UPPSALA
UNIVERSITET

Department of Linguistics and Philology
Språkteknologiprogrammet
(Language Technology Programme)
Master's thesis in Computational Linguistics

14 juni 2006

Classification into Readability Levels

Implementation and Evaluation

Patrik Larsson

Supervisor:
Beata Megyesi, Uppsala Universitet

Abstract

The use for a readability classification model is mainly as an integrated part of an information retrieval system. By matching the user's demands of readability to the documents with the corresponding readability, the classification model can further improve the results of, for example, a search engine.

This thesis presents a new solution for classification into readability levels for Swedish. The results from the thesis are a number of classification models. The models were induced by training a Support Vector Machines classifier on features that are established by previous research as good measurements of readability. The features were extracted from a corpus annotated with three readability levels. Natural Language Processing tools for tagging and parsing were used to analyze the corpus and enable the extraction of the features from the corpus. Empirical testings of different feature combinations were performed to optimize the classification model.

The classification models render a good and stable classification. The best model obtained a precision score of 90.21% and a recall score of 89.56% on the test-set, which is equal to a F-score of 89.88.

Contents

Acknowledgments

I would like to thank my supervisor at the Department of Linguistics and Philology, Beata Megyesi, for her advices and support, regarding both the text and work procedure.

Further I would like to thank Lina Ekerljung, for her help and suggestions of improvements of the text. I also would like to thank everyone who has helped me by providing texts or references.

A special thanks to Ulf Daneklev at Forskning.se, for giving me the idea of a readability classifier.

1 Introduction

Readability is an important field of research as the importance for companies, authorities and organizations to reach the public with their messages increases. The main channels to reach the public are often through written material, published either in newspapers or on the internet. The space is often limited so the importance of compact texts is crucial, which often leads to texts that are difficult to understand, in terms of unusual words, long sentences and rare syntactic constructions. Many people benefit from research in readability, not only big companies and authorities. School teachers, writers or just anyone who intends to write or select reading material for a specific group of readers, need to establish the readability of texts. The amounts of documents that are published on the internet increases every day. Search engines are used by people to find the interesting documents on the internet. The results from a search engine can be correct and concern the right topic according to the query, but the readability of the documents that are retrieved might not match the demands of the users.

To address all of these demands there is a need for an adaptive and well performing solution to tell whether a text is suitable for the target group or not. There are existing methods for measuring readability, but most of them are designed for manual analysis and has been proven to perform poorly in tasks like in a search engine [Collins-Thompson and Callan, 2004].

The factors that affect readability have been established by previous research. Many different factors influence readability for example, content of the texts, vocabulary and style of writing. The concept of *readable or not*, is not as important as the concept of *who can read this?*, since texts often are produced for a specific target group. The term *readability level*, is often used to describe the educational level a reader needs to understand a text.

New effective computerized tools for linguistic analysis of data provide a new approach to readability research. By letting the computer do the work, it is possible to analyze large amounts of data by looking at the readability factors established by previous research.

1.1 Purpose of the thesis

The purpose of the thesis is to create models for automatic classification of texts into different predefined readability levels. The models are created by machine learning techniques and well founded, linguistically motivated measurements of readability. Feature combinations, derived measurements from the text, are empirically tested during classification to see which combinations that are the most representative to predict each readability level. The model is created for classification of Swedish texts,

but the methods and much of the techniques are applicable for other languages as well. The evaluation of the model is done using standard evaluation techniques for classification of texts, the reason for this is to facilitate future comparisons with the model.

The models could function either as stand-alone applications for classifying text, useful for teachers or writers, or as an integrated part of an information retrieval system. The main use of the models created in this study is in an information retrieval system.

1.2 Outline

The thesis consist of four main parts. In chapter 1, there is an introduction to the subject and the purpose of the thesis. The background and previous research of both readability and classification into readability levels are presented in chapter 2. In chapter 3 the methods, tools and resources, used in this thesis are presented. Chapter 4 and chapter 5 provide the method for evaluation, the results of the evaluation and a discussion of the results.

2 Readability

Below follows a brief overview of the research of readability and readability levels, a section focusing on the traditional formulas for measuring readability and a section describing data-driven research.

2.1 Readability research

Readability can be defined in several ways, Björnsson [1971] defines it as:
The sum of linguistic properties in a text, that makes the text more or less available to the reader.
This definition separates readability from issues concerning the layout and how interesting the text is. There are three ways the term readability is used in research [Klare, 1963]:

1. To indicate legibility of either handwriting or typography.
2. To indicate ease of reading due to the interest-value or the pleasantness of writing.
3. To indicate ease of understanding or comprehension due to the style of writing.

