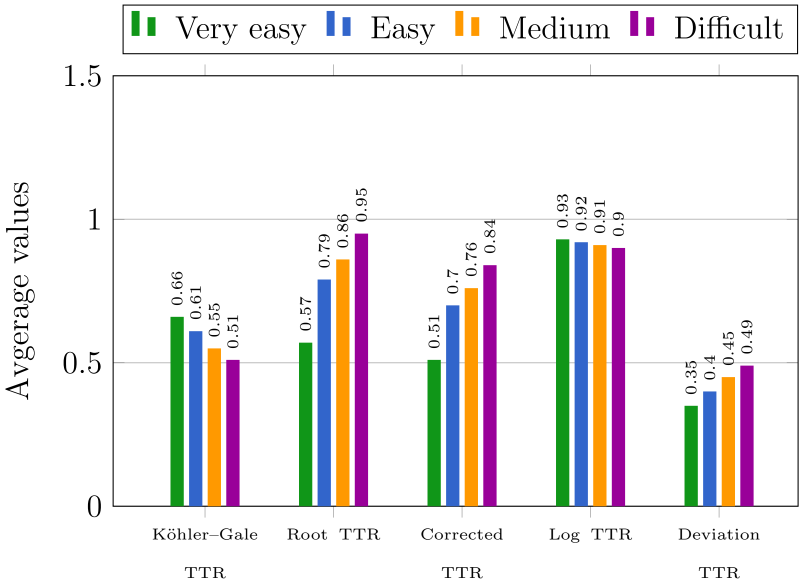


Figure 8.14: Different TTR features in the Bangla textbook corpus.

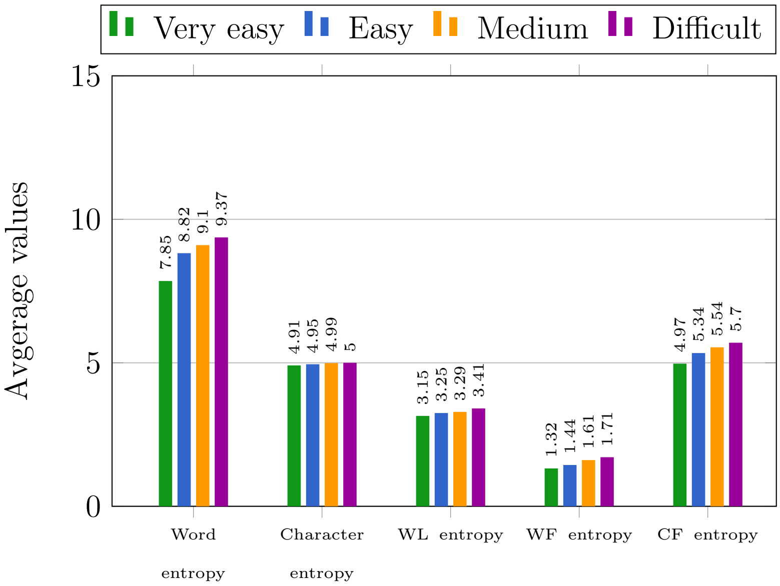


Figure 8.15: Entropy-based features in the Bangla textbook corpus.

readability corpus, their behavior was the opposite of our hypothesis. However, in the Bangla corpus these features behaved as we expected. A text with higher entropy will be more difficult than a text with a lower entropy. *Character entropy* perform similarly in the Bangla corpus. As we consider the Bangla readability corpus as an ideal resource for readability classification, we can say that easier texts will have lower entropy than difficult texts and will contain more function words. Figure 8.15 shows the average values of different entropy related features.

Information transmission-based features also performed well, as we expected. Figure 8.16 shows the behavior of two of these features. The table shows that a text becomes more difficult when the sentences are longer and contain more difficult words.

Lexical features constitute the best performing subset of features. Further, the average SL is a good indicator in Bangla texts in terms of readability. The classification performance increases up to 86% when information-theoretic features are combined with lexical ones.

Sinha, Sharma, Dasgupta, and Basu (2012) proposed two computational models for Bangla text readability. They proposed models by performing user experiments in which users identified structural characteristics of Bangla texts. We also compared their models with our readability measures. Table 8.1 summarizes the results. It shows that polysyllabic words and consonant-conjuncts should be considered as features when building readability classifiers for Bangla. These features together perform better than

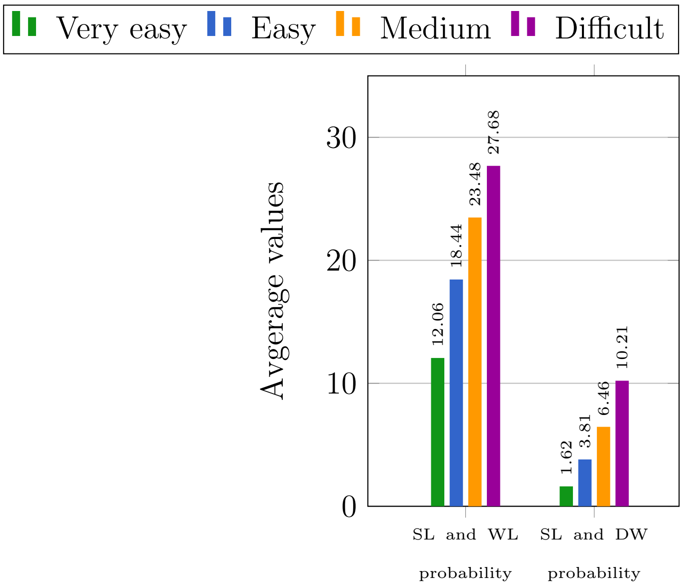


Figure 8.16: Information transmission-based features in the Bangla textbook corpus.

Features	Accuracy	F-Score
Sinha, Sharma, Dasgupta, and Basu (2012) model 3	64.99%	60.61%
Sinha, Sharma, Dasgupta, and Basu (2012) model 4	67.58%	66.18%
Together	74.57%	73.73%

Table 8.1: Evaluation of features proposed by Sinha, Sharma, Dasgupta, and Basu (2012) in the Bangla textbook corpus.

information-theoretic features. However, information-theoretic and lexical features perform far better when they are combined. It remains an open question what happens to performance if we additionally consider the feature list of Sinha, Sharma, Dasgupta, and Basu (2012).

Average SL in the German GEO/GEOLino corpus behaves similarly to the Bangla textbook corpus. Average WL is the best performing feature for German texts (See Figure 7.27). The average SL of the *difficult* readability class is more than the *easy* difficulty class. The *average DW* is lower for the *easy* class. However, both classes have the same average TTR value (See Figure 8.17).

Similar to the TTR, the average *Köhler TTR* values are also the same for both classes. However, other variations of TTR formulas perform similarly to the Bangla readability corpus. Figure 8.18 shows the behavior of different TTR features.

Our proposed entropy based features perform well, as we expected for the German corpus. Entropy of a *difficult* document will be higher than entropy of an *easy* document.

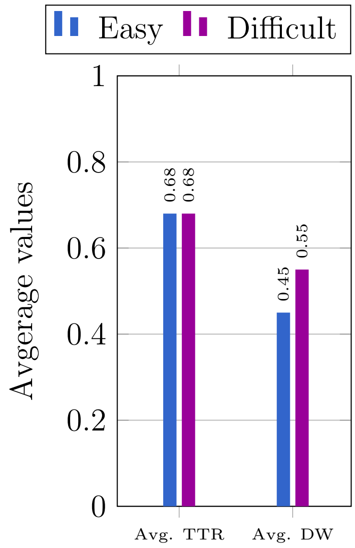


Figure 8.17: Lexical features in the German GEO/GEOLino corpus.

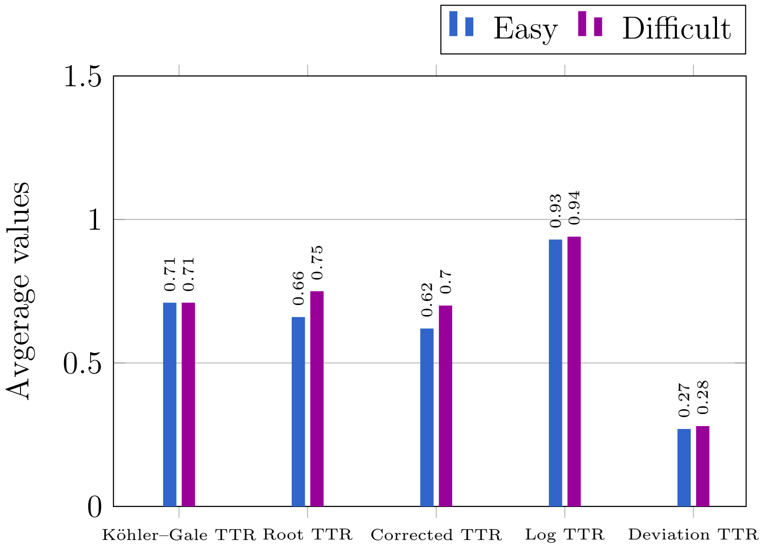


Figure 8.18: TTR features in the German GEO/GEOLino corpus.

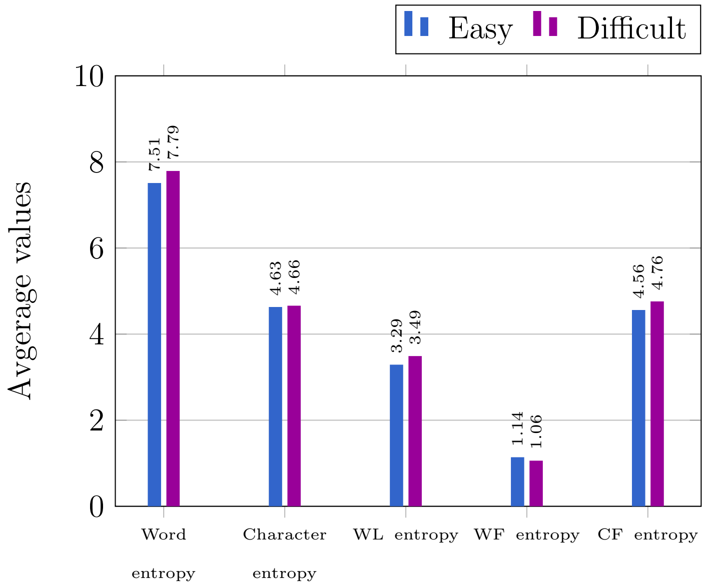


Figure 8.19: Entropy features in the German GEO/GEOLino corpus.

However, entropy of word frequency of a *difficult* class is lower. Figure 8.19 shows average values of these features in the German corpus.

Information transmission-based features also performed similarly to the Bangla corpus. However, information transmission based on sentence length and difficult words is not as distinguishable as their behavior in Bangla corpus. Figure 8.20 shows average values of these features.

The most recent readability analysis on German texts was performed by Hancke, Vajjala, and Meurers (2012). They achieved 89.7% classification using 155 lexical and linguistically motivated features. Their top ten features were able to achieve 84.3% classification accuracy. However, a classifier using our proposed 18 features achieves a accuracy of 86.38. The features we proposed in this thesis do not require any kind of linguistic knowledge. It should be noted that we use the same German readability corpus as used by Hancke, Vajjala, and Meurers (2012).

All of the observations described above show that our proposed features have good indicative power to measure reading difficulty. Figure 2.1, Figure 2.2 and Figure 2.3 show the non-parametric estimation of entropy in three different corpora. These figures satisfy our hypothesis that a difficult document will have higher entropy than an easier document. These figures also satisfy the claim made by Genzel and Charniak (2002) and Genzel and Charniak (2003). That is, the local measure of *entropy* will increase with the number of sentences. Entropy based features are cognitively motivated, espe-

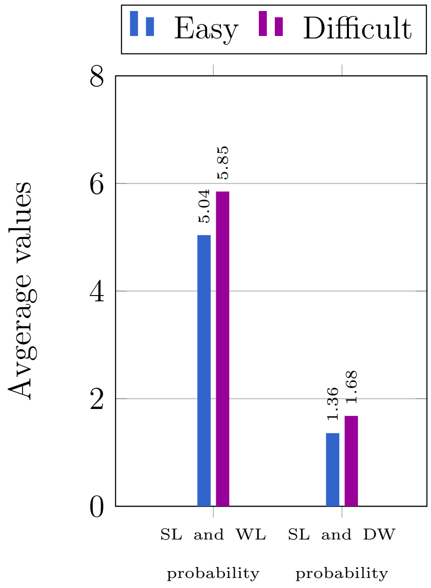


Figure 8.20: Information transmission-based features in the German GEO/GEOLino corpus.

cially the *word entropy.* This feature reflect the uncertainty in text that is related with predictability of next words. The cognitive model which is built during reading is a sequential process (Kintsch 1998). As part of the cognitive model, a network of *propositions* is built. When a text is more predictable it will be easier to build the cognitive networks for readers. This is the one reason that *word entropy* is a good indicator or reading difficulty. Similarly other information-theoretic features are also good indicators for the same reason. But, these features have a significant drawback. These features fail to measure reading difficulty at the sentence level. Document level readability measurement will be easier for these features. Therefore, it is necessary to know how many sentences a document should have at least in order to get a reasonable classification using this set of features.

We perform an experiment using the English textbook corpus to find out the threshold. The experiment started with each of the documents contain 1 sentence. Our hypothesis was that we will keep adding sentences into documents until the classification *F-Score* reaches 60%. Finally, each of the documents required 26 sentences to achieve the desired classification *F-Score*. So it is necessary that each of the documents contain at least 26 sentences for a better performance of the readability classifier using our proposed features.

9 Source and Translation Analysis

Translation is a process of rewriting an original text in a different language (Lefevere 2002). The process leaves behind latent traces in its textual output. These traces might come from different factors such as the source language, the translator's background, the translation environment and other sources. Although it is a very old linguistic process, in recent days translation scholars use corpora for exploring different properties of translation.

Corpus based translation studies is a new direction of research in the field of translation studies, starting in the early 1990's (Kruger, Munday, and Wallmach 2011). In recent years, corpus based research has become very popular among scholars in the area of translation studies; it has undergone a rapid development in linguistic investigation. As (Laviosa 1998, p:474) puts it: "the corpus based approach is evolving, through theoretical elaboration and empirical realization, into a coherent, composite and rich paradigm that addresses a variety of issues pertaining to theory, description, and the practice of translation". The translation process is influenced by environment, culture and the cognitive process of reading, understanding and rewriting. All of these put some markers into the translated texts that make them distinguishable from the respective source texts.

Currently, it is necessary to identify which documents are original (source) and which documents are translated from others. We can use NLP and machine learning techniques for this task. This is one of the applications of corpus based translation studies.

Source and translation classification is useful for different NLP applications. *Statistical machine translation* (SMT) uses training data of source texts and translated texts to build *translation models*, and uses *language models* to produce the final output in the target language. Generally, machine translation experts are not concerned about the data that they use to build the models. Lembersky, Ordan, and Wintner (2011) have shown that a *language model* from the translated text improves the performance of a Machine Translation (MT) system. A source and translation classifier can also be used to identify translated text as a part of a plagiarism detection tool. There are many plagiarism detection tools currently available. These tools perform reasonably well when the plagiarized texts come from the same language. However, these tools struggle to identify plagiarized text when the candidate text is a translation from another language (Franco-Salvador, Gupta, and Rosso 2013). Plagiarism detection tools divide a document into chunks in order to find candidate chunks for plagiarism. Any chunk that behaves like a translated text could be a potential candidate for multilingual plagiarism. In this case, a *source and translation* classifier would be useful to identify these potential candidates.

Many translation scholars have described the properties of the translation process itself as well as of the relation between source and target texts of translations. They proposed some translation properties (Baker 1996; Olohan 2001; Laviosa 2002; Hansen 2003; Pym 2005) by exploring monolingual or bilingual corpora. These translation properties are often called *translation universals.* These properties are subsumed under five keywords: *explicitation*, *simplification*, *normalization*, *levelling out* and *interference.* However, these properties of translations are issues of debate in the field of translation studies. Tymoczko (1998), Bernardini and Zanettin (2004), and Mauranen and Kujamäki (2004) claim the universality aspect of proposed translation properties. But, Toury (2004) stated that the value of these assumptions stands in their explanatory power. Therefore, it is necessary to investigate the validity of the translation properties. The following sections describe them in detail.

9.1 Translation properties

9.1.1 Explicitation

Generally a translated text tends to be more explicit than the corresponding non-translated text. A translator always tries to render any implicit content in the source language text more explicitly during the translation process. Translators are biased to make translations more *explicit* in order to resolve ambiguities that might be inherited in the translated text. Vinay and Darbelnet (1958) used the term *explicitation* as " a process of introducing information into the target language which is present only implicitly in the source language, but which can be derived from the context or situation"(Vinay and Darbelnet 1995; Pym 2005). However, Blum-Kulka (1986) first claimed *explicitation* as a translation universal when she studied translated *French* texts from *English* originals by professional and non-professional translators. According to her study, non-professional translators produce more explanatory or redundant texts. That is why a translated text by a non-professional translator is more *explicit* than a translated text by a professional translator. Séguinot (1988) provides an empirical study using two translated texts from *French* to *English*. There is a greater level of *explicitness* in the translated texts as linking words and conversion of subordinate clauses into coordinate clauses. Later, Baker (1995) and Baker (1996) observed similar characteristics by translators trying to fill the cultural gap.

The evidence of *explicitation* can be found in the length of the texts. Through the addition of new words to make a translated text more explicit, the length of a translated text will be longer than the corresponding source text. Hansen (2003) stated that *explicitation* can also be analyzed in view of lexis and syntax, using a corpus of translated texts and a comparable corpus of original texts in the same language. In the translated texts in English, the use of some words like: 'that' and 'the' show explicitation translation properties.

9.1.2 Simplification

The *simplification* translation property shows the tendency of a translator to simplify a text in order to improve the readability of the translated text. Blum and Levenston (1978) mention the term *simplification* as part of the lexical simplification using a small data set of *English* and *Hebrew* translations. According to them, translators use techniques such as *avoidance* and *approximation* in the translation process to make the translated text simple for the target readers. Later, Baker (1996) also observed this tendency in the translated texts.

To make a translated text simpler, the translator often breaks up complex sentences into two or more sentences. This tendency can be found in the average sentence length. That is, the average sentence length in a translated text will be shorter than a source text.

Punctuation is another linguistic feature that reflects *simplification* (Hansen 2003). Translators change punctuation from a weaker to a stronger mark. For instance, they tend to use *semicolons* or *periods* instead of commas and *periods* instead of *semicolons* (Hansen 2003).

Lexical density is another feature that also reflects the *simplification* translation property. It is the ratio of *lexical words* to *function words*. According to Hansen (2003), the *lexical density* tends to be lower in a translated text than the corresponding source text. More specifically, a translated text will contain more *function words* and fewer *lexical words* than the corresponding source text. *Type-token ratio* also reflects *simplification*. The *type-token ratio* in a translated text will be lower than its counterpart source text. Ilisei, Inkpen, Pastor, and Mitkov (2009) and Ilisei, Inkpen, Pastor, and Mitkov (2010) showed additional features such as different POS that also reflect the *simplification* translation property.