The first definition focus on the presentation of the text, typeface, colors and the placing of pictures are important factors. Modern research focus on the first definition since the use of web pages has increased and more people act as web-designers. The advanced lay-out options, available for web-designers and software developers makes the research important from a commercial point of view.

The second definition covers the content of the text. If the text is interesting, exciting or annoying, it influences the readability. Most research concerning the second definition is performed by studying children's comprehension of texts.

The third definition defines the readability through looking at the structure of sentences, words and phrases, and based on that decide how difficult the text is to read.

In some ways the second definition intersects with the third, since it is difficult to separate the views in test situations. If the text is difficult to read because of the style of writing, this will influence the interest value of the text. The definition of readability used in this thesis is the third definition.

There are methods for determining the readability, defined as above, of texts automatically. Methods that analyze linguistic properties of the text and returns a measurement of the readability, are called readability formulas. Readability formulas are traditionally constructed by linguists for manual analysis of texts, but some of them have been computerized. The usage is traditionally for someone who are supposed to

write or select a text for a specific group of readers, where a tool for classifying texts into a readability level can help the user to adjust the complexity of the text to an appropriate level. Readability formulas are used by teachers to guide in the selection of reading material for their students. Some well-known word processor programs have implementations of one of the most popular readability formulas; *Flesch-Kincaid Reading Ease* (see section 2.2.1), which gives the user a tool to measure the readability. The readability formulas perform poorly when integrated in a search engine, mainly because they often need at least 100 words and well formed sentences to be able to analyze the text [Collins-Thompson and Callan, 2004].

There is a lot of readability research available, of which most were conducted between 1930 and 1960. The research has resulted in numerous, more or less used and accepted, formulas for deciding the readability of a text for different languages. Since readability formulas are language dependent it is not possible to take a formula for English or German and apply it to Swedish [Klare, 1963]. Most existing formulas are designed for American English. Furthermore, many of the existing readability formulas are designed to cope with a small portion of the written language, e.g. a formula classifies into a level that correlates to an American school grade or one of different adult levels. The fact that the formulas are so unadaptable, makes readability formulas less suitable to computerize. Readability formulas are mathematical formulas initially designed to determine the suitability of books for students at a certain age or grade level. As an implication of the fact that the readability was measured manually and not by a computer, readability formulas tend to contain as few features as possible. The different features were evaluated against each other to see which ones that correlate and capture the same phenomenon. A feature that is a good indicator of readability may still be left out because of the increased workload to measure it [Björnsson, 1971].

Modern research on traditional readability formulas concentrates on updating existing formulas and adapting them to the new techniques available.

There have been recent attempts to use techniques from computer science to solve the problem of establishing the readability of texts. These techniques use preclassified data to compare the unclassified data with, therefore the techniques are called data-driven.

In the next two sections there is an overview of the traditional solutions of establishing the readability along with an overview of the data-driven solutions.

2.2 Existing methods for measuring readability

2.2.1 Traditional formulas

The common structure of a readability formula is that it consists of a limited amount of parameters that reflect either the reading difficulties on the word level, or on the sentence level [Björnsson, 1971]. For example the parameters *average sentence-length* and *number of different words* are often used. Almost every readability formula consists of parameters that represent either semantic or syntactic complexity. When a readability formula is constructed, every parameter is calculated on classified reading material. The parameters get a score of how well they correlate with the classified data and then the parameters, or combinations of parameters, with the highest score are used in the formula. This is called a *multiple correlation analysis*

[Cedergren, 1992]. The parameters are assigned a constant that reflects the importance of the feature in the multiple correlation analysis [Björnsson, 1971]. In the formulas below, the constants are the numbers, multiplied with the different parameters. The output of a readability formula is usually a classification into the level of education the reader needs to be able to understand the text. Some readability formulas return a score, where a higher score indicates more difficult text, and has to be interpreted with a scale to get the exact difficulty.

Many of the traditional formulas were originally designed in the first half of the 20th century and have since then been revised several times. These are well summarized in *The measurement of readability* [Klare, 1963].

A common way to measure the reading difficulty of a text is by assuming that unusual words are more difficult to read than common words. Many readability formulas estimate the proportion of common words. Both *Lorges-* (see 1 below) and *Dale-Chall* (see 2 below) formulas use lists of the most common words for English. There are different lists containing the most common words for English, most of them are subsets of the *Dale list of 3000*, which contains 3000 words claimed to be understandable by 80% of american fourth graders [Dale and Chall, 1948]. The *Dale 769-word list* is a subset of the bigger Dale list.

Below is a description of the most well-known readability formulas, they are designed for American English if nothing else is stated in the description.

1. **Lorges Formula** (1939) revised in 1948, classifies texts into grades 3-12.

   $Grade = .07sl + .1073w_d + .1301pp + 1.6126$ where

   - sl = Average sentence length.
   - w_d = Number of different difficult words per 100 words. Difficult words are words not on the *Dale 769-word list*.
   - pp = Number of prepositional phrases per 100 words.

   Lorges formula is considered as the best among the earliest formulas. It is relatively straightforward to calculate, which made it popular.

2. **Dale-Challs Formula** (1948) revised in 1995, classifies texts into grades 3-12.

   $Grade = .0596sl + .1579w_d + 3.6365$

   - sl = Average sentence length.
   - w_d = The percentage of words not occurring on the *Dale list of 3000*.

   In 1995, the *Dale list of 3000* was updated and the formula was changed. The reason why this formula is less known and used than, for example Fleschs Reading Ease formula is that it is more difficult to calculate, since checking of the 3000 words on the list is a time consuming task.

3. **Flesch Reading Ease formula** (1948) revised several times, returns a score where a higher score indicates a more difficult text.

   $ReadingEase = 206.835 - 1.015sl - .846wl$

- sl = Average number of word per sentence.
- wl = Number of syllables per 100 words.

Flesch Reading Ease formula returns a number between 0-100, where a higher score indicates that the text is harder to read. The formula is very simple to calculate, since the text passage to analyze has to be only 100 words and only two features need to be investigated. Reading Ease has become U.S governmental standard and most states in the U.S, require insurance forms to score at a certain level (around 40-50) to be valid. The formula is also used in several word processors as a service to test the documents readability. Flesch Reading Ease formula is the most used and well-known formula and it has influenced formulas for other languages because of its high correlation score and the simple calculation.

4. **Flesch-Kincaid Grade Level** (1975), classifies texts into American school grades.

   $$Grade = .39sl + 11.8wl - 15.59$$

   - sl = Average number of word per sentence.
   - wl = Average number of syllables per word.

   Flesch-Kincaid Grade Level is a modification of Fleschs *Reading Ease Formula*. It translates the former formula to an U.S grade level. A score of 10.2 indicates that the text is understandable for a 10th grade American student. Flesch-Kincaid Grade Level formula is the most used formula that classifies texts into a grade level.

5. **LIX (Läsbarhetsindex)** (1968) revised a few times, developed for Swedish.

   $$LIX = wl/s + 100 * w_d/wl$$

   - wl = Number of words in the text.
   - s = Number of sentences in the text.
   - w_d = Number of difficult words in the text, where difficult words is defined as words consisting of more than six letters.

   The value from *LIX* has to be interpreted with a *LIX-interpreter*. There are several available, two examples of interpreters is represented in Figure 2.1 [Björnsson, 1971]. Depending on which *LIX-interpreter* is used, *LIX* can be applied to any level of text, just by adjusting the scale. *LIX* have successfully been applied to several other languages with the good results, by simply adjusting the scale of the interpreter [Cedergren, 1992].

Evaluation of traditional methods

To evaluate traditional readability formulas, a reference criterion is usually used [Klare, 1963]. The outcome of the readability formula is then compared with the reference criterion and a score of how well it correlates is calculated. Below are some of the most used reference criteria for evaluating the traditional formulas.

LIX value	Description	LIX value	Description
20	Very easy	20-25	Children's books
30	Easy	31-35	Fiction
40	Average	40-45	Newspapers
50	Hard	50-55	Science reports
60	Very Hard	60-	Government texts, law texts ...

Figure 2.1: Two LIX-interpreters, the left is a simple version and the right an adapted interpreter.

1. *Cloze-procedure* is a way to determine the readability of a text by letting readers fill in left-out words in a text. Usually, every fifth word is excluded. There are differences in what accuracy to demand for a text to be classified as readable, but 50% accuracy on the left-out-words is a common threshold [Backman, 1976]. The level of education of the reader is used as the readability level of the text.

2. *Test questions* is a way of determining the readability level of the text by letting a reader answer a number of questions about the text. If the reader can answer a certain amount of the questions correctly, the reader's educational level is used to classify the text. [Backman, 1976]

3. *McCall-Crabbs Standard Test-lessons* is a series of reading tests/practice material for different levels of U.S. education. McCall-Crabbs Test-lessons are probably the most used criterion as reference material [Klare, 1963]. Both the *Dale-Chall formula* and the Flesch Reading Ease (described in section 2.2.1) used this as their reference criterion.

4. *Existing readability formulas* as a reference criterion is a common way to see which formulas that roughly cover the same data [Klare, 1963]. The *Flesch formulas* have in some tests correlated as high as .98 with the *Dale-Chall formula* (the formulas are described in 2.2.1).