9.1.3 Normalization

The *normalization* property shows a translator's effort to meet the normative criteria of the target language. It is a translator's tendency to conform to patterns and practices that are typical of the target language, even to exaggerate their use. This property can be observed in a translated text that contains very little trace of the source language. However, the opposite scenario can be seen as well, where the translation is influenced by the source language. In that case *normalization* will be weakened. The influence of *English* can be visible in many software manuals that are translated from *English*. Hansen (2003) stated that this contrary tendency also can be seen in *interpreting*, where the interpreter tries to finish an unfinished sentence and to render an ungrammatical structure into something grammatical.

9.1.4 Leveling out

Baker (1996) explains *levelling out* as "“the tendency of translated text to gravitate towards the center of a continuum". That is also known as *convergence* (Laviosa 2002). She explained *convergence* as a "relatively higher level of homogeneity of translated texts with regard to their own scores on given measures of universal features" such as lexical density or sentence length, in contrast to source texts. If we have a sub-corpus of translated texts from different languages to the same language and source texts in the same language then, translated texts from different languages will be similar too each other in terms of *lexical density*, *type-token ratio* and average sentence length, and will differ from the source texts.

9.1.5 Interference

Toury (1995) has a theory that deviates from the translation properties already listed. He states that "in translation, phenomena pertaining to the make-up of the source text tend to be transferred to the target text". That is, some *interference* effects will be observable in translated texts that are carried from source texts. These effects will be in the form of *negative transfer* or in the form of *positive transfer*. As an example, specific properties of the English language are visible in user manuals that have been translated to other languages from English (for instance, word order) (Lzwaini 2003). We can summarise this translation property in a way that translated texts from different source languages will be sufficiently different from each other.

Pastor, Mitkov, Afzal, and Pekar 2008; Ilisei, Inkpen, Pastor, and Mitkov 2009; Ilisei, Inkpen, Pastor, and Mitkov 2010, for example, provided empirical evidence for properties of the translation relation using a comparable corpus of English and Spanish in the medical domain. Obviously, corpora of this sort, which focus on a single language pair, are not adequate for claiming universal validity of properties of the translation relation. Currently, scholars in the area of translation studies lack corpora by which they can validate their theoretical claims. Also a classifier is needed that can classify source and translated texts. A real example of the *interference* effect in translation is available in Koppel and Ordan (2011). They have shown that an under-represented word (e.g., too) in a translated text from Spanish to English is also available in synonym form (e.g., also) in translations from cognate languages such as French and Italian to English.

9.2 Cognitive model of translation process

Translation is a *cross-language* and *cross-cultural* communication activity where mediation between languages has to considered. This is a task of rephrasing a communication that is expressed in a source language (SL) to a target language (TL). Sometimes it refers to an operation that turned a SL unit to a TL unit without considering modality of input and output text. It also refers to reformulating text in a SL to a text of TL.

It is important to produce translations that are well formed, accurate with respect to source texts and socially appropriate in their cultural contexts (Shreve 1997). The translation process consists of three fundamental cognitive tasks, including source language comprehension, meaning transfer from the source language to the target language and finally production of target language texts (Angelone 2010). The Figure 9.1 depicts an ideal translation process. Each of these tasks require different cognitive skills.

Holmes (1988) has proposed a mental approach to the translation process, that is called *mapping theory*. He stated that " I have suggested that actually the translation process is a multi-level process; while we are translating sentences, we have a map of the original text in our minds and at the same time, a map of the kind of text we want to produce in the target language. Even as we translate serially, we have this structural concept so that each sentence in our translation is determined not only by the sentence in the original but by the two maps of the original text and of the translated text which we are carrying along as we translate." So, the translation process is a complex process where *understanding*, *processing* and *projection* of the translated text are independent parts of one structure.

By considering *mapping theory*, Hönig (1991) proposed a model of an ideal translation process. Based on his model, translators read the source text in way that differs from the reading of a ordinary reader. The translator's reading in influenced by the translation task they have in mind. The *source text* has to be moved from its *natural* surroundings and projected into the *mental reality* of the translator in order to be translated (Hönig 1991; Göpferich 2009). The projected *source text* will become an object of mental processing. This process is performed in two different workspaces: the *controlled workspace* and the *uncontrolled workspace.* These workspaces are interdependent. The *understanding* of the *source text* takes place in the *uncontrolled work-space* that involves the activation of *frames* and *schemes.* These *schemes* and *frames* are structured domains of long-term memory in associative processes. The processes give rise to expectations of a prospective target text. These expectations are oriented to the target text and related to structure, style and content.

Using the prospective *target texts*, projected *source text* and collected data from the *uncontrolled workspace* a translator develops a translation *macro-strategy.* The *macro-strategy* includes decisive characteristics for the target text, its functions, its audience, the options that the translator has for searching information to verify their subjective associations. The development of a *macro-strategy* may happen more or less automatically on the basis of the translator's professional experience (Göpferich 2009). The *macro-strategy* precedes the actual translation phase. In the translation phase both workspaces are are involved. The actual translation is the *transfer competence* in Figure 9.1

All of the cognitive processes take place in the *controlled workspace.* According to Hönig (1991), the projected *source text* becomes the *target text* in four different ways:

1. As a linguistic reflex stimulated by the first contact between the projected *source text* and the semantic association in the *uncontrolled workspace.*

2. As an automatic transfer from the *uncontrolled workspace* after a *macro-strategy* has been worked out.

3. As a product of a *micro-strategy* applied in the *controlled workspace* that has been

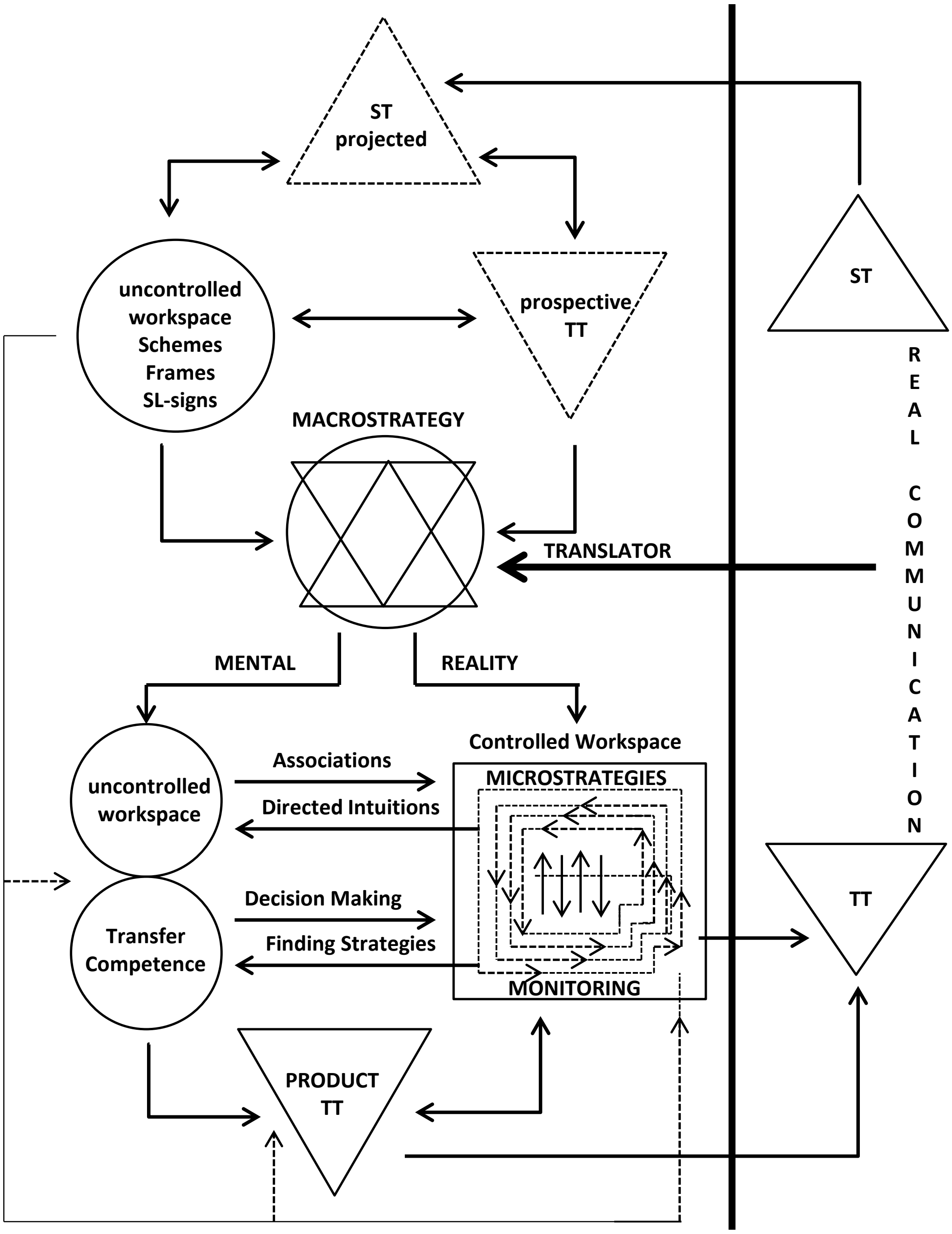


Figure 9.1: Hönig's model of an ideal translation process (Hönig 1991).

approved by monitoring.

4. As a product of interdependent processes taking place in the *controlled* and *uncontrolled* workspace, whereby the final approval can be made either by uncontrolled or controlled processes.

With the decision based on a *macro-strategy*, the *source text* is then translated into the *target text* where each of the sentences is evaluated by deciding whether it fulfills the translator's requirement. Finally, the target text leaves the translator's mental reality and becomes part of a real communication.

Generally, there are two parties involved when text is a medium of communication: *writers* and *readers*. The quality of a text is measured by how good the text conveys information to readers as it was intended by the writer. In the process of translation, another party is involved: *translators*. These three parties differ in language, cognitive abilities and cultures. Their physical environments may overlap in parts, and that might help them converge somewhat on inferential abilities (Chang 2009). However, experience and culture can vary among them according their cognitive environment. In the translation environment, both *writers* and target *readers* are absent. Translators depend on their assumptions about the writer's and readers' shared cognitive model in order to understand the original texts and produce a text that is better for the target readers. A translated text is a product, in which the translator's cognitive environment plays a vital role in the interaction between writers and readers. The translator needs to expand his/her cognitive environment until he/she finds a context that is able to fill the cognitive and cultural gap between a *writer* and a *reader*. In order to make a successful translation, the translator needs to bring out all of the *missing links* in the source text. There are many words that take different meanings in different contexts. It is necessary for a translator to decide the correct meaning that is suitable to the context. Translators need to make a choice in an environment where two languages and two cultures are involved. It is therefore very different than a communication environment where a single language and the same culture are involved. Translators have to consider differences in language and culture during the process of choice making. In the process of translation the cultural difference is more crucial than language differences (Chang 2009).

10 Related Work of Translation Analysis

This is a relatively new sub-field that is being explored in the field of NLP. That is why there are not many previous works available in the area. The field lacks a suitable multilingual corpus that can be used by translation scholars. That is why translation scholars try to compile their own corpora for their study.

At the beginning of corpus based translation studies, Baker (1995) distinguished three types of corpora that are suitable for empirical research on translations, namely: *comparable corpora*, *parallel corpora* and *multilingual corpora*[1]. Fernandes (2006) revisited Baker's typology and rejected the necessity of *multilingual corpora* in translation studies. He claims that Baker's tripartite classification can be rearranged under the categories of *comparable* and *parallel* corpora. As a reason for this binarism, Fernandes (2006) claims that the term *multilingual* is not contrastive enough to distinguish corpora from the other two categories. Moreover, he argues that corpus size is relativized by qualitative aspects, which are sometimes more relevant than quantitative ones. Fernandes (2006) also introduced the attribute *multi-directional* to denote corpora of more than two languages, where the translation direction between language pairs is not predetermined. In this line of terminology, Olohan (2004) focused on *comparable* and *parallel* corpora.

Another example is Lzwaini (2003) who presented a specialized corpus of three languages (English, Arabic and Swedish) in the domain of *Information Technology* (IT). The corpus contains user manuals in English, the online help of Windows 98 and of Microsoft Office 2000 together with their translations into Arabic and Swedish. Additionally, Lzwaini (2003) harvested bilingual text from websites of IT companies. Another example is the *Translational English Corpus* (TEC), which contains contemporary translational English (Baker 2004). The TEC was designed for the purpose of studying translations whose target language is English. It comprises texts of four types of

[1] As Baker (1995, p:232) states: " sets of two or more monolingual corpora in different languages, built up either in the same or different institutions on the basis of similar design criteria"

a variety of European and non-European source languages and contains fiction, biography, newspaper articles and in-flight magazines all of which were translated by native speakers of English. These corpora do not require any kind of customization.

Linguistic annotation is important to build reference corpora for translation studies. Hansen and Teich (2002) show how to build a reference corpus that contains such annotations. They also discuss typical problems that occur in translations from English to German and to French.

None of these corpora contains texts of more than two languages. Thus, claims about the universal validity of properties of the translation relation cannot be tested by means of these corpora. A central deficit of these corpora is that they disregard the diversity of language families and sub-families. Thus we are in need of a *multilingual* and *multi-directional* corpus in order to validate hypotheses in this field of research.

Corpus-based translation studies is a recent field of research with a growing interest within the field of computational linguistics. Translation properties have been studied by different scholars by comparing different patterns between original and translated texts. Gellerstam (1986) compares texts written in Swedish and texts translated from English into Swedish. He observes differences between these two type of texts that do not indicate poor translation but rather a statistical phenomenon. He terms these statistical phenomena as *translationese*.

Baroni and Bernardini (2006) started corpus-based translation studies empirically, where they work on a corpus of geo-political journal articles. They found that a computer system is able to distinguish between original texts and translated texts in the Italian language. A Support Vector Machine (SVM) was used for their experiment and *lexical cues*, distribution of *n-grams* of function words and *morpho-syntactic* properties were used as features. According to their results, word *bigrams* play an important role in the classification task. They noticed that *personal pronouns* and *adverbs* are influential for the task. So, shallow data representation can be sufficient to build a classifier that can identify original and translated texts. Van Halteren (2008) uses the *Europarl* corpus for the first time to identify the source language of text for which the source language marker was missing. He focused on texts from six of the most common European languages such as English, German, French, Italian, Spanish and Dutch. For each of these languages, he extracted a parallel six-lingual subcorpus of original texts and their translations into the other five languages. The task is to identify the source language of translated texts, and the reported results are excellent. Support vector regression was the best performing method.

Pastor, Mitkov, Afzal, and Pekar (2008) use a corpus based approach that tests statistical significance of their proposed features in order to investigate *simplification* translation properties on Spanish texts. The experimental texts belong to the medical domain and were produced by professional and semi-professional translators. Their experimental results show that a translated text exhibits lower lexical density and richness than the corresponding original text. Also, the translated text is more readable and contains a smaller proportion of simple sentences and appears to be significantly shorter. The *simplification* properties are more visible in texts translated by professional translators compared to those translated made by semi-professionals.

Ilisei, Inkpen, Pastor, and Mitkov (2009) and Ilisei, Inkpen, Pastor, and Mitkov (2010) also use Spanish original and translated text in order to check *simplification* translation properties. They use some features (e.g., ratio of grammatical words to content words) that are able to capture *general* characteristics of texts. Then they have added nine features that relate to *simplification* translation properties. Among all of their features, *lexical variety*, *sentence length* and *lexical density* are the most informative features.

Ilisei and Inkpen (2011) investigate the effect of translation properties on Romanian newspaper texts. They have trained a classifier that is able to identify translated and non translated texts. The classifier uses 38 language independent features from a POS tagger's output in addition with some features that reflect simplification translation properties. Different classification algorithms were used and SVM performs the best. Ilisei (2013) experiments with *explicitation* translation properties using texts from Spanish and Romanian. She uses features whose values are proportion of the some POS categories in texts.

Koppel and Ordan (2011) have built a classifier that can identify the correct source of the translated text (given different possible source languages). They have built another classifier which can identify source text and translated text. Furthermore, they have shown that the degree of difference between two translated texts, translated from two different languages into the same target language, reflects the degree of difference of the source languages. They have shown impressive results for both of the tasks. However, the limitation of this study is that they only used a corpus of English original text and English text translated from various European languages. A list of 300 function words (Pennebaker, Francis, and Booth 2001) was used as a feature vector for these classifications.

Popescu (2011) uses *string kernels* (Lodhi, Saunders, Shawe-Taylor, Cristianini, and Watkins 2002) to study translation properties. A classifier was built to classify English original texts and English translated texts from French and German books that were written in the nineteenth century. The *p-spectrum* normalized kernel was used for the experiment. The system works on a character level rather than on a word level. The system performs poorly when the source language of the training corpus is different from the one of the test corpus.

Volansky, Ordan, and Wintner (2013) classify original English texts and translated English texts from ten European languages. They compiled a corpus from the same source we use in this thesis. They used SVM as the classification algorithm and 10 fold cross validation for testing. This study tested several translation properties by proposing different feature sets that indicate those properties. They also used some features that were proposed in a previous study of Popescu (2011) and Ilisei, Inkpen, Pastor, and Mitkov (2010).

In the same line of research, recently Avner, Ordan, and Wintner (2014) investigated translation properties in Hebrew original and translated texts from English. They proposed linguistically motivated feature sets that are able to capture translation properties in Hebrew. As features, they have explored *function words*, *word unigrams*, different *morphological features*, *POS-tags* and more.

Some of our proposed features are already used in some of the previous studies listed above. Many of these features are indicators of *simplification* translation properties. Translated texts are more readable compared to original texts (Pastor, Mitkov, Afzal, and Pekar 2008). That is why our proposed feature will be useful for this task. The following chapter describes the corpus we use to study translation.

11 Customized Europarl Corpus

At the beginning of this chapter, it was necessary to disambiguate the term *corpus*. Generally, in the field of translation studies a *corpus* is a resource that can aid a translator during the translation process. In this study, the term *corpus* is different than a resource for translators. We define the term *corpus* as a resource that can be used by *translation scholars* to validate their finding in theoretical research. In recent years, many translation scholars proposed some translation properties (universals) using some corpora that are *monolingual* or *bilingual*. However, the field of translation studies lacks such a corpus that contains *translated* and *source* texts from many different languages. In this thesis, we provide a corpus that contains parallel *source* and *translated* texts from 21 European languages. The corpus is extracted and fine tuned from a well known resource called the *Europal* corpus (Koehn 2005). The following subsections describe the corpus in detail.

11.1 Customized Europarl for translation studies

The *Europarl* corpus is a *multilingual*, *parallel* corpus that has been collected from the proceedings of the *European Parliament* since 1996. The corpus contains the minutes of the European parliament, where members of the parliament speak in their native language. All of the speeches are transcribed, edited and translated into other official languages of the parliament. The official languages of European parliament changed over time. The corpus we use in this thesis contains texts from 21 European languages. The corpus is compiled by Koehn (2005) with an intention to use the data for statistical machine translation. The latest version of the corpus contains about 60 million words for each official language of the *European Union* (EU). According to Cartoni and Meyer (2012), in 2004, *English* was introduced as a pivot language where all statements are first translated to English and then to the other languages. The *European parliament* provides different information[1]. According to them, the translation process is *direct*,

[1] http://www.europarl.europa.eu/multilingualism/trade_of_translator_en.htm

Language (sub-)family	Language names
Germanic	English, German, Dutch, Danish and Swedish
Romance	French, Italian, Spanish, Portuguese and Romanian
Slavic	Czech, Bulgarian, Polish, Slovak and Slovenian
Baltic	Latvian and Lithuanian
Finnic	Finnish and Estonian
Ugric	Hungarian
Hellenic	Greek

Table 11.1: Sub-families of languages and their members that are included in the *Europarl* corpus.

that is, translation is done from one language to another. If there is no translator for a language pair, *German*, *French* or *English* is used as a pivot language. In most cases, the translation service in the *European Parliament* lets translators translate into their native language(Van Halteren 2008). The corpus is annotated with `<CHAPTER id>` to identify documents, with `<SPEAKER id name language>` to identify source languages and with `<P>` to segment paragraphs. The procedure of corpus collection is described in Koehn (2005). Sentences in the *Europarl* corpus are aligned by using the sentence alignment algorithm described in Gale and Church (1993).

The customizing procedure of the *Europarl* corpus for translation studies was done in multiple steps – see Figure 11.1 for a visual depiction of this procedure. Although it has been available since 2001, this corpus has not been used by translation scholars. A reason might be its deficient customization regarding the task of corpus-based translation studies. Our goal was to customize this corpus in a way that translation scholars can use it without further pre-processing. Note that the *Europarl* corpus is diverse as it contains texts from 21 languages from 7 language (sub-)families. Table 11.1 shows these languages and their family memberships. Figure 11.1 outlines the procedure of corpus customization. The following subsections describe this procedure in more detail.

The selection of language pairs is decisive when building a multilingual and multi-directional corpus that reflects the diversity of natural languages. There is neither theoretical nor practical research in the field of corpus-based translation studies on how to select such pairs for building *multilingual* parallel corpora. We address this deficit in this thesis. Our intention is to make the corpus as diverse as possible by considering a broad range of language (sub-)families. We considered all possible language pairs.

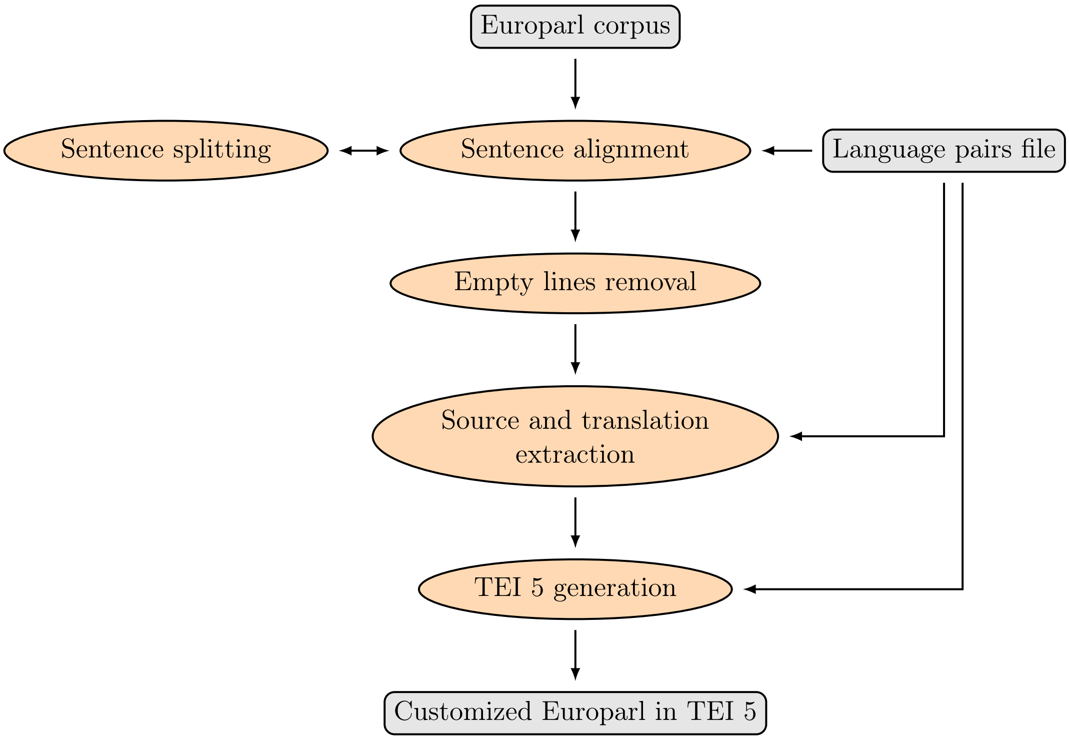


Figure 11.1: Building a customized version of the *Europarl* corpus: extraction steps.

11.1.1 Speaker name and native language

The language annotation in the original *Europarl* corpus is not reliable because of erroneous annotations. There are many cases where one speaker has multiple speeches in different languages that cause problems when trying to identify the speaker's native language. Table 11.2 depicts the language variation of a single speaker. The language codes are presented using the *ISO 639-1* language codes (Alvestrand 1995). The table shows that *Christian Silviu Buşoi* from *Romania* speaks 21 languages. Also *Alain Lipietz* from the *French Green Party* and *Danuta Jazłowiecka* from *Poland* have given speeches in 13 different languages. As a solution to the problem of identifying the native language, we can count the frequency of the languages for each speaker. We can assign the most frequent language as the speaker's native language. Table 11.3 shows the language frequencies of the same speakers as in Table 11.2. According to table 11.3, *Romanian* is the native language of *Christian Silviu Buşoi*. However, further problems arise due to variations in the name of a speaker. We use the *Levenshtein distance* (Levenshtein 1966) to identify name variations. The allowed distance is *one-fourth* of the candidate name. We experimented with different values for the distance but *one-fourth* of the candidate name gives the best result. Table 11.4 shows the name variation of *Christian Silviu Buşoi*.

Speaker	Languages
Cristian Silviu Buşoi	LV, CS, IT, LT, FR, RO, FI, SV, PT, PL, SL, SK, DE, HU, ES, DA, GA, EN, EL, NL, BG
Andersson e Waidelich	DE, FI, FR, EN, IT, PT, ES
Alain Lipietz	FR, HU, DE, SV, ES, PT, NL, EN, EL, SK, PL, LT, IT
Catherine Bearder	FR, DE, SV, PT, DA, NL, EN, EL, PL, SK, BG, IT
Danuta Jazłowiecka	RO, CS, FR, DE, ES, GA, FI, PT, NL, EN, EL, PL, IT
George Lyon	CS, FR, HU, DE, SV, ES, PT, EN, EL, PL, IT

Table 11.2: Speaker with speeches in multiple languages.

According to the table, different languages are selected as the most frequent language for the different name variations, which leads to another problem. Again, if we consider language frequency, then *Romanian* is the language of *Christian Silviu Buşoi*. But this is not always the case. Many speakers (including name variations) are assigned the wrong language if we resolve the *name* and *language* annotations in this way. According to manual observation, typing mistakes are the main reason for these name variations.

We also extracted name variations first and counted frequency of languages afterwards to find the native language of a speaker. However, this process also failed due to the *language annotation* problem. Finally we decided to collect the name of the member of the *European parliament* and their native language manually. We collected names from the current members list page [2] of the *European parliament*. Names of former members are collected from the corresponding *Wikipedia* pages. The official language of the country of each member is assigned as the native language of a speaker. Members from *Belgium* and *Luxemburg* are not considered as we are not sure about the language they speak in the parliament. Each member from *Finland* is assigned to the *Finnish* language. By this criteria, we drop speeches where any parliament member from Finland speaks any other language than Finnish. Finally, the list contains 2,125 member names and their native languages. This list was extended by finding name variations in the *Europarl* corpus. Any name with a *Levenshtein distance* of three was considered a variation of that name. The final speaker list was checked manually and erroneous speaker names were deleted.

11.1.2 Extracting source sentences and their translations

The *Europarl* corpus comes as plain text with additional marking of the document, speaker and paragraph ids. The *Europarl* community also provides a pre-processor (e.g., a sentence splitter, tokenizer) together with a *sentence aligner* based on Gale and Church (1993). The details of the pre-processor are described in Koehn (2005). *Sentence*

[2] http://www.europarl.europa.eu/meps/en/full-list.html

Speaker	Languages frequencies
Cristian Silviu Buşoi	LV = 272, CS = 86, IT = 464, LT = 48, FR = 690, RO = 4580, FI = 12, SV = 196, PT = 94, PL = 222, SL = 4, SK = 2, DE = 1226, HU = 100, ES = 296, DA = 104, GA = 8, EN = 516, EL = 386, NL = 266, BG = 38
Andersson e Waidelich	DE = 5, FI = 7, FR = 22, EN = 10, IT = 5, PT = 7, ES = 28
Alain Lipietz	FR = 1274, HU = 36, DE = 304, SV = 32, ES = 78, PT = 46, NL = 8, EN = 33, EL = 8, SK = 10, PL = 36, LT = 40, IT = 72
Catherine Bearder	FR = 312, DE = 56, SV = 72, PT = 104, DA = 18, NL = 108, EN = 330, EL = 18, PL = 6, SK = 6, BG = 96, IT = 96
Danuta Jazłowiecka	RO = 32, CS = 8, FR = 100, DE = 130, ES = 24, GA = 4, FI = 4, PT = 46, NL = 120, EN = 156, EL = 316, PL = 484, IT = 148
George Lyon	CS = 8, FR = 434, HU = 160, DE = 236, SV = 6, ES = 112, PT = 474, EN = 566, EL = 136, PL = 12, IT = 286

Table 11.3: Speaker with speech frequencies in different languages.

Name Variation	Most frequent language
Christian Silviu Buşoi	LV = 20
Cristian Silviu Buo̲oi	RO = 2
Cristian Silviu Buŗoi	PL = 8
Cristian Silviu Buşoi	RO = 4580
Cristian Silviu Buṣoi	DE = 36
Cristian Silviu BuÍoi	RO=2

Table 11.4: Variations in speaker name with highest language frequency.

alignment is the first step of corpus customization, and starts with reading language pairs from the input corpus (see Figure 11.1). Source texts and their translations are iteratively processed to align their sentences. In this stage, the sentence splitter is used to detect sentence boundaries. At the beginning of the extraction process, the corpus is aligned by using the tool provided by the *Europarl* community. However, the tool generates empty lines in cases where there is no parallel sentence for a candidate sentence. These empty lines and their corresponding sentences must be deleted. The next step is the core of the corpus extraction procedure, namely the step of extracting source sentences and their translations. In this step, we extract sentences only for those speakers who are in the input speaker list. The language of the speaker in the list and in the corpus has to be the same. Van Halteren (2008) showed that there are many cases where language annotations are missing in the source side of the corpus. To circumvent this problem, we solely extracted pairs of sentences for which the source language marker is available. As an output of this customization step, we got 2,646,765 source sentences together with their corresponding translations. It should be noted that there are many re-occurring sentences on both sides of the corpus. Table 11.5 shows the results of the corpus extraction process.

There is no translation available from *English* to eight of the languages due to missing language annotations in the translation. To circumvent this problem, we selected only speakers whose native language is *English*. Speeches of these speakers were extracted from both sides of the corpus. The number of extracted sentences are much larger than some mostly spoken language pairs. This is another example of a language annotation problem in the *Europarl* corpus.

	BG	CS	DA	DE	EL	EN	ES	ET	FI	FR	HU	IT	LT	LV	NL	PL	PT	RO	SK	SL	SV
BG	–	494	655	618	601	599	561	427	423	497	610	570	578	514	437	688	552	643	517	483	540
CS	1,632	–	2,048	1,616	1,513	1,692	1,576	1,331	1,375	1,363	1,315	1,414	1,544	1,525	218	1,666	2,529	1,477	1,826	1,326	1,480
DA	688	787	–	1,715	1,084	1,662	1,591	818	1,543	1,586	606	1,955	975	1,060	951	832	1,440	746	1,183	272	1,516
DE	11,645	10,945	30,344	–	13,259	20,921	20,647	9,719	17,235	19,795	9,618	20,831	12,439	12,298	14,320	13,045	21,495	11,356	14,363	8,690	19,572
EL	600	1,249	1,627	1,235	–	2,264	1,978	1,197	2,021	1,843	1,175	1,954	1,181	1,343	2,515	1,016	2,418	587	1,829	992	1,937
EN	8,209	–	21,344	6,387	5,350	–	6,438	–	9,184	11,839	–	14,145	–	–	8,114	–	13,095	–	–	–	9,983
ES	1,118	1,470	3,890	3,735	1,506	2,493	–	1,241	2,096	2,328	1,196	2,829	1,636	1,605	2,302	1,649	2,745	1,049	1,742	1,011	2,302
ET	287	273	481	320	273	326	266	–	254	249	200	328	316	278	385	317	325	294	443	151	270
FI	1,031	1,047	2,910	2,101	1,169	2,102	1,878	600	–	1,901	646	2,147	1,125	924	3,059	1,010	1,987	967	1,539	402	1,873
FR	10,266	10,986	20,271	15,344	11,157	16,847	14,396	8,414	13,997	–	7,707	16,170	11,401	10,131	9,533	11,336	15,581	10,039	12,712	7,713	14,250
HU	1,593	1,587	2,758	1,930	1,499	1,768	1,584	1,361	1,416	1,470	–	1,749	1,667	1,679	468	1,620	1,759	1,546	1,995	952	1,619
IT	1,785	2,993	5,646	4,122	2,999	4,744	3,798	2,897	3,610	3,527	2,835	–	3,650	2,946	2,823	3,334	5,036	2,518	4,069	2,769	4,073
LT	363	608	788	557	504	878	654	571	666	682	688	658	–	771	691	586	825	299	885	612	756
LV	110	140	233	149	157	184	167	166	145	179	183	201	177	–	176	193	161	131	202	122	187
NL	3,437	4,150	13,557	7,126	4,539	7,873	7,067	3,558	6,235	7,330	3,681	7,859	4,481	3,720	–	4,273	9,056	3,448	5,887	3,105	7,325
PL	3,397	3,942	4,795	3,847	3,717	4,419	4,226	3,235	3,644	3,347	3,070	3,912	3,982	4,035	1,442	–	4,334	3,235	4,583	2,784	3,950
PT	5,401	5,461	10,071	6,478	6,520	8,494	7,850	4,726	5,783	6,556	4,916	7,708	6,199	5,526	5,292	6,075	–	5,070	6,067	3,906	7,035
RO	3,766	3,178	3,545	3,745	3,564	3,842	3,547	2,757	3,386	2,932	2,426	3,232	3,610	3,289	2,502	3,821	3,545	–	3,783	1,871	3,487
SK	968	857	1,284	1,023	877	1,080	984	896	921	711	718	853	1,012	1,055	691	936	1,607	915	–	290	849
SL	182	183	267	227	165	197	161	183	51	156	217	159	195	170	104	184	207	182	204	–	193
SV	1,763	3,298	7,456	5,083	3,678	5,495	4,583	2,404	5,032	5,660	1,342	5,845	3,565	3,910	3,065	3,627	5,466	3,461	3,627	3,138	–

Table 11.5: Corpus statistics of the customized Europarl corpus for translation studies.

Listing 11.1: Corpus sample

```
<TEI>
 <teiHeader>
  <srcLang>de</srcLang>
  <trnLang>en</trnLang>
 </teiHeader>
 <documents>
    <document title="ep-09-12-15-015">
      <segment id="42060">
              <srcSent id ="3">Diese Antwort haben wir im Prinzip gerade erhalten .</srcSent>
              <trnSent id ="3">We have just received this response in principle .</trnSent>
      </segment>
    </document>
 </documents>
</TEI>
```

11.1.3 TEI generation

The next step is to provide the data in a machine readable way that can be easily processed by translation scholars. TEI P5 TEI Consortium 2008 is used to represent the extracted corpus. Listing 11.1 shows a sample of the customized corpus. Translation scholars can use the whole or a subset of some specific language pairs. The customized corpus consists of a single file in which information about the source and target language of a translation is specified in the header of each <TEI> section. The complete corpus contains 2,646,765 segments and, thus, 2,646,765 sentences of 21 languages and their translations.

The compiled corpus is suitable for translation studies where translation scholars validate their theoretical findings. Translation scholars already have proposed some translation properties, and the features selected in this study are indicators of some of these properties. The following chapter describes the experiments in detail.

12 Experiment of Source and Translation Analysis

Translation scholars proposed different translation properties that are described in the previous sections. Chapter 6 describes a variety of features for both of the tasks we consider in this thesis. The proposed features have already been shown to be useful for text readability analysis. In this chapter, we use the same feature set for two different experiments. The first experiment intends to validate the claim by Toury (1995), where is stated that there will be some transfer into the translated text from the original text. That is why translated texts into the same language from different languages will be sufficiently different. Section 12.1 shows the details experimental result.

The second experiment is about the translation properties proposed by Blum-Kulka (1986), Baker (1993), Vinay and Darbelnet (1995), Baker (1995), and Baker (1996). According to them, a translated text is very different from the corresponding original text. The following sections describe in detail experimental results of our proposed features in *source and translation* analysis.

12.1 Source identification

In order to perform the source identification experiment, we selected ten European languages from seven languages familie. Koppel and Ordan (2011) have shown that a list of function words is useful to identify the original language of a translated text. A list of function words is language dependent. A classifier will require a new list of function words whenever it needs to deal with a new language. In this thesis, we try to resolve the problem by using features that are language and linguistic tools independent. In this experiment, our hypothesis is that the *interference* translation property does not exist if the classification *F-score* is close to 11%. That is the the probability of translating a text by chance.

The *Europarl* corpus contains text from 21 language that belong to 7 different language

	Czech	German	Greek	Spanish	Finish	French	Hungarian	Lithuanian	Dutch	Polish
Czech	–	236,572	33,032	42,769	21,840	270,224	35,386	13,436	85,003	85,741
German	43,994	–	36,690	119,537	47,365	423,198	50,537	10,001	169,363	97,398
Greek	42,755	348,350	–	50,098	27,733	332,599	44,345	10,912	118,527	103,892
Spanish	42,920	570,848	62,398	–	47,056	423,010	42,915	17,496	189,830	119,310
Finish	25,167	327,370	48,344	51,793	–	290,762	29,760	11,259	113,454	68,955
French	38,542	572,932	63,148	79,858	50,499	–	44,552	17,571	206,790	94,341
Hungarian	30,099	201,643	28,348	31,556	12,511	191,462	–	13,774	75,919	67,935
Lithuanian	31,733	250,034	25,539	45,534	19,738	261,740	37,345	–	87,802	79,402
Dutch	5,748	361,042	81,661	77,320	70,393	276,494	10,971	16,177	–	38,438
Polish	39,096	277,903	28,928	49,544	21,342	283,099	40,005	9,059	90,751	–

Table 12.1: Source language identification corpus words per language pair.