5. *Expert classification* means that one or more experts of estimating readability, assign readability levels for a collection of texts.

The fact that there is no standardized way of evaluating readability formulas can make it hard to compare different formulas. Backman [1976] performed a critical evaluation of the constituents of the Swedish readability formula *LIX* (described in section 2.2.1), he also examined the evaluation procedures of readability formulas. He stated that *LIX*, and most other readability formulas, are in fact not measuring the readability. Backman's main critique about the constituents concerns the measurement *sentence-length*. He illustrates this by performing empirical tests that indicates that the often used parameter *sentence-length* does not affect the readability for Swedish and probably not for other languages either. If it was a good measurement then all sentences with equal length would be equally difficult to understand, which is not the case. *Frequency of difficult words* is the other often used parameter in formulas. In *LIX* it is defined as words containing more than six letters. Backman means that even though it is hard to make an obvious decision about where to draw the line for the number of letters forming a difficult word, it is probably a pretty good reflection of unfrequent words. Long, frequent words have a tendency to be shortened by

time. Backman states that *LIX* and most other formulas are not good measurements of readability but rather, good measurements of how to pass the evaluation with good scores.

Backman criticize the construction of both *Cloze-procedure* and *Test questions*. He states that in *cloze-procedures*, the factors of guessing, and words frequently occurring in the same context, makes the results unreliable. He refers to independent studies of *Test questions* which states that the tests seldom fulfill the most basic requirements of measuring a reader's understanding of the text.

Traditional formulas are widely used, despite the disadvantages presented here. That indicates that the need of an automatic way of determining the readability is great.

2.2.2 Data-driven methods

Recent studies of classification into readability levels origin from the regular classification techniques used in, for example, text classification. These methods differ from the traditional readability formulas and so is the use of them. The data-driven methods can be applied to the same tasks as a traditional readability formula if the classifier is trained on a representative data-set for each class. However, the use of data-driven methods require a large amount of data. The data consists of texts, classified into readability levels for training of the model. Other requirements are an algorithm that creates the model, a set of features to represent the data and different tools for extracting the features from the texts. The use of a data-driven classifier gives the user a chance to adjust the readability levels to the preferred ones, compared to traditional readability formulas that already have fixed levels.

Data-driven classification models can easily be retrained on new data depending on the task. That makes data-driven methods well suited for use in classification of readability, as languages change rapidly and so are the types of text to analyze.

There are two studies in automatic classification regarding readability that involves data-driven techniques:

1. *A Language Modeling Approach to Predicting Reading Difficulty* [Collins-Thompson and Callan, 2004] is an approach to solve the classification into readability levels by using multiple language models to estimate the most likely grade level. This is done by using a multinomial naïve Bayes classifier and a language model based on unigrams of words. The model performed significantly higher than the reference methods used for evaluation (unknown types of the *Dale 3000 word-list*) on texts collected from the internet belonging to different grades, but the model was outperformed when classifying reading test documents. This model showed a reliable classification of passages containing 5-8 words into low, medium or high levels of difficulty, which is useful when integrating it as part of a search engine.

2. *Automatic Recognition of Reading Levels from User Queries* [Liu and Oh, 2004] is an attempt to classify texts into readability levels based on the queries to a search engine. This requires training the model on authentic user queries, since user queries tend to be very short and incomplete sentences. The model was induced by using Support Vector Machines trained on a number of syntactic and semantic features derived from the authentic user queries. Examples of

the features are *sentence-length*, *average number of syllables per word* and the output from traditional readability formulas. The model obtained significantly higher correlation scores in the evaluation compared to traditional formulas when applied to user queries.

3 Automatic classification of readability

A data-driven solution for classifciation into readability levels is presented here. The quality of the classification depends on a number of crucial factors which affect the process.

The first important factor is the selection of the data-set to induce the model from. The corpus has to be partitioned into different readability levels, where each partition clearly represents a level of readability. It is the different partitions of data that defines the readability levels. The stability and reliability of the classification depends on the amount of data. A large amount of data renders a stable and reliable model.

The second factor is the algorithm. An efficient and robust classification algorithm is needed to induce the model and classify the data. The algorithm need to be fast in the classification task to function in an application and it also has to be able to handle multiclass classification and multidimensional data.

The third crucial factor is the selection of the features to be extracted from the data. The features have to be good measurements of readability. That is why the features have to be motivated by previous research to be a measurement of readability and also evaluated empirically. The features also have to be possible to extract automatically with high quality and the extraction has to be reasonably fast.

3.1 Data

3.1.1 Corpus Development

There is no corpus available for Swedish that is annotated with reading difficulty levels. The corpus used in the project is assembled for research purposes and not for usage in any particular application. For use in an application, the data-set needs to be created based on the field of the application. The data used in this project can be divided into three readability levels.

- **Morning paper texts**
  The data-set consist of articles taken from the data used in the SCARRIE project [Dahlqvist, 1999]. The articles are from two Swedish morning papers, Uppsala Nya Tidning (UNT) and Svenska Dagbladet (SvD). The number of articles available from the SCARRIE project is more than 100.000. The texts are aimed at adult readers and written by professional writers, probably well-educated in writing. This is the main reason why the data-set is considered as the most difficult readability level in this study.

  This level of readability is related to as the *Difficult*-level in the rest of the thesis.

An example of a sentence from the morning paper texts is shown below. The sentence is long and complex due to the subordinated clauses and the rich use of attributes:

```
Det närbelägna fjällpartiet, med växter från våra egna fjäll, är
nu smyckat av en av mina favoriter, klippveronikan, blåblommig
med rött öga. (“The nearby mountain, with plants from our own
mountains, is now decorated by one of my own favorites, the Rock
Speedwell, blue flowery with a red eye”.)
```

- **High school student texts**
  A collection of texts, written by students at the age of 16-18 years old as a part of an exam in Swedish. The data-set consists of 418 texts, written about a topic given in the exam. The quality of the texts varies greatly between the writers but most of the documents in this group are clustered at an equal level of writing. The documents in this class are written by people who write at a daily basis, but have not yet developed a professional writer's vocabulary and style of writing. That is the reason why this data-set is considered as the medium readability level.

  This readability level is related to as the *Medium*-level in the rest of this thesis.

  An example from the high school student texts which illustrates the unusual high frequent use of prepositions and pronouns:

```
Detta kan man göra på så sätt, att man läser för dem på ett
intressant och roligt sätt. ("This one can do in such a way, that
one reads to them in an interesting and funny way")
```

- **Easy newspaper texts**
  The texts are newspaper texts from *Sesam*, a newspaper that provides news for people with problems to read Swedish. The data-set consists of 787 articles downloaded from the web-site of Sesam[1]. The articles are written by professional writers but for people with problems of reading, either for beginners of the language or for people with some reading handicap. The data-set is considered as the easiest readability level.

  This readability level is related to as the *Easy*-level in the rest of the thesis.

  An example of the easy newspaper text is shown below. The example illustrates the short and simple sentences, often used in the data-set:

```
Dario Fos pappa var stins. Han arbetade vid järnvägen. Ibland
fick han jobb på en ny station. ("Dario Fo’s father was a
stationmaster. He worked at the railroad. Sometimes he got a job
at a new railway station")
```

[1] The web-site is located at `http://sesam.nu/`

3.1.2 Preprocessing

The corpus was scaled to contain an equal amount of documents from each class. The smallest class is the exam class which contains 418 documents, and the two other classes were trimmed to the exact same size. The reason for scaling the corpus is that it makes the process of finding optimized parameters for the SVM to use in the Kernel function easier (described in section 3.2).

The data consists of a number of documents belonging to the different readability levels described in section 3.1.1. The different features extracted from the texts are not affected by the length of a document. In spite of that, documents containing less than ten sentences are disregarded in the analysis. The reason for this is to simulate the use of the model in a search engine, for example on a company web site, where documents of shorter length hardly would be indexed. The SCARRIE data contains newspaper articles with just a few sentences.

3.2 Support Vector Machines and LIBSVM

Support Vector Machines (SVM) [Vapnik, 1998], is a new technique for automatic classification with it's foundation in statistical learning theory. SVM has shown promising results in many practical solutions, e.g. in text categorization. It is an efficient algorithm that often finds the global minimum of the objective function where many other classifiers find local optimum solutions. It has been shown that SVM works very well with high dimensional data and avoids the problem known as *the curse of dimensionality* [Tan, 2006]. In machine learning, a dimension simply is a feature in a feature space, and the curse of dimensionality is about finding the important features amongst hundreds or maybe thousands of features.

Support Vector Machines use *vectors* to represent the data. *Vector* is a term that origins from geometry. A vector is a quantity that involves magnitude (size or length) and direction. Vectors are usually identified by an arrow, and a vector, $v = \overrightarrow{AB}$ is an arrow with the tail in point A and the head in point B. Every feature of a vector can be regarded as a dimension, $\overrightarrow{AB}$ is thus two-dimensional and $\overrightarrow{ABCD}$ is a four-dimensional vector [Adams, 1999]. Using vectors to represent data is popular because of the efficient algebraic operations defined for vectors. An example of a popular operation in classification and clustering, is the similarity or the distance between two vectors. Any data record that consists of features can be represented as a vector and thereby data records of any kind can be compared with the operations available for vectors.