	Czech	German	Greek	Spanish	Finish	French	Hungarian	Lithuanian	Dutch	Polish
Czech	-	199	200	200	200	200	200	200	200	200
German	200	-	200	200	200	199	200	200	200	200
Greek	200	196	-	200	200	197	200	200	200	200
Spanish	200	196	200	-	200	198	200	200	200	199
Finish	200	195	200	200	-	198	200	200	196	200
French	200	197	197	196	200	-	200	200	200	200
Hungarian	200	196	200	200	200	199	-	200	200	200
Lithuanian	200	196	200	200	200	196	200	-	199	200
Dutch	200	199	200	198	200	199	200	200	-	200
Polish	200	197	200	200	200	199	200	200	200	-

Table 12.2: Chunks for the source language identification task.

families. We we could not use all the language pairs due the missing texts for some pairs. Also, some of language pairs have very few sentences. That is why we consider texts from 10 languages. These 10 languages belong to 7 language families. Table 12.1 shows the languages used for this experiment and the words per language pair. In this experiment, our goal is to build a classifier using our proposed features that is able to classify translated texts according to their original language. We assume that the data used here contains little or no errors. We use a similar strategy as used by Koppel and Ordan (2011) in dividing the data into chunks. Table 12.2 shows the number of chunks per language pairs. The following sub-sections describe the performance of different feature sets.

12.1.1 Baseline system

Generally, a translated text is more readable than its original counterpart (Pastor, Mitkov, Afzal, and Pekar 2008). That is why we could use traditional readability formulas as features. However, some of the traditional formulas used in this thesis require a syllable count in a word. As we are using texts from different languages, identifying syllables in the different languages will be different. Therefore, we only use three readability formulas that do not require syllable information for building the baseline system. These are: the *Gunning fog index* (Gunning 1952), the *Dale-Chall readability formula* (Dale and Chall 1995) and the *ARI* (Senter and Smith 1967).

	Czech	German	Greek	Spanish	Finish	French	Hungarian	Lithuanian	Dutch	Polish
Czech	-	33.44	34.27	45.67	08.36	38.80	00.32	18.17	28.97	20.58
German	00.45	-	33.53	42.08	22.47	38.13	00.83	37.69	37.57	10.00
Greek	06.22	33.88	-	39.97	01.96	34.38	01.53	33.56	12.44	02.04
Spanish	00.47	29.14	36.30	-	25.92	28.71	00.00	17.29	24.35	04.58
Finish	03.45	42.44	35.45	40.11	-	40.28	00.96	43.73	33.97	21.70
French	02.25	34.00	31.32	38.07	18.43	-	01.95	18.76	28.07	19.58
Hungarian	15.35	35.99	15.06	43.65	28.57	35.28	-	15.13	24.13	24.71
Lithuanian	13.50	36.41	33.88	42.53	25.43	34.36	00.62	-	26.33	24.15
Dutch	41.90	50.87	42.76	32.21	41.72	42.73	14.43	09.32	-	00.42
Polish	11.15	38.44	33.42	39.76	23.89	39.63	01.09	26.39	03.37	-

Table 12.3: Traditional formulas in source language identification (F-score).

Language	Accuracy	F-score
Czech	29.77	25.27
German	29.84	24.71
Greek	25.57	18.35
Spanish	23.33	18.46
Finnish	33.15	0.28.94
French	25.31	21.27
Hungarian	28.91	26.38
Lithuanian	29.37	26.31
Dutch	36.12	30.61
Polish	29.45	23.30

Table 12.4: Weighted average performance of the baseline system in source identification.

Another idea behind using traditional readability formulas to build a baseline system is that these traditional formulas use surface features that represent the different translation properties described above. Table 12.3 shows the *F-score* per language pair and Table 12.4 shows the weighted average of *accuracy* and *F-score*. The performance of the baseline system is poor. The classifier does not achieve a 50% of *F-score* for any language pair. Also, the classifier is unable to identify the correct source of a single document for some language pairs. The most important observation is that the classifier does not work for the translated texts to any of the languages from *Hungarian* and *Czech*. For many language pairs the classifier achieves less than 1% classification accuracy.

12.1.2 Lexical features

Lexical features have proven to be useful for text readability classification. Some of them are indicators of different translation properties described above. For example: in order to make a translation simpler, a translator will tend to make sentences shorter instead of longer. So, average SL will be a good indicator to identify source and translated texts. Also, a translated text exhibits lower lexical density and richness than the corresponding original text. Table 12.5 shows the *F-score* per language pair and Table 12.6 shows the

	Czech	German	Greek	Spanish	Finish	French	Hunga-rian	Lithu-anian	Dutch	Polish
Czech	-	78.04	84.33	66.00	91.43	75.08	66.96	99.94	75.40	70.72
German	93.65	-	99.25	80.86	80.75	100.00	76.65	99.81	100.00	81.82
Greek	46.25	100.00	-	57.99	99.66	100.00	24.01	100.00	100.00	100.00
Spanish	54.98	100.00	79.47	-	78.35	100.00	53.58	100.00	100.00	100.00
Finish	88.40	99.94	67.47	66.15	-	99.94	88.30	99.97	100.00	100.00
French	97.34	100.00	77.83	98.80	79.68	-	96.57	100.00	100.00	100.00
Hungarian	96.49	100.00	61.16	58.35	58.54	100.00	-	52.92	99.77	99.77
Lithuanian	85.73	94.97	69.33	64.87	74.67	94.59	63.41	-	99.79	99.56
Dutch	99.92	99.79	83.85	83.65	100.00	99.79	99.15	99.24	-	100.00
Polish	46.05	99.83	74.49	60.92	76.39	99.82	28.80	100.00	100.00	-

Table 12.5: Lexical features in source language identification (F-score).

Language	Accuracy	F-score
Czech	78.81	78.66
German	90.76	90.73
Greek	81.63	80.77
Spanish	85.10	85.03
Finnish	89.95	89.94
French	94.47	94.49
Hungarian	80.74	80.67
Lithuanian	82.87	82.88
Dutch	96.18	96.17
Polish	76.67	76.12

Table 12.6: Weighted average performance of lexical features in source identification.

weighted average of *accuracy* and *F-score* of the lexical features.

Eleven lexical features are used for this experiment. The classifier achieves better classification accuracy when the source and translated languages belong the same language family. The classification accuracy of translated texts from *Czech* and *Hungarian* improves. However, the classifier achieves only a low classification *F-score* for translated texts from *Hungarian* to *Greek* and *Polish.* For some language pairs the classifier using these lexical features achieves a 100% classification *F-score.* The weighted average of *accuracy* and *F-score* of a classifier using lexical features for all translated languages is more than 75%. It is hard to find the best performing lexical features in this kind of experiment where texts from many different language pairs are used.

12.1.3 Information-theoretic features

Information density could be different in translated texts translated from different languages. This type of property can be captured by *information-theoretic* features. *Information theoretic* features have proven useful for text readability classification. In this experiment, we have merged *entropy-based* and *information transmission-based* features. In total seven features are used in this experiment: *word entropy*, *word frequency en-*

	Czech	German	Greek	Spanish	Finish	French	Hunga- rian	Lithua- nian	Dutch	Polish
Czech	-	68.89	51.06	57.32	65.21	68.61	39.04	85.95	64.99	63.42
German	38.21	-	78.04	74.75	57.15	100.00	84.01	99.16	100.00	73.20
Greek	28.97	90.56	-	50.76	78.24	90.74	36.88	95.33	86.69	87.40
Spanish	38.56	99.38	58.73	-	52.68	99.38	37.55	97.37	100.00	100.00
Finish	56.25	96.82	57.88	62.42	-	96.80	52.81	94.07	100.00	98.98
French	57.64	100.00	54.21	69.84	48.44	-	31.52	94.87	100.00	87.23
Hungarian	53.99	90.75	29.52	39.04	57.20	90.93	-	46.72	78.98	79.89
Lithuanian	39.37	71.36	55.93	56.78	66.31	70.14	44.80	-	75.62	76.54
Dutch	78.56	98.54	59.50	63.30	81.68	98.55	57.07	71.04	-	95.56
Polish	34.55	84.99	63.47	57.15	64.40	85.44	34.90	94.07	100.00	-

Table 12.7: Entropy-based features in source language identification (F-score).

Language	Accuracy	F-score
Czech	63.03	62.70
German	74.54	74.21
Greek	72.15	71.62
Spanish	76.15	75.76
Finnish	79.53	79.41
French	71.81	71.50
Hungarian	63.41	62.85
Lithuanian	62.01	61.78
Dutch	66.49	66.28
Polish	58.44	58.05

Table 12.8: Weighted average performance of information-theoretic features in source identification.

tropy, *word length entropy*, *character entropy*, *character frequency entropy*, *information transmission of SL and WL probability* and *information transmission of SL and DW probability*. Table 12.7 shows the *F-score* per language pair and Table 12.8 shows the weighted average of *accuracy* and *F-score* of the information-theoretic features.

These features do not perform as well as lexical features but the performance is comparable for some language pairs. Although the *F-score* is very low, the classifier performs better for *Polish* translated texts from *Hungarian* than the classifier using *lexical features*. The classifier also achieved a 100% classification *F-score* for some language pairs. A classifier using *information-theoretic* features achieves better classification than the baseline system. One important observation is that there is a big drop in the classification *F-score* for translated *French* texts from *Hungarian* compared to the classifier using lexical features.

Table 12.9 and Table 12.10 show the performance of all features together. The classifier achieves satisfactory classification accuracy for most of the languages pairs. However, the classifier does not perform well for four language pairs such as: *Czech* to *Greek*, *Czech* to *Polish*, *Hungarian* to *Greek* and *Hungarian* to *Polish*. It should be noted that

	Czech	German	Greek	Spanish	Finish	French	Hungarian	Lithuanian	Dutch	Polish
Czech	-	77.90	85.06	65.65	93.07	75.23	65.61	99.58	76.08	71.20
German	90.60	-	99.24	80.56	85.01	100.00	75.14	99.75	100.00	81.95
Greek	44.40	100.00	-	57.48	99.27	100.00	30.31	100.00	99.74	99.75
Spanish	52.88	100.00	78.83	-	78.58	100.00	53.84	100.00	100.00	100.00
Finish	87.19	99.97	68.46	68.03	-	99.97	86.91	99.85	100.00	99.98
French	94.72	100.00	76.81	97.39	78.92	-	93.77	100.00	100.00	100.00
Hungarian	94.99	99.98	58.99	55.03	60.45	99.98	-	55.99	99.91	99.91
Lithuanian	82.61	93.87	68.56	63.86	72.97	93.53	64.78	-	99.75	99.51
Dutch	99.89	99.96	83.25	82.84	99.44	99.96	99.05	99.17	-	100.00
Polish	47.33	99.04	72.80	60.32	74.93	99.41	34.22	100.00	100.00	-

Table 12.9: All features in source language identification.

Language	Accuracy	F-score
Czech	78.94	78.82
German	90.26	90.22
Greek	81.55	81.15
Spanish	84.90	84.77
Finnish	89.98	89.96
French	93.49	93.53
Hungarian	80.51	80.47
Lithuanian	82.05	82.05
Dutch	95.98	95.96
Polish	76.65	76.36

Table 12.10: Weighted average performance of all features in source identification.

the proposed features are language and linguistic tools independent.

The classifier using all of the proposed features perform the least well for the translated texts to *Greek* and *Polish* from *Hungarian*. The *F-score* of these two language pairs is around 30%.

Based on the experimental results, we can say that *interference* translation properties might exist. For some of the language pairs, that classifier performed poorly. This could happen due to the amount of data available for these language pairs. More specific experiments have to be considered in order to find out the reason. The following section describes the other experiment.

12.2 Source and translation identification

As we described above, the task of source and translation identification is itself a multilingual application. It can be a classification task, attempting to place *source* and *translation* in their proper classes. We used a similar strategy in building a baseline system using traditional readability formulas. The following subsections will show the usefulness of different feature sets. However, these sections are focused on general trans-

Readability formula	Accuracy	F-score
Gunning Fog Index	51.90%	40.56%
Dall-Chall formula	56.26%	56.06%
ARI Index	52.93%	46.23%
All together	46.51%	45.23%

Table 12.11: Evaluation of traditional readability formulas in source and translation identification.

lation properties. In this kind of experiment, our hypothesis is that if the classification accuracy is close to 100 percent then we can say that there are some general effects on translation that make them distinguishable from original texts.

The customized *Europarl* corpus contains 2,646,765 parallel sentences from 412 language pairs of 21 European languages. There are many documents that only contain one or two sentences. In order to avoid a data sparseness problem, we merged all documents and later divided them into chunks. A chunk contains 100 sentences. Each of the categories (source or translation) contains 26,467 chunks. More specifically, 26,467 chunks were *source* texts and the same number of chunks were *translations*. There are many chunks that contain texts from two languages. We followed a similar strategy where hundreds of sets of data were randomly generated, in which 80% of the corpus is used for training and the remaining 20% for evaluation.

12.2.1 Baseline system

The baseline system uses similar features as source identification. Only three formulas are used. Surface features such as average SL and average WL are quantitative indicators of the *simplification* translation property. Table 12.11 shows the evaluation of traditional readability formulas in source and translation identification. Here our hypothesis is that a feature is useful when the classifier achieves close to 100% and a feature is not useful if the classification accuracy is around 50%. Among three of these features the *Dale-Chall* (Dale and Chall 1948; Dale and Chall 1995) readability formula performs best. However, the classifier using these features achieves a classification accuracy below 50%.

12.2.2 Lexical features

Lexical features performed best for source language identification from translated texts. Table 12.12 shows the evaluation of lexical features in identifying source and translated texts. Here, average SL and average WL do not perform as well as they performed in text

Features	Accuracy	F-score
Average SL	54.01%	53.29%
TTR (document)	59.83%	59.18%
TTR (sentence)	58.93%	57.42%
DW (document)	52.61%	45.74%
DW (sentence)	52.52%	45.83%
Average WL	56.15%	49.43%
Köhler–Gale TTR	59.58%	58.89%
Root TTR	62.67%	62.67%
Corrected TTR	62.61%	62.61%
Bi-logarithmic TTR	62.23%	62.08%
TTR deviation	60.54%	60.00%
All together	78.47%	78.40%

Table 12.12: Evaluation of lexical features in source and translation identification.

readability classification. TTR related features perform better in this experiment, which show that *lexical density* is an important factor in original and translated texts. Ilisei, Inkpen, Pastor, and Mitkov (2010) also observed similar behavior of features related with *lexical density.* Among all of these features the *Root TTR* performs best. The classification accuracy using these 11 features improves around 73% compared to the baseline system.

12.2.3 Information-theoretic features

One of our several assumptions about translated texts is that a translator always tries to make a translation simpler; that is, more readable. So, the word entropy of a translated text will be lower than the corresponding original text. Table 12.13 shows the performance of individual information-theoretic features, and then all together. In this experiment, this feature set performed slightly better than the *lexical* feature set. However, in the previous experiment source identification, a classifier using *lexical* features performs better than a classifier using an *information-theoretic* feature set. *Entropy* related features perform better than *information transmission-based* features. Among all these, *word entropy* is the best performing individual feature. Seven *information-theoretic* features achieve an accuracy more than 78%. We achieve 86.63 classification accuracy when we add *information-theoretic* features with lexical features.

A classifier using our proposed features is able identify original and translated text with a reasonable accuracy. There are many documents that were translated centuries ago. Currently, the origins of these kinds of documents are unknown or doubtful. We use a model using our proposed features and the customized corpus in order to find the source language of a historical document. The following chapter describes this in detail.

Features	Accuracy	F-score
Word entropy	62.02%	61.92%
Word frequency entropy	63.36%	63.39%
Word length entropy	53.81%	50.94%
Character entropy	57.78%	56.58%
Character frequency entropy	57.93%	57.28%
SL and WL probability	52.93%	50.26%
SL and DW probability	54.41%	53.86%
Information-theoretic features	78.87%	78.85%
All features	86.63%	86.62%

Table 12.13: Evaluation of information-theoretic features in source and translation identification.

13 Identifying the Origin of a Historical Document

The gospels are historical documents that were first translated almost 2,000 years ago. There are many versions of each gospel, some that have been translated from the original, and some that are translations of translations. Nowadays, it is often unclear what the language of the original documents was. Historians and linguists are unsure how to decide the origin of such historical documents. The *Georgian* gospels (Adysh gospels) are an example of such documents. The *Georgian* gospels were translated from the *Armenian* or *Greek* gospels (Lang 1957). There are about 300 manuscripts of the four Gospels in *Georgian* that are translated from different languages (Kharanauli 2000). Linguists are able to narrow down potential origins by looking at different linguistic properties, but it is still not possible to decide the single origin. We have three such gospels in *Georgian*, *Armenian* and *Greek*, where linguists believe that *Armenian* or *Greek* are the potential origin. In this thesis we use a supervised machine learning technique to find out the correct source of a version of the *Georgian* Gospel.

One of the challenges of dealing with historical data is it requires specific knowledge of a language that is not spoken currently, or, in the case of a language that is in current use, we know that many properties change due to evolution. Finally, it is also more challenging for a machine learning technique when the data set is very small.

Before finding the correct source of the *Georgian* gospel, it is important to find out that the document itself is a translated document. Then we can go one step further to find the correct source of that gospel.

13.1 The Georgian Gospel

The gospels were among the very first documents that were translated into Georgian following the invention of the Georgian alphabet (Lang 1957). The history begins with the palimpsest manuscripts from the *fifth* or *sixth* centuries and ends with the manuscripts

Language	Sentences	Avg. sentence length	Avg. word length
Georgian	3738	18.96	4.71
Armenian	3738	19.15	4.00
Greek	3738	20.40	4.24

Table 13.1: Historical corpus statistics

from the *eighteenth* century. Scholars are still debating the origin of the Georgian translation of the holy scripture. According to Blake Blake (1932), many translations were made from the *Syrian* and *Armenian* gospels.

However, recent studies show two more sources. The first one is the *Palestinian* and other one is the *Antiochian/Constantinopolian* (Kharanauli 2000).

The precise dates of these translations are unknown, but the earliest translations of the Georgian Gospel are presented in the lower script of palimpsests, the so-called *Xanmeti* fragments. The term *Xanmeti* was introduced by the famous Georgian monk, religious writer and translator *George the Athonite*[1]. He denotes the text where the *x-prefix* is employed to mark the second subject and the third person objects of the *Georgian* verb. This prefix did not occur in inscriptions since the seventh century. Based on philological data, these fragments are dated from the *fifth* to the *seventh* centuries. Codicological study of the folio size reveals that they are fragments of quite large codices, and it can be assumed that these codices included several books of the Bible.

Currently, there are about 300 manuscripts of the four gospels in *Georgian* (Kharanauli 2000). Among these, about 40 codices include the oldest text version of the Georgian gospels. The gospel considered for this study is believed to be translated from *Armenian* or *Greek*. These gospels were digitized and aligned manually by linguists. Table 13.1 shows the statistics of the gospels.

13.2 Approach

Section 9.1 in Chapter 9 describes the properties of translation. Based on these properties, a translated text is different than the corresponding source text. Properties proposed by translation scholars focus on texts and the translation process. Our assumption is that even though historical texts were translated many hundreds of years ago, there are some properties that are common to modern texts and the recent translation process.