SVM uses several of the defined algebraic operations available for vectors. For example, calculating the similarity between two vectors - the *dot product*. The dot product is the sum of the products of two vector's corresponding components. If there are two vectors $\overrightarrow{1,3}$ and $\overrightarrow{2,4}$, then the dot product is calculated as:
$(1,3) \cdot (2,4) = 1*2+3*4 = 14$

SVM classifies data using the notion of *Maximum Margin Hyperplanes* (MMH) to induce a model and for classification [Tan, 2006]. Hyperplane is a geometric definition. In one-dimension, a hyperplane is a point that divides a line into two rays. In two-dimensions a hyperplane is a line dividing the area in two, and so on. MMH is the hyperplane in the data-set that separates the classified data points most correct, and is at equal distance from the classified sets, thus maximized. When the data is

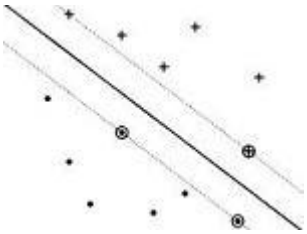

Figure 3.1: Linear Maximum Margin Hyperplane

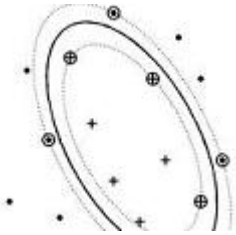

Figure 3.2: Nonlinear Maximum Margin Hyperplane

separable into their respective class by a straight line, the hyperplane is called *linear* (see Figure 3.1).

When it is not possible to separate the data into their respective class by a linear decision boundary, the hyperplane is called *nonlinear*. To be able to find a hyperplane in a data-set like in Figure 3.2, the data is transformed from its original coordinate space into a new space where a linear decision boundary can be used to separate the data. This is a computationally expensive task that involves solving the *Quadratic Programming Optimization problem* [Platt, 1998] and calculating the dot product between pairs of vectors in the transformed space.

Calculating the dot product is a possibly computationally expensive task because the high dimensional space may even be infinite-dimensional. The solution is a method called the *kernel trick*. The *kernel trick* transforms algorithms that use dot products, by replacing these with the *Kernel function*. The *Kernel function* is a function that measures similarity and is computed in the original attribute space [Tan, 2006]. There are many different *kernel functions* that can be used in SVMs, for example, *linear*, *Polynomial* and *Sigmoid*. The most used is the *Radial Basis Function* (RBF) because it generally provides higher accuracy than the other functions [Chang and Lin, 2001].

LIBSVM is an integrated software for Support Vector Machines (SVM) classification. The software includes everything that is needed for parameter selection, model training and classification [Chang and Lin, 2001]. *LIBSVM* has efficient solutions to multi-class classifications and an advanced module for cross-validation for model selection. The tool uses a well-known effective algorithm for solving the optimization problem, the *Sequential Minimal Optimization* algorithm [Platt, 1998], known as *Quadratic Programming* during training of SVM.

3.3 Feature Selection

The most basic feature to use is the relative frequencies of words, unigrams, in the text. The use of unigrams in classification into readability levels have been studied by Collins-Thompson and Callan [2004]. The features used here are well motivated and previously used in readability formulas (described in section 2.2.1) or in other research as tools in text-analysis. Many of the features used in this study and in text analysis in general origin from the study performed by [Biber, 1988]. Not all the features are claimed by previous research to be influential but chosen anyway to verify that. The features might be measurements of the exact same phenomena but with different methods. For example *sentence-length* and *syntactic depth* (described in 1 and 2 in the list below) are used as measurements of the same phenomenon; complexity of the sentence. Average sentence-length is used as a measurement of syntactic complexity in traditional formulas and average syntactic depth also describes the complexity of a sentence. The features that best models each class are assigned the highest importance by the SVM when training the classifier.

Below are descriptions of the features extracted from each text, together with a short motivation of why the feature was chosen.

1. Syntactic depth: the maximum depth of every sentence, calculated as an average throughout all sentences in the text. Complex sentences are less readable than simple sentences and the proportion of complex sentences in a text is often used as a feature in traditional readability formulas [Klare, 1963].

2. Sentence-length: the average number of tokens per sentence. Sentence-length is a good indicator of the complexity of a text Melin and Lange [1995].

3. Prepositional phrases: the average number of prepositional phrases per sentence. The *pp-attachment* problem is well-known to give rise to ambiguity. Many of the traditional formulas use a measurement of how many prepositions there are in a text. An ambiguous text is less readable because of the workload to disambiguate it.