We model the task as a classification task where we use an SVM implementation

[1] Wikipage: http://en.wikipedia.org/wiki/George_the_Athonite

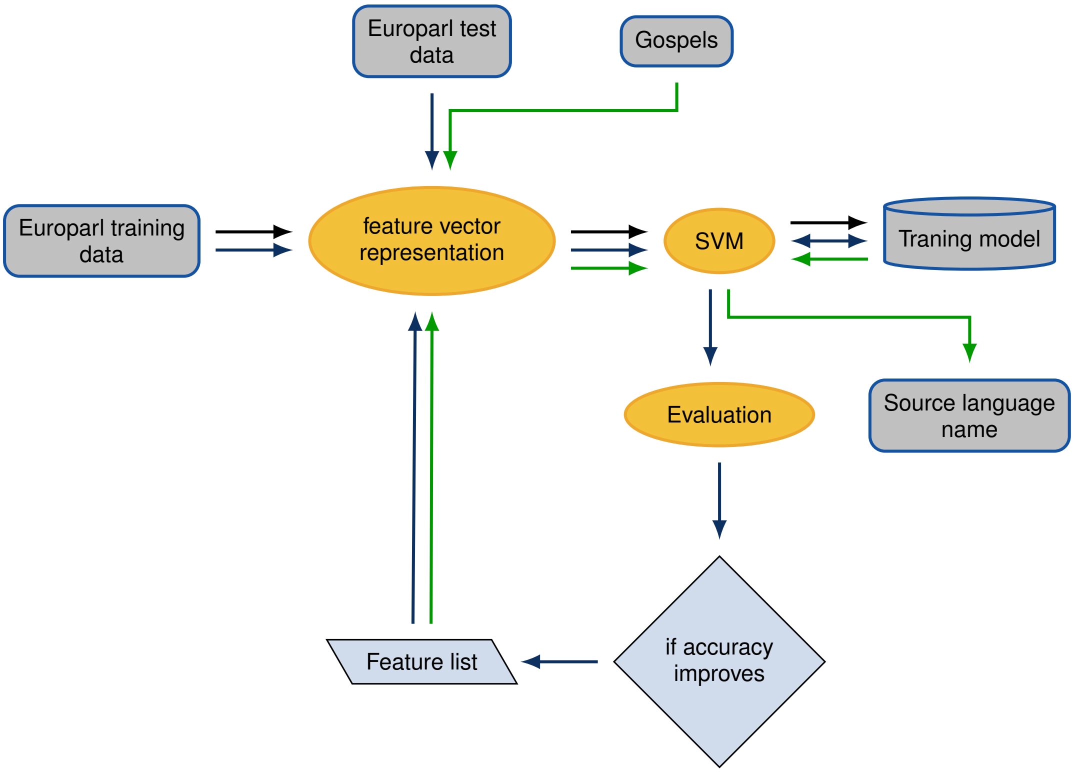


Figure 13.1: Machine learning approach for finding the source of the Georgian gospel.

in Weka to find the correct source of the *Georgian* Gospel. Linguists believe that the *Georgian* gospel is a translated document. They narrowed down potential origins by looking at different linguistic properties compared to the *Greek* and *Armenian* gospel. Before finding the source of the *Georgian* gospel, it is necessary to check that the gospel itself is a translated document or not. If the gospel is classified as a translated document then we can move further to find the source. The gospel that has properties of an original document will be the closest candidate for the origin *Georgian* gospel.

As features we use our proposed features that are listed in Section 6.5 of Chapter 6. We consider the whole *Europarl* corpus (See Table 11.5) in order to build the training model for the SVM. Figure 13.1 shows the visual representation of our approach.

13.3 Experiment with Gospels

In order to experiment with the target corpus, we also followed a similar strategy to the training corpus. The target corpus is divided similarly into chunks. Each gospel was divided into 37 chunks. Each chunk contains 100 verses. Then, this data was divided

	Source	Translation
Armenian	0	37
Georgian	1	36

Table 13.2: Confusion matrix of the *Armenian–Georgian* gospels.

	Source	Translation
Greek	20	17
Georgian	1	36

Table 13.3: Confusion matrix of *Greek–Georgian* Gospels.

into two sets. The first set contains chunks from *Armenian* and *Georgian*. The other contains chunks from *Greek* and *Georgian*.

As we stated earlier, the first task was to classify the *Georgian* gospel into the binary classes of *translated* or *original*. Table 13.2 shows the confusion matrix of the first set. The confusion matrix shows that 36 out of 37 chunks of the *Georgian* gospel were identified as translated text. We get the same result for the second set also (See Table 13.3).

So, both of the confusion matrices show that the *Georgian* gospel is a translated document. Now we can move to find the source of the *Georgian* gospel. The source of the gospel has to be identified as an original document. Table 13.2 shows that all of the chunks from the *Armenian* gospel were identified as a translation. That is why we can rule out the *Armenian* gospel as the origin of the *Georgian* gospel. Now the *Greek* gospel remained as the candidate for the original. However, the chunks of the gospel had to be identified as translated documents. Table 13.3 shows the confusion matrix of the second set where 20 out of 37 chunks of the *Greek* gospel were identified as source. Therefore, these two confusion matrices show that *Greek* is more likely to be the origin of the *Georgian* gospel. It becomes more clear when we have a look on Table 13.4. Here *Armenian* and *Greek* chunks are labeled as *source* and *Georgian* chunks are labeled as translation. The *accuracy* and *F-Score* of the *Armenian–Georgian* pair is below 50%. But the *accuracy* and *F-Score* of the *Greek–Georgian* pair is above 75%. So, our experimental results suggest that the *Greek* Gospel is more likely to be the origin of the *Georgian* Gospel.

Source-translation	Accuracy	F-Score
Armenian–Georgian	48.64%	32.73%
Greek–Georgian	75.67%	74.48%

Table 13.4: Classification results of Gospels

14 Discussion of Translation Analysis

A translated text contains some latent traces that make the text very different than the corresponding source text. These traces can be used to separate a source text and a translated text. By using some features that are representative of these traces, it is possible to train a machine that can identify source texts and translated texts.

The task of identifying source and translated texts is itself a multilingual task where input texts could be from any language. But this study has a limitation. That is, input texts have to be word separated. The word segmentation process has to be performed before classification of texts from languages where words are not segmented, such as *Chinese*, *Japanese* and *Thai*.

In recent years, translation scholars have proposed different translation properties using monolingual or comparable corpora. It is necessary to validate the claims by translation scholars. But, the field lacks a multilingual corpus. In this thesis, we compiled such a corpus that can be used by translation scholars.

There are properties of translated texts that make them distinguishable from the original texts. Also, translated texts from different source languages to the same language have identical properties that make them separable according their origin (Koppel and Ordan 2011). However, we observe a different scenario when we consider source and translated texts from different languages. Translated texts could be influenced by the language family of the source texts and regional or demographic influence among languages. Source and translated texts from *Slavic* languages have different properties than texts from *Germanic* and *Romance* languages. However, we found similar findings to Koppel and Ordan (2011), i.e., that a translated text is very different than a source text.

In this thesis, we analyzed some *information-theoretic* and *lexical features* for source and translation classification that are also shown useful for text readability classification. A classifier using these features is able to separate translated texts into the same language from different source languages. However, we observe some language family and regional influences among texts. Table 12.9 shows that translated texts from *Hungarian* and *Czech* to different languages are difficult for classification according to their source

	Czech	German	Greek	Spanish	Finish	French	Hungarian	Lithuanian	Dutch
Czech	103	0	0	40	0	0	57	0	0
German	0	195	0	0	0	2	0	0	0
Greek	0	0	140	0	60	0	0	0	0
Spanish	29	0	0	132	0	0	39	0	0
Finish	0	0	43	1	156	0	0	0	0
French	0	0	0	0	0	199	0	0	0
Hungarian	79	0	0	57	1	0	63	0	0
Lithuanian	0	0	0	0	0	0	0	200	0
Dutch	0	0	0	0	0	0	0	0	200

Table 14.1: Confusion matrix of *Polish* translation texts from different language sources.

languages. In order to show the classification errors, we generated a confusion matrix of *Polish* translated texts from different languages using a 10-fold cross validation, shown in Table 14.1. It shows that the classifier makes mistakes for translated texts in *Czech*, *Hungarian* and *Spanish*. Many of the texts from *Hungarian* are classified as *Czech*. A potential reason of this miss-classification could be the *loanwords* in *Hungarian* from *Slavic* languages. István (1989)[1] shows that about 20 word roots in the *Hungarian* language come from *Slavic* languages.

Ilisei, Inkpen, Pastor, and Mitkov (2009) and Ilisei, Inkpen, Pastor, and Mitkov (2010) showed that different lexical features that are quantitative indicators of translation properties are able to classify source texts and translated texts. Average SL and average WL are among such features. A higher value of average SL in translated texts shows *explicitation* translation properties. A lower value of average SL in translated texts might represent *simplification* properties, where translators divide a single complex sentence into several simple sentences in order to make the translation simple. However, this is not the case here because the customized *Europarl* corpus is parallel and the sentence numbers on both sides are equal. Figure 14.1 shows average SL in source and translated texts of ten European languages. For most of the languages in Figure 14.1, the average SL of the translation has a higher value than the source. Translators tried to make the translated texts more explicit than the source texts. However, the behavior is not uniform. The value of the feature is lower for some languages. The average WL shows similar behavior.

We observe the same for *word entropy*. For most of the languages *word entropy* has a higher value. In the preceding chapter, we have shown that a good readable text has lower entropy than a less readable text. Figure 14.2 shows *word entropy* in source and translated texts. Translated texts of most of the languages have higher *word entropy* than corresponding source texts. That means that source texts are more readable when we consider the *word entropy* feature. However, the behavior of this feature is not

[1] The reference is taken from: `http://en.wikipedia.org/wiki/Hungarian\_language`

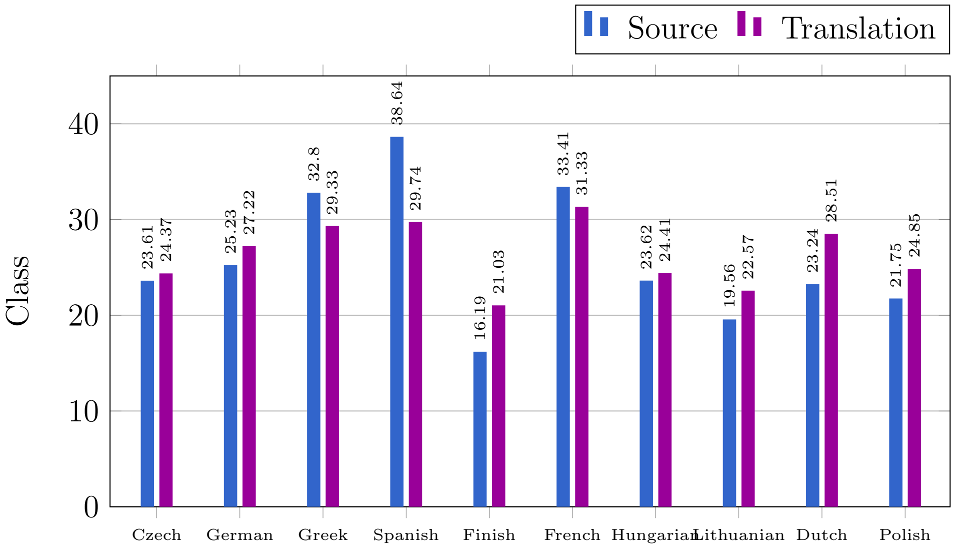


Figure 14.1: Average sentence length in source and translation

uniform either.

Figure 14.3 shows *character entropy* in source and translated texts. The behavior of this feature is uniform. Translated texts have less or equal *character* entropy than source texts. By considering the behavior of this feature, translated texts are more readable than original texts.

A classifier using our proposed features achieves classification accuracy of 86.63%, when we use modern texts. However, the history of translation started a long time ago. There are many documents that were translated many hundreds of years ago. We have trained a model on modern texts from the *Europarl* corpus and used that model to classify three historical documents. Our assumption was that there are some similarities between the translation task performed hundreds of years ago and the translation task performed more recently. The classifier gives us the potential origin of the *Georgian* Gospel, that is the *Greek* Gospel. However, it would be more constructive if we could measure the accuracy of the classifier using a *gold standard* for historical documents.

It is hard to collect *gold standard* historical translated and original documents where the original and translated documents are annoted. However, we have collected five original books and their translations from the *project Gutenberg*[2] in the literature domain as negative examples. These books are written by *Johann Wolfgang von Goethe*, *Gotthold Ephraim Lessing* and *Friedrich Nietzsche*. We are calling them negative examples because these authors use great freedom when writing books. Also, styles vary

[2] http://www.gutenberg.org/

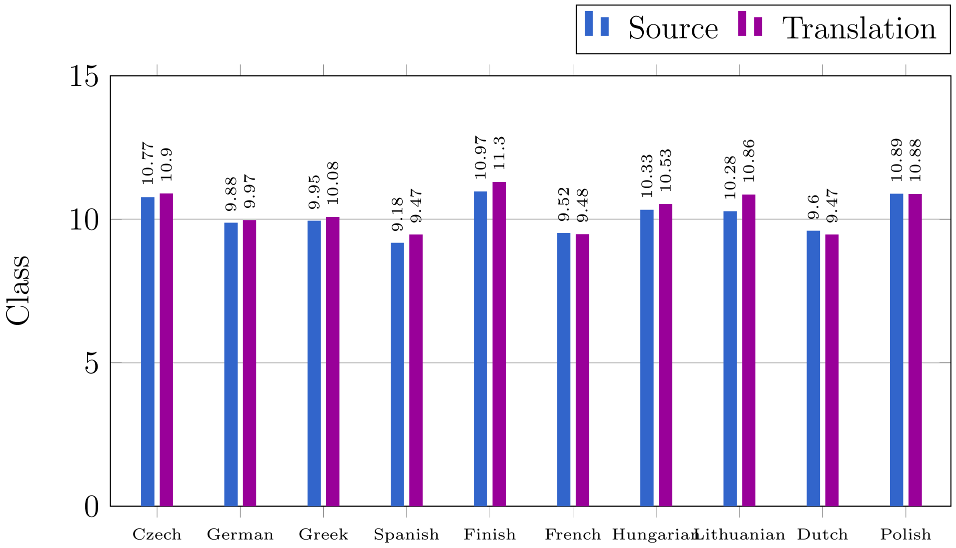


Figure 14.2: Word entropy in source and translation

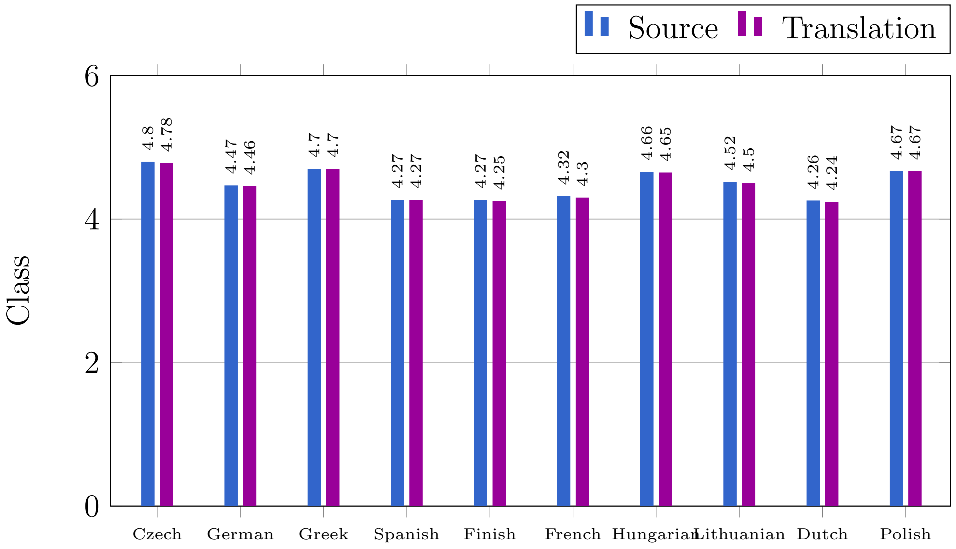


Figure 14.3: Character entropy in source and translation

	Source	Translation
Source	0	138
Translation	32	233

Table 14.2: Confusion matrix of books from the *project Gutenberg*

significantly among writers. It is very challenging for a classifier to classify these texts. The test corpus contains in total 403 chunks where 138 chunks are from original books. Our hypothesis is that the classifier will not perform well for this test set due to greater freedom authors enjoy during their writing. Table 14.2 shows the confusion matrix of the books from the *project Gutenberg*. The classifier does not work as we expected. All of the source documents are identified as translation.

We also explored the *attribute selection* procedure for both *source identification* and *source and translation* classification. The best classification accuracy of 87.74% was achieved for *source and translation* classification using *ChiSquaredAttributeEval* (Bouckaert, Frank, Hall, Kirkby, Reutemann, Seewald, and Scuse 2013) as the attribute selection algorithm and *Ranker* (Bouckaert, Frank, Hall, Kirkby, Reutemann, Seewald, and Scuse 2013) as the search algorithm. All of the proposed features are selected for this setting. For *source identification* of source and translated texts from ten European languages, the *ConsistencySubsetEval* (Liu and Setiono 1996) performs best as the attribute selection algorithm with *Genetic Search* (Goldberg et al. 1989) as the search algorithm. This setting uses 14 among 18 proposed features.

The experimental results and observations shown above demonstrate that our proposed features are not only useful for text readability analysis, but also for source and translation analysis. Many of these proposed features are indicators of different translation properties described in Section 9.1 of Chapter 9. *Lexical* features are found more useful than *information-theoretic* features for this task. However, *information-theoretic* features have more predictive power to measure readability. In the translation analysis, a translated text is more readable than an original text (Pastor, Mitkov, Afzal, and Pekar 2008). That is why this feature set is also useful for this task.

15 Conclusions

Multilingualism is one of the latest trends in NLP application development. In this thesis we have considered two multilingual tasks: *readability classification* of texts from three different languages and *source and translation* classification of texts from 21 languages. Both of the tasks lacked appropriate resources that could be used by the research community. So, the first challenge was collect appropriate and suitable corpora for both of the tasks. In this thesis we provide four such corpora to the research community, three of them for *text readability* classification and one for corpus based translation studies. The first readability corpus was collected from the English Wikipedia by considering the result of the *article feedback tool*. The corpus contains 641 documents under four difficulty classes. The *well written* attribute in the *article feedback tool* is a measure of the readability of an article. However, our experimental results showed that this attribute is influenced by other attributes of the *article feedback tool* such as *complete* and *trustworthy*. A very recent study on Wikipedia articles (Flekova, Ferschke, and Gurevych 2014) based on readability found similar findings. So, it was necessary to collect another English corpus that is more appropriate for readability.

We compiled another English corpus from textbooks that have been used in schools in Bangledesh for students who want to study in English. The respective government agency controls all aspects of these textbooks including readability. That is why these textbooks are ideal resources for compiling a corpus for text readability analysis. We also compile a readability corpus for Bangla from textbooks from the same source. This corpus contains 582 documents under four difficulty classes.

The last corpus we provide is a customized version of the *Europarl* corpus. The *Europarl* corpus contains many erroneous annotations. These erroneous annotations could lead to erroneous results. We resolved these problematic annotations and extracted a corpus where source sentences and translated sentences are annotated with the correct language. The corpus contains 2,646,765 parallel sentences from 412 language pairs of 21 European languages. All of these corpora are freely available at: `http://www.hucompute.org/ressourcen/corpora`.

For the text readability analysis many traditional formulas have been proposed that are still being used in different commercial tools. Many recent studies have been proposed, using different features that are linguistically motivated. These features have been shown to be useful for text readability classification (Schwarm and Ostendorf 2005; Feng, Elhadad, and Huenerfauth 2009; Heilman, Collins-Thompson, and Eskenazi 2007; Aluisio, Specia, Gasperin, and Scarton 2010; Feng, Janche, Huenerfauth, and Elhadad 2010). Barzilay and Lapata (2008), Brück and Hartrumpf (2007), Brück, Hartrumpf, and Helbig (2008a), and Brück, Hartrumpf, and Helbig (2008b) proposed different features based on semantics. They have shown that semantic features also have good predictive quality. We also proposed several semantic features that are useful for readability classification. The evaluation of all linguistic features suggests that features should be calculated per document instead of per sentence for a document-level readability classification. However, all of these features are dependent on different linguistic tools. There are many languages in this world that are considered *low-resource* languages due to a lack of written resources available even though these languages are widely spoken, often in fact by more people than the well-resourced languages.

Features that require linguistic pre-processing can not be used for these *low-resourced* languages. In this thesis we proposed 18 *information-theoretic* and *lexical* features that are language independent and do not require any kind of linguistic pre-processing. We have shown that a classifier using these features is able to achieve a higher classification accuracy compared to a classifier using 40 linguistic features. These features are not only useful for readability classification of an *English* corpora, but also for texts from the German and Bangla languages. There are very few related works available on Bangla readability analysis. Das and Roychoudhury (2004) and Das and Roychoudhury (2006) found that traditional readability formulas are useful for Bangla readability classification. Recently Sinha, Sharma, Dasgupta, and Basu (2012) proposed two formulas that are like English traditional readability formulas. These formulas considered two important factors of Bangla texts: *polysyllabic words* and *consonant-conjuncts*. However, the classifier we present in this thesis is the best performing readability classifier for Bangla texts. A readability classifier using our proposed features also achieves a classification accuracy that is comparable with Hancke, Vajjala, and Meurers (2012). They have achieved 89.7% classification using 155 lexical and linguistically motivated features. However, a classifier using our proposed 18 features achieves 85.82% correct classification.

We performed a study to find the best performing features for readability classification of these three corpora. Only two features are selected for all four corpora and both of them belong to the *information-theoretic* feature set. Seven of our 18 proposed features are selected for the three corpora and four of them are *information-theoretic* (See Table 7.1 in Chapter 7). A text is easier when the next word occurrences are easier to predict. This kind predictability can be measured by the *information-theoretic* feature we proposed in this thesis. This evaluation shows that our proposed *information-theoretic* features should be considered for readability classification of texts from any language.

Corpus-based translation studies is a relatively new field. Nowadays, translation scholars are interested in investigating the properties of texts as translation as well as original. Using monolingual or comparable corpora scholars proposed several properties of translated texts. However, the field lacks corpora where scholars can validate these translation properties. This thesis provides a corpus that contains texts from 21 European languages. This corpus could be an ideal resource for scholars who wish to validate their theoretical findings.

The translation process leaves some latent traces in the translated texts. These traces could be used to separate translated texts from original texts. Along with the corpus the field also lacked such features that are able to separate original and translated texts. Recently, Ilisei, Inkpen, Pastor, and Mitkov (2009) and Ilisei, Inkpen, Pastor, and Mitkov (2010) proposed some features that separate original and translated texts in Spanish. However, many of their features are linguistically motivated and require linguistic preprocessing. Our proposed features, which proved useful for text readability classification, can also separate translated texts from original texts. Some of our proposed features are good indicators of different translation properties proposed recently by translation scholars. A classifier using these features achieves a classification accuracy of 86.63% in source and translation classification. This feature set also achieves satisfactory classification accuracy for *source language identification* from translated texts. However, the classifier struggles with translated texts from *Czech* and *Hungarian*. Our experimental results suggest that translated texts are influenced by language family and loan words from neighboring languages.

In summary, this thesis considered multilingual text readability classification for the first time. We proposed *information-theoretic* features that are useful for *text readability* classification and *source and translation* classification. Classifiers using these features achieve reasonable classification accuracy for all of the tasks mentioned in this thesis. Some of the proposed features are cognitively motivated. Along with these features, we also provide three corpora for readability analysis and one corpus for translation studies.

16 Thesis Summary in German

Eine Sprache kann als Medium menschlicher Kommunikation bezeichnet werden, welches Informationen codiert und überträgt. Alle natürlichen Sprachen folgen den Gesetzen der Evolution, (Nowak, Komarova, and Niyogi 2002; Lieberman, Michel, Jackson, Tang, and Nowak 2007). Obwohl das Codieren von Information hochkomplex ist, kann es durch Bilden von Wortsequenzen erreicht werden. Diese Sequenzen bilden Sätze und Abfolgen von Sätzen wiederum bilden Texte. Daher ist ein Text ein Medium der Kommunikation zwischen Autoren und ihrer Leserschaft. Parallel dazu ist auch ein übersetzter Text ein Kommunikationsmedium, wobei hier gleichzeitig der Übersetzer involviert ist. Ein Text kann monolingual oder multilingual sein.

Die Anzahl multilingualer Texte im World Wide Web (WWW) wächst dramatisch an und ein mehrsprachiger Wirtschaftsraum wie die Europäische Union (EU) ist auf die Verfügbarkeit multilingualer Werkzeuge aus dem Feld des Natural Language Processing (NLP) angewiesen. Dank einer rapiden Entwicklung solcher NLP-Werkzeuge sind heutzutage eine Vielzahl an lexikalischen, syntaktischen, semantischen und anderen linguistischen Merkmalen in unterschiedlichen NLP-Applikationen in Gebrauch. Allerdings gibt es einige Umstände unter denen diese Merkmale aufgrund des Typs der Applikation oder der mangelnden Verfügbarkeit von NLP-Ressourcen für manche der Sprachen nicht genutzt werden können. Aufgrund dessen sollte eine Applikation, die für mehrsprachige Texte konzipiert wurde, Merkmale verarbeiten, die nicht von einer Einzelsprache abhängig sind und auch keine spezifisch einzelsprachigen Werkzeuge verwendet. In der vorliegenden Arbeit werden zwei solcher Applikationen fokussiert: *text readability classification* oder Lesbarkeitsklassifikation und *source and translation classification* oder Quell- und Zielsprachenklassifikation. Für beide Aufgaben fehlen geeignete wissenschaftliche Ressourcen. Daher war es die erste Herausforderung, passende und passgenaue Korpora für beide Aufgaben zu erstellen. In dieser Arbeit werden der Wissenschaftsgemeinschaft vier solcher Korpora zugänglich gemacht, drei davon für die Lesbarkeitsklassifikation und eines für die korpusbasierte Übersetzungsforschung. All diese Korpora sind unter `http://www.hucompute.org/ressourcen/corpora` frei verfügbar.

Die Entwicklung beider Applikationen beinhaltet vier Schritte. Diese sind: *Korpuszusammenstellung*, *Merkmalsextraktion*, *Implementierung* und *Evaluation*. Diese Aufgaben werden als überwachte Kategorisierungsaufgabe behandelt. D.h. jedem Eingabedokument wird das Label einer der vortrainierten Klassen zugewiesen. Die Testdaten sind von den Trainingsdaten getrennt. Ziel war es, diejenigen Merkmale, die sprachunabhängig sind und keine linguistische Vorverarbeitung benötigen zu sondieren. Es wurde dabei mit einer Reihe an Merkmalen experimentiert. Schließlich wurden 18 Merkmale, die sich in beiden Applikationen als nutzbringend herausstellten, ausgewählt. Sieben davon fallen in die Klasse der *informationstheoretischen* Merkmale, während die restlichen *lexikalisch* sind.

Es wurde das für seine hohe Qualität bei Multiklassen-Klassifikation bekannte WEKA Toolkit, (Hall, Frank, Holmes, Pfahringer, Reutemann, and Witten 2009), verwendet. Das Toolkit umfasst eine Vielzahl an Algorithmen des maschinellen Lernens. Sequential Minimum Optimization (SMO), (Platt 1998; Keerthi, Shevade, Bhattacharyya, and Murthy 2001), ist ein Algorithmus, der zum Trainieren von SVM mit dem *Pearson VII function-based universal kernel* PUK, (Üstün, Melssen, and Buydens 2006), verwendet wird. Beim Vergleich verschiedener Klassifikationsalgorithmen in den Experimenten war SMO der performanteste. Verschiedene Modelle wurden anhand von Merkmalen trainiert und mittels *Akkuratheit* und *F-Wert* evaluiert. Bei jedem der Experimente wurde das Korpus in ein *Trainings-* und ein *Testkorpus* aufgespalten. Das *Trainingskorpus* enthält dabei 80% des Originalkorpus und der Rest des Korpus wird als *Testkorpus* verwendet. Jedes Experiment wird 100 Mal wiederholt, wobei *Trainings-* und *Testkorpus* per Zufallsgenerator zusammengestellt werden. Schließlich wird der gewichtete Durchschnitt von *Akkuratheit* und *F-Wert* aus den Ergebnissen dieser Experimente berechnet.

Bei Lesbarkeit geht es um den Schwierigkeitsgrad des Textes beim und für das Lesen. Ein Text der für einen Leser leicht zu lesen ist, kann für einen anderen schwierig sein. Das Problem besteht daher darin, für die verschiedenen Lesefähigkeitsstufen der Leser passende Texte zu finden. Genauer gesagt, Lesbarkeit bedeutet, die Schwierigkeitsstufen des Materials zu variieren und an den jeweiligen Leser je nach dessen Fähigkeiten anzupassen.

In der Lesbarkeitsanalyse wurden viele traditionelle Formeln publiziert, die noch immer in verschiedensten kommerziellen Tools in Gebrauch sind (DuBay 2004). Viele neuere Studien wurden veröffentlicht, bei denen Merkmale gebraucht wurden, die linguistisch motiviert waren. Es konnte für diese Merkmale gezeigt werden, dass sie für die Lesbarkeitsklassifikation brauchbar sind, (Schwarm and Ostendorf 2005; Feng, Elhadad, and Huenerfauth 2009; Heilman, Collins-Thompson, and Eskenazi 2007; Aluisio, Specia, Gasperin, and Scarton 2010; Feng, Janche, Huenerfauth, and Elhadad 2010). Barzilay and Lapata (2008) and Brück, Hartrumpf, and Helbig (2008a) konnten zeigen, dass semantische Merkmale ebenfalls eine gute Grundlage für Vorhersagbarkeit bilden. In der vorliegenden Arbeit werden ebenfalls semantische Merkmale vorgestellt, die für die Lesbarkeitsforschung brauchbar sind. Die Evaluation aller linguistischer Merkmale deutet darauf hin, dass diese per Dokument statt pro Satz berechnet werden sollten, wenn es sich um eine Lesbarkeitsklassifikation auf Dokumentenebene handelt. Allerdings sind diese Merkmale von verschiedenen linguistischen Werkzeugen abhängig. Viele Sprachen in der Welt werden als ressourcenarm angesehen, entweder weil die Bevölkerung, die diese Sprachen, spricht nicht sehr groß ist oder weil nicht genügend digitalisiertes Textmaterial und linguistische Werkzeuge verfügbar sind, (Islam, Tiedemann, and Eisele 2010). Den meisten der ressourcenarmen Sprachen fehlt ein Werkzeug zur Lesbarkeitsklassifikation. Eine universelle Lesbarkeitsklassifikation für solche Sprachen benötigt Merkmale, die sprachunabhängig sind und keinerlei linguistische Vorverarbeitung erfordern.

Beide Aufgaben, auf die sich die vorliegende Arbeit konzentriert, sind mit einem kognitiven Modell verbunden. Es ist wichtig, entsprechende kognitive Theorien zu berücksichtigen. Die hier vorgestellte Forschung findet im Rahmen theoretischer Überlegungen statt, die im Felde der Kognitionswissenschaft begründet liegen; ein Verständnis für kognitive Prozesse des Lesens und der Übersetzung eines Textes von einer in eine andere Sprache spielen dabei die entscheidende Rolle. Nach Kintsch (1998) besteht der Kernpunkt des kognitiven Modells des Lesens im sequentiellen Aufbau von Bedeutung Wort für Wort, Satz für Satz und Absatz für Absatz. Dieses ist ein inkrementeller Prozess. Gleichzeitig ist die effektivste Methode, Information gegen ein Grundrauschen zu übertragen, eine konstante Übertragungsrate, (Genzel and Charniak 2002; Genzel and Charniak 2003). Diese Grundregel muss in jeder Art Kommunikation befolgt werden, um diese effizient zu gestalten. D.h. in einem beliebigen Text wächst die Entropie eines Satzes mit seiner numerischen Position im Text, sofern sie ohne den Kontext zu berücksichtigen gemessen wird.

Die Lesbarkeit eines Textes hängt ebenso von der Vorhersagbarkeit des Textes ab. Ein Text mit einer geringeren Entropie ist vorhersagbarer und daher leichter zu lesen. Das Lesen eines Textes ist zudem von weiteren Textaspekten abhängig. Diese Aspekte beginnen mit lexikalischen Variablen wie der Wortfrequenz in einer Sprache, der *Average Word Length* (durchschnittl. Wortlänge) usw. Folglich müssen verschiedene lexikalische Merkmale bei der Messung des lesebezogenen Schwierigkeitsgrades beachtet werden.

Es wird gezeigt, dass ein Klassifizierer, der die hier vorgeschlagenen Merkmale benutzt, eine höhere Klassifikationsakkuratheit erreicht als ein Klassifizierer, der 40 linguistische Merkmale kombiniert. Diese Merkmale sind nicht nur für die Lesbarkeitsklassifikation *englischer* Korpora nützlich, sondern auch für deutsche und bengalische Texte. Der Lesbarkeitsklassifizierer erreicht einen *F-Wert* von 74.21% für das englische Wikipediakorpus, einen *F-Wert* von 75.47% für das englische Textbuchkorpus, einen *F-Wert* von 86.46% für das Bangla (Bengalisch) Textbuchkorpus und einen *F-Wert* von 86.26% für das deutsche *GEO/GEOLino*-Korpus.

Übersetzung ist eine sprach- und kulturübergreifende Kommunikationsaktivität, bei der die Mediation zwischen Sprachen eine wichtige Rolle spielt. Dabei handelt es sich um die Aufgabe der geeigneten Reformulierung von Kommunikationsstrukturen der Quell- in der Zielsprache. Teilweise bezieht sich dies auf eine Operation, die eine Einheit der Quellsprache in eine der Zielsprache überführt ohne die Modalität des Ursprungs- und Zieltextes zu beachten. Generell sind zwei Parteien involviert, wenn es um Text als Kommunikationsmedium geht: der *Schreiber* und der *Leser*. Beim Prozess des Übersetzens ist jedoch noch eine dritte Partei involviert: der *Übersetzer*. Diese drei Parteien unterscheiden sich in ihren Sprachen, kognitiven Fähigkeiten und Kulturen. Ihre physikalischen Umgebungen können teilweise überlappen, was ihnen hilft inferentiell ähnliche Fähigkeiten zu entwickeln, (Chang 2009). Der Übersetzungsprozess hinterlässt einige latente Spuren im übersetzten Text. Diese Spuren können genutzt werden, um übersetzte Texte von Originaltexten zu unterscheiden.

Korpusbasierte Übersetzung ist ein relativ junges Feld. Heutzutage sind Übersetzungsforscher daran interessiert, die Charakteristika von übersetzten sowie originalen Texten zu verstehen. Anhand von monolingualen oder vergleichbaren Korpora haben Forscher einige Charakteristika übersetzten Textes erarbeitet. Sie schlugen einige Charakteristika von Übersetzungen vor indem sie monolinguale oder bilinguale Korpora untersuchten, (Baker 1996; Olohan 2001; Laviosa 2002; Hansen 2003). Diese werden oft Übersetzungsuniversalien (*translation universals*) genannt. Diese Merkmale werden unter fünf Schlüsselworten zusammengefasst: *explicitation* (Explizierung), *simplification* (Vereinfachung), *normalization* (Normalisierung), *levelling out* (Ausklingen) und *interference* (Interferenz). Vor Kurzem schlugen (Ilisei, Inkpen, Pastor, and Mitkov 2009; Ilisei, Inkpen, Pastor, and Mitkov 2010) Merkmale vor, die Original und Übersetzung im Spanischen charakterisieren. Allerdings waren viele ihrer Merkmale linguistisch motiviert und benötigen linguistische Vorverarbeitung.

Der Forschungsgemeinschaft fehlen bisher multilinguale Korpora an denen Wissenschaftler Übersetzungscharakteristika validieren können. Neben einem solchen Korpus fehlt der Forschungsgemeinschaft des Weiteren das Wissen um Merkmale, die generell Originaltexte von übersetzten Texten unterscheiden helfen. Die vorliegende Arbeit bietet ein multilinguales Korpus, welches aus dem Europarl corpus, (Koehn 2005), heraus kompiliert wurde. Das Korpus ist vielschichtig, da es Texte aus 21 Sprachen und 7 Sprach (unter) familien enthält. Das Korpus enthält 2646765 parallele Sätze aus 412 Sprachpaaren in 21 europäischen Sprachen.

Die hier vorgeschlagenen Merkmale, die sich in der Lesbarkeitsklassifikation als brauchbar erwiesen haben, können auch übersetzte von originalen Texten unterscheiden. Einige der vorgestellten Merkmale sind gute Indikatoren für verschiedene der von Übersetzungsforschern vorgeschlagenen Charakteristika. Der Klassifizierer erreicht mit denselben 18 Merkmalen eine Klassifikationsakkuratheit von 86.63%. Diese Merkmale wurden zudem beim Bau eines Ursprungssprachenklassifikators verwendet. Für 10 europäische Sprachen erreichte er eine Akkuratheit von 75%.

Zusammenfassend lässt sich sagen, dass die vorliegende Arbeit die erste ist, in der multilinguale Lesbarkeitsklassifikation erforscht wird. Es werden *informationstheoretische* Merkmale vorgestellt, die sowohl für eine Lesbarkeitsklassifikation, als auch für die Klassifikation von *Quell- und Zielsprache* nutzbar sind. Klassifizierer, die diese Merkmale nutzen, erzielen akzeptable Akkuratheiten für alle Aufgaben, die in der vorliegenden Arbeit besprochen werden. Manche der vorgeschlagenen Merkmale sind kognitiv motiviert. Zusammen mit diesen Merkmalen werden drei Korpora für die Lesbarkeitsanalyse und ein Korpus für die Übersetzungsforschung erarbeitet und verfügbar gemacht.