4. Subordinating conjunctions: the average number of subordinating conjunctions per sentence. Every Swedish clause that starts with a subordinated conjunction are subordinated clauses. Subordinate clauses per sentence are good indicators of readability, especially if the subordinated clause is nested and negations are used [Björnsson, 1971]. This is illustrated by a typical Swedish sentence [2]:
   `...inte roligt [att du inte berättade [att du inte kunde komma]].`

5. Difficult words: the number of words containing more than six letters, calculated as an average number of difficult words per sentence. The measurement is used in *LIX* (described in section 2.2.1). This parameter reflects readability by being a reflection of unusual words [Backman, 1976]. Biber [1988] uses the proportion of long words as a measurement of how specific the texts are.

6. Vowels: the average number of vowels per word. This represents the number of syllables per word. The number of syllables per word is often used in readability formulas as a measurement of morphological complexity [Klare, 1963],

[2] ...not funny that you did not tell me that you could not come

where more syllables makes the words more complex. The number of vowels and number of syllables is not a one to one relationship, but gives a rough estimation of the number of syllables.

7. Nominal quotient (NQ): the number of nouns, prepositions and participle divided by the number of pronouns, adverbs and verbs, calculated as a measurement per document. This is a measurement of how much information there are in a text. The normal NQ value is 1.0 [Melin and Lange, 1995]. The amount of information in a text has an influence on the readability of the text, a less informative text is more readable than a highly informative.

   $$NQ = \frac{Nouns + prepositions + participle}{pronouns + adverbs + verbs}$$

8. Noun/Pronoun quotient: the number of nouns divided by the number of pronouns in the text, calculated as a measurement per document. This also gives a measurement of the amount of information in the text since nouns are a part-of-speech with high information value and pronouns often repeat previous information [Melin and Lange, 1995].

9. Attributes: the number of attributes on the left-hand side of a nounphrase, calculated as the number of attributes per noun phrase. A large number of attributes decreases the readability of the sentence. A head noun in a noun phrase can have a number of different attributes. For example, quantifiers, genitive attributes, and adjectives can act as attributes. Attributes make the sentences long and the reader has to keep the information in mind until the phrase ends. The text gets more detailed with many attributes.

10. Phrase-length: the average number of constituents per phrase gives a measurement of how general or detailed the text is [Melin and Lange, 1995]. Short phrases indicate undetailed descriptions and generality which could be a good indicator of the readability of the text. It is difficult to know in which way the generality measure affects readability. If something advanced and technical is described in general terms it could decrease readability, while in other cases it might increase readability.

11. Definite articles: the number of definite articles per sentence. This provides a measurement of how abstract the text is since abstract texts have less definite nouns and articles. In Klare [1963], a formula for measuring abstraction of a text by P.J Gillie is presented and the core feature of the formula is the amount of definite articles.

3.4 Tools for feature selection

The different tools used for preprocessing the data and extracting the features necessary to model the readability levels, are described below along with a description of why they are used.

3.4.1 Trigrams'nTags

Trigrams'nTags (TnT) [Brants, 2000] is an efficient statistical part-of-speech (PoS) tagger based on Hidden Markov Models. It is trainable on many different languages and different tag sets. It is an implementation of the Viterbi Algorithm for second order Markov models based on trigrams. The probabilities for a tag are based on the relative frequency of trigrams in training data. The TnT is robust and provides every word with a part-of-speech tag [Brants, 2000]. It is a fast implementation, training speed is around 100.000 tokens/second. According to Megyesi [2002], TnT has the highest overall accuracy when tagging both known and unknown words for Swedish, when trained on Stockholm Umeå Corpus, compared to other freely available taggers, that is the major reason for using this PoS-tagger. TnT is free of charge for non-commercial research purposes.

3.4.2 SPARKchunk

SPARKchunk [Megyesi, 2002] is a parser based on a context-free-grammar for Swedish, where the parser is an Earley parser implemented in python. SPARKchunk parses PoS-tagged data into a number of different phrase types that represent the hierarchical structure of a sentence. Some of the categories are not phrases in a classic sense, such as the category Verb Cluster described below, but more like chunks with a content word acting as head of the chunk, and a number of function words matching a fixed template. Here is a brief description of the phrases used in SPARKchunk [Megyesi, 2002]:

- Adverb Phrase (ADVP): adverbs that modify either adjectives or numerical expressions.
- Minimal Adjective Phrase (AP): the adjectival head, possibly with modifiers.
- Maximal Adjective Phrase (APMAX): more than one AP separated with a delimiter or a conjunction.
- Numeral Expression (NUM): numerals, possibly with modifiers.
- Noun Phrase (NP): head noun and its modifiers to the left.
- Prepositional Phrase (PP): one or more prepositions delimited by a conjunction and one or more NP/NPMAX.
- Verb Cluster (VC): a group of continuous verbs belonging to the same verb phrase without constituents in between.
- Infinitive Phrase (INFP): an infinitive verb together with the infinite particle, possibly with ADVP and/or verbal particles.

SPARKchunk is easy to use and robust. It is free of charge for non-commercial use.

4 Evaluation

4.1 Method

In this thesis, the problem of classification into readability levels is considered as a classification problem. For such problems there are well-established evaluation methods.

The corpus was divided into two non intersecting data-sets; a training set and a test-set. The test-set consists of 20 percent of the scaled corpus (described in section 3.1), where each class is represented by the same number of documents. The remaining 80 percent was used as training set. This renders a test-set containing 249 documents and a training set containing 1005 documents.

The LIBSVM package (see section 3.2) has a tool for automatically selecting optimized parameters for the algorithm. The tool was used on the training set to get the best parameters for each combination of features. A 40-fold cross validation was performed on the training data to validate the feature combinations. The 40-folded cross validation verifies the results and reduces the risk that a model performs well by chance.

Three models are picked for the final evaluation on the separate test-set. The selection of the models which are further evaluated on the test-set, is solely based on the performance of the model during the validation. When using the classification model as a part of a search engine, the most important is that the model select all the relevant documents and select some irrelevant documents rather than select only correct documents and miss some of the relevant document. Thus all decisions are based on the recall scores during the validation.

The final results are presented as an overall recall, precision and F-score, along with the same measurements per readability level. The precision score is defined as, *the fraction of documents that actually turns out to be correct in the group of documents that the model has declared as a class*. The precision score reveal how reliable the model is when classifying a document to a certain class. The recall score is defined as, *the fraction of documents correctly predicted by the model compared to what actually should be detected*. The recall score measure how many documents the system detects. The combination of the precision and recall score into one measurement is called *F-score*. The F-score represents the harmonic mean between the recall and the precision. The F-score is calculated as:
$2 * Precision * Recall / (Recall + Precision)$

4.2 Results

4.2.1 Baseline

There are several ways to define and calculate a baseline; the simplest feature (unigrams), the best single predictor of readability or the most used feature in traditional readability formulas.

In this study there were to obvious candidates, the first candidate feature, average sentence-length of each document, is the most frequent measurement of traditional readability formulas. This means that sentence-length has been established by previous research to be a good measurement of readability. In the listing of the most famous formulas made in Klare [1963], more than 50% of the formulas use average sentence-length as a component.

The other candidate is the feature that by itself models the data best. With that feature as baseline, one can be sure of having a strict comparison during the evaluation of the model. In Table 4.1, the results from classification using each feature by

Feature	Total	Easy	Medium	Difficult
Syntactic depth	61.04	81.92	32.50	68.67
Sentence-length	58.02	**93.97**	2.40	79.49
Prepositional phrases	63.45	86.75	28.92	74.70
Subjunctions	52.61	62.65	59.03	36.14
Difficult words	62.25	81.92	28.91	75.90
Vowels	62.25	43.37	72.29	71.08
NQ	**69.73**	55.41	69.67	**84.12**
Nouns/Pron	69.44	60.24	**74.70**	73.39
NP-attr.	61.04	43.37	72.28	67.47
Phrase-length	59.04	46.99	54.21	75.90
Definite articles	49.40	73.49	53.01	21.69

Table 4.1: The overall recall of each feature used by itself along with the recall for each readability level.

itself is shown. The table also shows how good each feature is to model the different readability levels. As can be seen in the table, the feature that best modeled the data by itself was Nominal quotient (NQ) (described in section 3.3). It is notable that all classification models induced from only one single feature performed on a level significantly higher than guessing, since guessing would render a recall score around 33%. It is also notable to see that there are such differences between the readability levels, sentence-length has a recall of 94% when classifying the Easy-level but only 2.4% when classifying the Medium-level.

Since NQ outperformed the sentence-length measurement by as much as 11.71 percentage points, it was selected to be the baseline. Sentence-length is disregarded despite that it is widely used in previous research.

4.2.2 Validation

To find the feature combinations that model the data best by empirical testing, means that the ideal would be to try all possible feature combinations. Even though the number of features to choose from is relatively small (the number of features are 11),

it is impossible to try more than a few of the thousands of combinations because of the time consuming process of inducing and validating the models.

The strategy was to use the information about the single feature classification ratio. The hypothesis was that the best overall classification models were to be found by combining the features that model the readability levels best and the features with the best overall recall score. A number of feature combinations were selected to be validated in an initial validation. The purpose was to get an indication of, for example, the amount of features to use. The results from the initial 40-folded cross validation are shown in Table 4.2.

Motivation	Feature combination	Recall
All features	All	**88.26**
10 best single features