Bibliography

Abedi, Jamal, Robert Bayley, Nancy Ewers, Kimberly Mundhenk, Seth Leon, Jenny Kao, and Joan Herman (2012). "Accessible reading assessments for students with disabilities". In: *International Journal of Disability, Development and Education* 59.1, pp. 81–95.

Abedi, Jamal, Jenny C Kao, Seth Leon, Ann M Mastergeorge, Lisa Sullivan, Joan Herman, and Rita Pope (2010). "Accessibility of segmented reading comprehension passages for students with disabilities". In: *Applied Measurement in Education* 23.2, pp. 168–186.

Aha, David W and Richard L Bankert (1996). "A comparative evaluation of sequential feature selection algorithms". In: *Learning from Data*. Springer, pp. 199–206.

Alan, Keith (2001). *Natural Language Semantics*. Blackwell Publishers Ltd, Oxford.

Aluisio, Ra, 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*.

Alvestrand, Harald Tveit (1995). "Tags for the Identification of Languages". In: *IETF, RFC 1766, Mar.* URL: `http://www.ietf.org/rfc/rfc1766.txt`.

Andrews, Sally (1997). "The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts". In: *Psychonomic Bulletin & Review* 4.4, pp. 439–461.

Angelone, Erik (2010). "Uncertainty, uncertainty management and metacognitive problem solving in the translation task". In: *Translation and Cognition* 15, pp. 17–40.

Arends-Kuenning, Mary and Sajeda Amin (2004). "School incentive programs and children's activities: The case of Bangladesh". In: *Comparative Education Review* 48.3, pp. 295–317.

Argamon, Shlomo and Shlomo Levitan (2005). "Measuring the Usefulness of Function Words for Authorship Attribution". In: *The Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing*.

Avner, Ehud Alexander, Noam Ordan, and Shuly Wintner (2014). "Identifying translationese at the word and sub-word level". In: *Literary and Linguistic Computing*, fqu047.

Baker, Mona (1993). "Corpus Linguistics and Translation Studies - Implications and Applications". In: *Text and Technology. In Honour of John Sinclair.* Ed. by Mona Baker, Gill Francis, and Elena Tognini-Bonelli. John Benjamins, pp. 233–354.

— (1995). "Corpora in Translation Studies. An Overview and Suggestion for Future Research". In: *Target* 7(2), pp. 223–243.

— (1996). "Corpus-based Translation Studies: The Challenges that Lie Ahead". In: *LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager.* Amsterdam & Philadelphia: John Benjamins, pp. 175–186.

— (2004). "A corpus-based view of similarity and difference in translation". In: *International Journal of corpus Linguistics* 9(2), pp. 167–193.

Balasubrahmanyan, VK and S Naranan (2005). "Entropy, information and complexity". In: *An International Handbook of Quantitative Linguistics*, pp. 878–891.

Baroni, Marco and Silvia Bernardini (2006). "A New Approach to the Study of Translationese: MachineLearning the Difference between Original and Translated Text". In: *Literary and Linguistic Computing* 21(3), pp. 259–274.

Barry, Christopher, Catriona M Morrison, and Andrew W Ellis (1997). "Naming the Snodgrass and Vanderwart pictures: Effects of age of acquisition, frequency, and name agreement". In: *The Quarterly Journal of Experimental Psychology: Section A* 50.3, pp. 560–585.

Barzilay, Regina and Mirella Lapata (2008). "Modeling Local Coherence: An Entity-based Approach". In: *Computational Linguistics* 21(3), pp. 285–301.

Bates, Elizabeth (1999). "On the nature and nurture of language". In: *Frontiere della Biologia The Brain of Homo sapiens, Rome: Giovanni Trecani.*

Bell, Alan, Jason M. Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky (2008). "Predictability effects on durations of content and function words in conversational English". In: *Elsevier Journal of Memory and Language* 60, pp. 92–111.

Bernardini, Silvia and Fedrico Zanettin (2004). "When is a Universal not a Universal? Some limits of corpus-based methodologies for the investigation". In: *Translation Universals. Do They Exist*, pp. 51–64.

Biere, Bernd Ulrich (2000). "Der Einfluss der Textlinguistik auf die praktische Verständlichkeitsforschung". In: ed. by K. Brinker, W. Antos G.and Heinemann, and S.F. Sager. deGruyter, pp. 859–870.

Blake, Robert Pierpont (1932). *Khanmeti palimpsest fragments of the Old Georgian version of Jeremiah.* Cambridge Univ Press.

Blum-Kulka, Shoshana (1986). "Shifts of cohesion and coherence in translation". In: *Interlingual and intercultural communication: Discourse and cognition in translation and second language acquisition studies*, pp. 17–35.

Blum, Shoshana and Eddie A Levenston (1978). "Universals of lexical simplification". In: *Language learning* 28.2, pp. 399–415.

Borst, Alexander and Frédéric E. Theunissen (1999). "Information theory and neural coding". In: *Nature Neuroscience* 2, pp. 947–957.

Bouckaert, Remco R, Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann, Alex Seewald, and David Scuse (2013). "WEKA Manual for Version 3-7-10". In:

Brainerd, Barron (1976). "On the Markov nature of text". In: *Linguistics* 14.176, pp. 5–30.

Brants, Thorsten (2000). "TnT: a statistical part-of-speech tagger". In: *Proceedings of the sixth conference on Applied natural language processing.* Association for Computational Linguistics, pp. 224–231.

Bruce, Bertram C, Andee Rubin, and Kathleen S Starr (1981). "Why readability formulas fail". In: *IEEE Transactions on Professional Communication.*

Brück, Tim Vor der and Sven Hartrumpf (2007). "A Semantically Oriented Readability Checker for German". In: *3rd Language and Technology Conference, Poznań, Poland.*

Brück, Tim Vor der, Sven Hartrumpf, and Hermann Helbig (2008a). "A Readability Checker with Supervised Learning using Deep Indicators". In: *Informatica* 32.4, pp. 429–435.

— (2008b). "A Readability Checker with Supervised Learning using Deep Syntactic and Semantic Indicators". In: *International Multiconference: Information Society - IS 2008 - Language Technologies, Ljubljana, Slovenia.*

Brück, Tim Vor der, Alexander Mehler, and Md. Zahurul Islam (2014). "ColLex.en: Automatically Generating and Evaluating a Full-form Lexicon for English". In: *Proceedings of LREC 2014.* Reykjavik, Iceland.

Carroll, John Bissell (1964). *Language and thought.* Prentice-Hall Englewood Cliffs, NJ.

Carson, Lorna (2003). "Multilingualism in Europe". In: *A Case Study Organized by the Madariaga European Foundation.*

Cartoni, Bruno and Thomas Meyer (2012). "Extracting Directional and Comparable Corpora from a Multilingual Corpus for Translation Studies". In: *LREC 2012.*

Chang, Zixia (2009). “A Cognitive-Pragmatic Model for Translation Studies Based on Relevance and Adaptation”. In: *Canadian Social Science* 5.1, pp. 88–111.

Collins-Thompson, Kevyn (2014). “Computational Assessment of Text Readability: A Survey of Past, Present, and Future Research”. Online document. URL: `http://www-personal.umich.edu/~kevynct/pubs/ITL-readability-invited-article-v6-4review.pdf`.

Collins-Thompson, Kevyn, Paul N Bennett, Ryen W White, Sebastian de la Chica, and David Sontag (2011). “Personalizing web search results by reading level”. In: *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, pp. 403–412.

Collins-Thompson, Kevyn and James P Callan (2004). “A Language Modeling Approach to Predicting Reading Difficulty”. In: *HLT-NAACL*.

Cortes, Corinna and Vladimir Vapnik (1995). “Support-vector networks”. In: *Machine learning* 20.3, pp. 273–297.

Crossley, Scott A., David F. Dufty, Philip M. McCarthy, and Danielle S. McNamara (2007). “Toward a New Readability: A Mixed Model Approach”. In: *The 29th annual conference of the Cognitive Science Society*.

Dagan, Ido, Yael Karov, and Dan Roth (1997). “Mistake-driven learning in text categorization”. In: *arXiv preprint cmp-lg/9706006*.

Dale, Edgar and Jeanne S. Chall (1948). “A Formula for Predicting Readability”. In: *Educational Research Bulletin* 27(1), pp. 11–20+28.

Dale, Edgar and Jeanne S Chall (1949). “The concept of readability”. In: *Elementary English* 26.1, pp. 19–26.

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

Das, Dipanjan and Noah A. Smith (2011). “Semi-Supervised Frame-Semantic Parsing for Unknown Predicates”. In: *The Annual Meeting of the Association for Computational Linguistics, Portland*.

Das, Sreerupa and Rajkumar Roychoudhury (2004). “Testing Level of Readability in Bangla novels of Bankim Chandra Chattopodhay w.r.t the Density of Polysyllabic Words”. In: *Indian Journal of Linguistics* 22, pp. 41–51.

— (2006). “Readabilit Modeling and Comparison of One and Two parametric fit: a case study in Bangla”. In: *Journal of Quantative Linguistics* 13(1).

Dash, Manoranjan and Huan Liu (1997). “Feature selection for classification”. In: *Intelligent data analysis* 1.3, pp. 131–156.

Davison, Alice and Robert N. Kantor (1982). "On the Failure of Readability Formulas to Define Readable Texts: A Case Study from Adaptations". In: *Reading Research Quarterly* 17.2, pp. 187–209.

De Belder, Jan and Marie-Francine Moens (2010). "Text simplification for children". In: *Prroceedings of the SIGIR workshop on accessible search systems*, pp. 19–26.

De Beni, Rossana and Paola Palladino (2000). "Intrusion errors in working memory tasks: Are they related to reading comprehension ability?" In: *Learning and Individual Differences* 12.2, pp. 131–143.

De Cock, Rozane (2012). "Children and online news: a suboptimal relationship. Quantitative and qualitative research in Flanders". In: *E-youth: Balancing between opportunities a risks.*

De Cock, Rozane and Eva Hautekiet (2012). "Children's News Online: Website Analysis and Usability Study Results (the United Kingdom, Belgium, and the Netherlands)". In: *Journalism and Mass Communication* 2.12, pp. 1095–1105.

Deerwester, Scott C., Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman (1990). "Indexing by latent semantic analysis". In: *JASIS* 41.6, pp. 391–407.

Dell'Orletta, Felice, Simonetta Montemagni, and Giulia Venturi (2011). "READ–IT: Assessing Readability of Italian Texts with a View to Text Simplification". In: *2nd Workshop on Speech and Language Processing for Assistive Technologies.*

Douma, W.H. (1960). "De lessbaarheid van landbouwbladen:een onderzoek naar en een toepassing van lessbaarheids-formules". In: *Bulletin.*

DuBay, William H (2004). "The Principles of Readability." In: *Online Submission.*

Duarte Torres, Sergio and Ingmar Weber (2011). "What and how children search on the web". In: *Proceedings of the 20th ACM international conference on Information and knowledge management.* ACM, pp. 393–402.

Eickhoff, Carsten, 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.*

Falkenjack, Johan, Katarina Heimann Mühlenbock, and Arne Jönsson (2013). "Features indicating readability in Swedish text". In: *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pp. 27–40.

Feng, Lijun (2010). "Automatic Readability Assessment". PhD thesis. Graduate Faculty in Computer Science, The City University of New York.

Feng, Lijun, 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.*

Feng, Lijun, 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).*

Fernandes, Lincoln (2006). “Corpora in Translation Studies: revisting Baker’s typology”. In: *Fragmentos: Revista de Língua e Literatura* 30, pp. 87–95.

Fillmore, Charles J. (1982). “Frame Semantis”. In: *Linguistics in the Morning Calm.* Hanshin Publishing Co., pp. 111–137.

Fillmore, Charles J., Christopher R. Johnson, and Miriam R.L. Petruck (2003). “Background to Framenet”. In: *International Journal of Lexicography* 16(3), pp. 235–250.

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning (2005). “Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling”. In: *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005).*

Fitzsimmons, PR, BD Michael, JL Hulley, and GO Scott (2010). “A readability assessment of online Parkinson’s disease information”. In: *The Journal of the Royal College of Physicians of Edinburgh* 40, pp. 292–296.

Flekova, Lucie, Oliver Ferschke, and Iryna Gurevych (2014). “What Makes a Good Biography? Multidimensional Quality Analysis Based on Wikipedia Article Feedback Data”. In: *Proceedings of the 23rd International World Wide Web Conference (WWW 2014).*

Flesch, Rudolph (1948). “A new readability yardstick.” In: *Journal of applied psychology* 32.3, p. 221.

Franco-Salvador, Marc, Parth Gupta, and Paolo Rosso (2013). “Cross-language plagiarism detection using a multilingual semantic network”. In: *Advances in Information Retrieval.* Springer, pp. 710–713.

François, Thomas L (2009). “Combining a statistical language model with logistic regression to predict the lexical and syntactic difficulty of texts for FFL”. In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop.* Association for Computational Linguistics, pp. 19–27.

François, Thomas and Cédrick Fairon (2012). “An AI readability formula for French as a foreign language”. In: *Proceedings of the 2012 Joint Conference on Empirical Meth-*

ods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, pp. 466–477.

Frankenberg-Garcia, Ana (2009). "Are translations longer than source texts". In: *A corpus-based study of explicitation In: Beeby, A., Rodríguez P., & Sánchez-Gijón, P.(eds.) Corpus use and learning to translate (CULT): An Introduction. Amsterdam & Philadelphia: John Benjamins*, pp. 47–58.

Friedman, Daniela B and Laurie Hoffman-Goetz (2006). "A systematic review of readability and comprehension instruments used for print and web-based cancer information". In: *Health Education & Behavior* 33.3, pp. 352–373.

Gale, William A. and Kenneth W. Church (1993). "A program for aligning sentences in bilingual corpora". In: *Computational Linguistics* 19(1).

Garciá López, Félix, Miguel Garciá Torres, Belén Melián Batista, José A Moreno Pérez, and J Marcos Moreno-Vega (2006). "Solving feature subset selection problem by a parallel scatter search". In: *European Journal of Operational Research* 169.2, pp. 477–489.

Gellerstam, Martin (1986). "Translationese in Swedish novels translated from English". In: *Translation studies in Scandinavia*, pp. 88–95.

Genzel, Dimitry and Eugene Charniak (2002). "Entropy Rate Constancy in Text". In: *Proceedings of the 40st Meeting of the Association for Computational Linguistics (ACL 2002).*

— (2003). "Variation of Entropy and Parse Trees of Sentences as a Function of the Sentence Number". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Ghani, Kamarulzaman Abdul, Ahmad Sabri Noh, and Nik Mohd Rahimi Nik Yusoff (2014). "Linguistic Features for Development of Arabic Text Readability Formula in Malaysia: A Preliminary Study". In: *Middle-East Journal of Scientific Research* 19.3, pp. 319–331.

Gibson, Edward (1998). "Linguistic complexity: Locality of syntactic dependencies". In: *Cognition* 68.1, pp. 1–76.

Gildea, Daniel and Daniel Jurafsky (2002). "Automatic Labeling of Semantic Roles". In: *Computational Linguistics* 28(3), pp. 245–288.

Giles, Jim (2005). "Internet encyclopaedias go head to head". In: *Nature* 438, 900:901.

Goldberg, David Edward et al. (1989). *Genetic algorithms in search, optimization, and machine learning.* Vol. 412. Addison-wesley Reading Menlo Park.

Göpferich, Susanne (2009). "Towards a model of translation competence and its acquisition: the longitudinal study TransComp". In: *Behind the mind: Methods, models and results in translation process research* 37, p. 11.

Graesser, Arthur C., Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai (2004). "Coh-Metrix: Analysis of text on cohesion and language". In: *Behavior Research Methods, Instruments, and Computers* 36, pp. 193–202.

Gray, William Scott and Bernice Elizabeth Leary (1935). *What makes a book readable.* Univ. Chicago Press.

Guiraud, Pierre (1960). *Problèmes et méthodes de la statistique linguistique.* Presses universitaires de France.

Güngör, Tunga (2003). "Lexical and morphological statistics for Turkish". In: *Proceedings of TAINN*, pp. 409–412.

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

Gutlein, Martin, Eibe Frank, Mark Hall, and Andreas Karwath (2009). "Large-scale attribute selection using wrappers". In: *Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on.* IEEE, pp. 332–339.

Halfaker, Aaron, Oliver Keyes, and Dario Taraborelli (2013). "Making peripheral participation legitimate: reader engagement experiments in wikipedia". In: *Proceedings of the 2013 Conference on Computer Cupported Cooperative Work.*

Hall, M. A. (1998). "Correlation-based Feature Subset Selection for Machine Learning". PhD thesis. Hamilton, New Zealand: University of Waikato.

Hall, Mark Andrew and Geoffrey Holmes (2003). "Benchmarking attribute selection techniques for discrete class data mining". In: *Knowledge and Data Engineering, IEEE Transactions on* 15.6, pp. 1437–1447.

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

Hancke, Julia, Sowmya Vajjala, and Detmar Meurers (2012). "Readability Classification for German using Lexical, Syntactic and Morphological Features". In: *24th International Conference on Computational Linguistics (COLING), Mumbai, India.*

Hansen, Silvia (2003). "The Nature of Translated Text: An Interdisciplinary Methodology for the Investigation of the Specific Properties of Translations". PhD thesis. University of Saarland.

Hansen, Silvia and Elke Teich (2002). “The creation and exploitation of a translation reference corpus”. In: *First International Workshop on Language Resources for Translation Work and Research, 3rd International Conference on Language Resources and Evaluation (LREC-2002).*

Harly, Trevor A. (2008). *The Psychology of Language.* Psychology Press, Taylor and Francis Group.

Harris, Theodore L and Richard E Hodges (1995). *The literacy dictionary: The vocabulary of reading and writing.* International Reading Assoc.

Hasnat, Md. Abul, 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).*

Hayes, Phillip J, Peggy M Andersen, Irene B Nirenburg, and Linda M Schmandt (1990). “Tcs: a shell for content-based text categorization”. In: *Artificial Intelligence Applications, 1990., Sixth Conference on.* IEEE, pp. 320–326.

Heilman, Michael, 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.*

— (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).*

Herdan, Gustav (1964). *Quantitative linguistics.* Butterworths.

Holler, Anke and Lisa Irmen (2007). “Empirically Assessing the Effects of the Right Frontier Constraint”. In: *Anaphora: Analysis, Algorithms and Applications.* Ed. by António Branco. Lecture Notes in Artificial Intelligence. Berlin and Heidelberg: Springer, pp. 15–27.

Holmes, James S (1988). *Translated!: papers on literary translation and translation studies.* 7. Rodopi.

Hönig, Hans G (1991). “Holmes’“Mapping Theory” and the landscape of mental translation processes”. In: *Translation Studies: the state of the art. Amsterdam-Atlanta: Rodopi*, pp. 77–89.

Hotho, Andreas, Andreas Nürnberger, and Gerhard Paaß (2005). “A Brief Survey of Text Mining.” In: *Ldv Forum.* Vol. 20. 1, pp. 19–62.

Ilisei, Iustina (2013). “A Machine Learning Approach to the Identification of Translational Language: An Inquiry into Translationese Learning Models”. PhD thesis. Wolverhampton, UK. URL: `http://clg.wlv.ac.uk/papers/ilisei-thesis.pdf`.

Ilisei, Iustina and Diana Inkpen (2011). “Translationese traits in romanian newspapers: A machine learning approach”. In: *International Journal of Computational Linguistics and Applications* 2.1–2.

Ilisei, Iustina, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov (2009). “Towards Simplification: A Supervised Learning Approach”. In: *Proceedings of Machine Translation 25 Years On, London, United Kingdom, November 21-22.*

— (2010). “Identification of translationese: A machine learning approach”. In: *Proceedings of CICLing 2010 LNCS 6008.* Springer, pp. 503–511.

Inhoff, Albrecht Werner, Ralph Radach, and Dieter Heller (2000). “Complex compounds in German: Interword spaces facilitate segmentation but hinder assignment of meaning”. In: *Journal of Memory and Language* 42.1, pp. 23–50.

Islam, Md. Saiful (2008). “Research on Bangla Language Processing in Bangladesh: Progress and Challenges”. In: *8th International Language & Development Conference.*

Islam, Md. Zahurul, Md. Rashedur Rahman, and Alexander Mehler (2014). “Readability Classification of Bangla Texts”. In: *15th International Conference on Intelligent Text Processing and Computational Linguistics (cicLing), Kathmandu, Nepal.*

Islam, Md Zahurul, Jörg Tiedemann, and Andreas Eisele (2010). “English to Bangla phrase-based machine translation”. In:

Islam, Zahurul and Armin Hoenen (2013). “Source and Translation Classification using Most Frequent Words”. In: *6th International Joint Conference on Natural Language Processing (IJCNLP).*

Islam, Zahurul and Alexander Mehler (2012). “Customization of the Europarl Corpus for Translation Studies”. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC).*

— (2013). “Automatic Readability Classification of Crowd-Sourced Data based on Linguistic and Information-Theoretic Features”. In: *14th International Conference on Intelligent Text Processing and Computational Linguistics.*

Islam, Zahurul, Alexander Mehler, and Rasherdur Rahman (2012). “Text Readability Classification of Textbooks of a Low-Resource Language”. In: *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation.*

Islam, Zahurul and Rashedur Rahman (2014). "Readability of Bangla News Articles for Children". In: *The 28th Pacific Asia Conference on Language, Information and Computation (PACLIC 2014).*

István, Kenesei (1989). "A nyelv és a nyelvek". In: *Bp., Gondolat.*

Jatowt, Adam and Katsumi Tanaka (2012). "Is Wikipedia too difficult?: comparative analysis of readability of Wikipedia, Simple Wikipedia and Britannica". In: *Proceedings of the 21st ACM international conference on Information and knowledge management.* ACM, pp. 2607–2610.

Joachims, Thorsten (1998). *Text categorization with support vector machines: Learning with many relevant features.* Springer.

— (1999). "Transductive inference for text classification using support vector machines". In: *ICML*. Vol. 99, pp. 200–209.

Kali, Robert V. (2009). *Children and Their Development.* Pearson Education.

Kate, Rohit J., 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).*

Keerthi, S.S., S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy (2001). "Improvements to Platt's SMO Algorithm for SVM Classifier Design". In: *Neural Computation* 13(3), pp. 637–649.

Keller, Frank (2004). "The Entropy Rate Principle as a Predictor of Processing Effort: An Evaluation against Eye-tracking Data." In: *EMNLP*, pp. 317–324.

Kharanauli, Anna (2000). "Einführung in die georgische Psalterübersetzung". In: ed. by Anneli; Quast Aejmelaeus. Vandenhoeck & Ruprecht, pp. 248–308.

Kidwell, Paul, Guy Lebanon, and Kevyn Collins-Thompson (2011). "Statistical estimation of word acquisition with application to readability prediction". In: *Journal of the American Statistical Association* 106.493, pp. 21–30.

Kincaid, J Peter, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.* Tech. rep. DTIC Document.

Kincaid, J., R. Fishburne, R. Rodegers, and B. Chissom (1975). *Derivation of new readability formulas for Navy enlisted personnel.* Tech. rep. US Navy, Branch Report 8-75, Cheif of Naval Training.

Kintsch, Walter (1998). *Comprehension: A paradigm for cognition.* Cambridge university press.

Kireyev, Kirill and Thomas K Landauer (2011). “Word maturity: Computational modeling of word knowledge”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1.* Association for Computational Linguistics, pp. 299–308.

Klein, Dan and Christopher D. Manning (2003). “Accurate Unlexicalized Parsing”. In: *Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL 2003).*

Klir, George Jiri (2005). *Uncertainty and Information.* Wiley-Interscience.

Knight, Kevin (1999). “Mining online text”. In: *Communications of the ACM* 42.11, pp. 58–61.

Koehn, Philipp (2005). “Europarl: A Parallel Corpus for Statistical Machine Translation”. In: *MT Summit.*

Köhler, Reinhard and Matthias Galle (1993). “Dynamic aspects of text characteristics”. In: *Quantitative text analysis*, pp. 46–53.

Koppel, Moshe and Noam Ordan (2011). “Translationese and Its Dialects”. In: *49th Annual Meeting of the Association for Computational Linguistics (ACL).*

Kornai, András (2008). *Mathematical Linguistics.* Springer.

Kruger, Alet, Jeremy Munday, and Kim Wallmach (2011). *Corpus-based Translation Studies: Research and Applications.* Continuum International Publishing Group.

Landauer, T and D Way (2011). “Improving text complexity measurement through the reading maturity metric”. In: *annual meeting of the National Council on Measurement in Education, Vancouver, BC.*

Landauer, Thomas K, Kirill Kireyev, and Charles Panaccione (2011). “Word Maturity: A new metric for word knowledge”. In: *Scientific Studies of Reading* 15.1, pp. 92–108.

Lang, David Marshall (1957). “Recent Work on the Georgian New Testament”. In: *Bulletin of the School of Oriental and African Studies* 19.01, pp. 82–93.

Laviosa, Sara (1998). “The Corpus-based Approach: A New Paradigm in Translation Studies”. In: *journal des traducteurs / Meta: Translators' Journal* 43(4), pp. 474–479.

— (2002). *Corpus-based translation studies. Theory, findings, applications.* Amsterdam/New York: Rodopi.

Learning, Renaissance (2001). “The ATOS readability formula for books and how it compares to other formulas”. In: *Madison, WI: School Renaissance Institute.*

Lefevere, André (2002). *Translation/history/culture: A sourcebook.* Routledge.

Lembersky, Gennadi, Noam Ordan, and Shuly Wintner (2011). "Language Models for Machine Translation: Original Vs. Translated Texts". In: *Empirical Methods in Natural Language Processing (EMNLP).*

Lennon, Colleen and Hal Burdick (2004). "The lexile framework as an approach for reading measurement and success". In: *electronic publication on www. lexile. com.*

Levenshtein, Vladimir I (1966). "Binary codes capable of correcting deletions, insertions and reversals". In: *Soviet physics doklady.*

Lieberman, Erez, Jean-Baptiste Michel, Joe Jackson, Tina Tang, and Martin A Nowak (2007). "Quantifying the evolutionary dynamics of language". In: *Nature* 449.7163, pp. 713–716.

Liu, H. and R. Setiono (1996). "A probabilistic approach to feature selection - A filter solution". In: *13th International Conference on Machine Learning*, pp. 319–327.

Livingstone, Sonia, Leslie Haddon, Anke Görzig, and Kjartan Ólafsson (2010). "Risks and safety for children on the internet: the UK report". In: *Politics* 6.2010, p. 1.

Lodhi, Huma, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins (2002). "Text classification using string kernels". In: *The Journal of Machine Learning Research* 2, pp. 419–444.

Lu, Xiaofei (2012). "The relationship of lexical richness to the quality of ESL learners' oral narratives". In: *The Modern Language Journal* 96.2, pp. 190–208.

Lzwaini, Sattar (2003). "Building specialised corpora for translation studies". In: *Workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives, Corpus Linguistics.*

Ma, Yi, Ritu Singh, Eric Fosler-Lussier, and Robert Lofthus (2012). "Comparing human versus automatic feature extraction for fine-grained elementary readability assesment". In: *NAACL-HLT 2012 Workshop on Predicting and Improving Text Readability for target reader populations.*

Mahowald, Kyle, Evelina Fedorenko, Steven T Piantadosi, and Edward Gibson (2013). "Info/information theory: Speakers choose shorter words in predictive contexts". In: *Cognition* 126.2, pp. 313–318.

Mann, William and Sarah Thompson (1988). "Rhethorical Structure Theory: Towards a Functional Theory of Text Organization". In: *Text* 8.3, pp. 243–281.

Martin, Gale L (2004). "Encoder: a connectionist model of how learning to visually encode fixated text images improves reading fluency." In: *Psychological review* 111.3, p. 617.

Mauranen, Anna and Pekka Kujamäki (2004). *Translation universals: do they exist?* Vol. 48. John Benjamins Publishing.

McLaughlin, G. Harry (1969). "SMOG Grading – a New Readability Formula". In: *Journal of Reading* 12.8, pp. 639–646.

Mccallum, Andrew and Kamal Nigam (1998). "A Comparison of Event Models for Naive Bayes Text Classification". In: *AAAI-98 Workshop on 'Learning for Text Categorization.*

Mehler, Alexander (2006). "Stratified Constraint Satisfaction Networks in Synergetic Multi-Agent Simulations of Language Evolution". In: *Artificial Cognition Systems.* Ed. by Angelo Loula, Ricardo Gudwin, and João Queiroz. Hershey: Idea Group Inc., pp. 140–174.

Mehler, Alexander and Christian Wolff (2005). "Einleitung: Perspektiven und Positionen des Text Mining". In: *LDV-Forum.* Vol. 20. 1.

Mikk, Jaan (2005). "Quantitative Linguistik/Quantitative Linguistics: Ein internationales Handbuch/An International Handbook". In: ed. by Reinhard Köhler, Gabriel Altmann, and Rajmund G Piotrowski. Walter de Gruyter. Chap. Text comprehensibility, pp. 909–921.

Miller, James R and Walter Kintsch (1980). "Readability and recall of short prose passages: A theoretical analysis." In: *Journal of Experimental Psychology: Human Learning and Memory* 6.4, p. 335.

Montemurro, Marcelo A and Damián H Zanette (2002). "Entropic analysis of the role of words in literary texts". In: *Advances in complex systems* 5.01, pp. 7–17.

— (2010). "Towards the quantification of the semantic information encoded in written language". In: *Advances in Complex Systems* 13.02, pp. 135–153.

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

Nigam, Kamal, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell (2000). "Text classification from labeled and unlabeled documents using EM". In: *Machine learning* 39.2-3, pp. 103–134.

Nowak, Martin A, Natalia L Komarova, and Partha Niyogi (2002). "Computational and evolutionary aspects of language". In: *Nature* 417.6889, pp. 611–617.

Olohan, Maeve (2001). "Spelling out the optionals in translation:A corpus study". In: *Corpus Linguistics 2001 conference. UCREL Technical Paper number 13. Special issue.*

Olohan, Maeve (2004). *Introducing Corpora in Translation Studies.* London/New York: Routledge.

Oosten, Philip van, Dries Tanghe, and Veronique Hoste (2010). "Towards an improved methodology for automated readability prediction". In: *7th Conference on International Language Resources and Evaluation (LREC 2010).* European Language Resources Association (ELRA), pp. 775–782.

Pastor, Gloria Corpas, Ruslan Mitkov, Naveed Afzal, and Viktor Pekar (2008). "Translation universals: do they exist? A corpus-based NLP study of convergence and simplification". In: *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (AMTA-08).*

Pazienza, Maria Teresa (1997). *Information Extraction: A multidisciplinary approach to an emerging information technology.* Springer.

Pennebaker, Jams W., Martha E. Francis, and Roger J. Booth (2001). *Linguistic Inquiry and Word Count (LIWC): LIWC2001 Manual.* Erlbaum Publishers.

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

Pitler, Emily 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).*

Pitt, Ellen and Richi Nayak (2007). "The use of various data mining and feature selection methods in the analysis of a population survey dataset". In: *Proceedings of the 2nd international workshop on Integrating artificial intelligence and data mining-Volume 84.* Australian Computer Society, Inc., pp. 83–93.

Platt, John C. (1998). *Fast training of support vector machines using sequential minimal optimization.* Ed. by Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola. MIT Press.

Plotkin, Joshua B. and Martik A. Nowak (2000). "Language Evolution and Information Theory". In: *Journal of Theoretical Biology* 205(1), pp. 147–159.

Polanyi, Livia (1988). "A formal model of the structure of discourse". In: *Journal of Pragmatics* 12.56, pp. 601–638.

Popescu, Marius (2011). "Studying Translationese at the Character Level". In: *Recent Advances in Natural Language Processing.*

Pym, Anthony (2005). "Explaining Explicitation". In: *New Trends in Translation Studies. In Honour of Kinga Klaudy.* Akadémia Kiadó, pp. 29–34.

Rayner, Keith (1998). "Eye movements in reading and information processing: 20 years of research." In: *Psychological bulletin* 124.3, p. 372.

Rayner, Keith, Alexander Pollatsek, Jane Ashby, and Charles Clifton Jr (2012). *Psychology of Reading.* Psychology Press.

Reichle, Erik D, Alexander Pollatsek, Donald L Fisher, and Keith Rayner (1998). "Toward a model of eye movement control in reading." In: *Psychological review* 105.1, p. 125.

Reichle, Erik D, Keith Rayner, and Alexander Pollatsek (2003). "The EZ Reader model of eye-movement control in reading: Comparisons to other models". In: *Behavioral and brain sciences* 26.4, pp. 445–476.

Rello, Luz, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion (2013). "Frequent words improve readability and short words improve understandability for people with dyslexia". In: *Human-Computer Interaction–INTERACT 2013.* Springer, pp. 203–219.

Rubin, Andee (1981). "Conceptual Readability: New Ways to Look at Text. Reading Education Report No. 31." In:

Sato, Satoshi, Suguru Matsuyoshi, and Yohsuke Kondoh (2008). "Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus." In: *LREC.*

Schank, Roger C and Robert P Abelson (1977). *Scripts, plans, goals and understanding: an inquiry into human knowledge structures. Hillsdale, NJ: L.* Erlbaum.

Schwarm, Sarah E. 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).*

Sebastiani, Fabrizio (2002). "Machine learning in automated text categorization". In: *ACM computing surveys (CSUR)* 34.1, pp. 1–47.

Seddah, Djamé, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, et al. (2013). "Overview of the SPMRL 2013 shared task: cross-framework evaluation of parsing morphologically rich languages". In: Association for Computational Linguistics.

Séguinot, Candace (1988). "Pragmatics and the explicitation hypothesis". In: *TTR: traduction, terminologie, rédaction* 1.2, pp. 106–113.

Senter, R.J. and E. A. Smith (1967). *Automated Readability Index.* Tech. rep. Wright-Patterson Air Force Base.

Shannon, Claude Elwood (1948). "A Mathematical Theory of Communication". In: *The Bell System Technical Journal* 27(1), pp. 379–423.

Sherman, Lucius Adelno (1893). "Analytics of literature: A manual for the objective study of english poetry and prose". In: *Boston: Ginn.*

Shreve, Gregory M (1997). "Cognition and the evolution of translation competence". In: *APPLIED PSYCHOLOGY-LONDON-SAGE-* 3, pp. 120–136.

Si, Luo and Jamie Callan (2001). "A Statistical Model for Scientific Readability". In: *Tenth International Conference on Information and Knowledge Management.*

Sinha, Manjira, Sakshi Sharma, Tirthankar Dasgupta, and Anupam Basu (2012). "New Readability Measures for Bangla and Hindi Texts." In: *COLING (Posters)*, pp. 1141–1150.

Slaney, John, Masayuki Fujita, and Mark Stickel (1995). "Automated reasoning and exhaustive search: Quasigroup existence problems". In: *Computers & mathematics with applications* 29.2, pp. 115–132.

Stockmeyer, Norman (2009). "Using Microsoft Word's readability program". In: *Michigan Bar Journal* 88, p. 46.

Stoyanov, Veselin, Claire Cardie, Nathan Gilbert, Ellen Riloff, David Buttler, and David Hysom (2010). "Coreference Resolution with Reconcile". In: *Conference of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), Short Paper.*

TEI Consortium (2008). *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* `http://www.tei-c.org/Guidelines/P5/`.

Temnikova, Irina (2012). "Text Complexity and Text Simplification in the Crisis Management Domain". PhD thesis. University of Wolverhampton.

Toury, Gideon (1995). *Descriptive Translation Studies and Beyond.* John Benjamins, Amsterdam/Philadelphia.

— (2004). "Probabilistic explanations in translation studies". In: *Translation Universals. Do They Exist*, pp. 15–32.

Tuldava, Juhan (1993). "The statistical structure of a text and its readability". In: *Quantitative text analysis*, pp. 215–227.

Tymoczko, Maria (1998). "Computerized corpora and the future of translation studies". In: *Meta* 43.4, pp. 652–659.

Üstün, B., 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". In: *Chemometrics and Intelligent Laboratory Systems* 81.1, pp. 29–40.

Vajjala, Sowmya and Detmar Meurers (2012). “On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition”. In: *The 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7), Association for Computational Linguistics.*

— (2013). “On the applicability of readability models to web texts”. In: *2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations, ACL.*

Van Halteren, Hans (2008). “Source Language Markers in EUROPARL Translations”. In: *International Conference inComputational Linguistics(COLING)*, pp. 937–944.

Vapnik, Vladimir (2000). *The nature of statistical learning theory.* springer.

Vinay, Jean-Paul and Jean Louis Darbelnet (1995). *Comparative stylistics of French and English: a methodology for translation.* Vol. 11. John Benjamins.

Vinay, Jean-Paul and Jean Darbelnet (1958). *Stylistique comparée de l’anglais et du français.*

Volansky, Vered, Noam Ordan, and Shuly Wintner (2013). “On the features of translationese”. In: *Literary and Linguistic Computing*, fqt031.

Waltinger, Ulli (2010). “On social semantics in information retrieval”. PhD thesis. Bielefeld University.

Whaley, CP (1978). “Word—nonword classification time”. In: *Journal of Verbal Learning and Verbal Behavior* 17.2, pp. 143–154.

Wimmer, Gejza (2008). “The Type-Token Relation”. In: ed. by Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski. deGruyter. Chap. Text/Fields an Phonema:Text, pp. 361–368.

Wimmer, Gejza and Gabriel Altmann (1999). “Review article: On vocabulary richness”. In: *Journal of Quantitative Linguistics* 6.1, pp. 1–9.

Yuka, Tateisi, Ono Yoshihiko, and Yamada Hisao (1988). “A computer readability formula of Japanese texts for machine scoring”. In: *Proceedings of the 12th conference on Computational linguistics-Volume 2.* Association for Computational Linguistics, pp. 649–654.

Zipf, George Kingsley (1935). “The psycho-biology of language.” In:

— (1949). “Human behavior and the principle of least effort.” In: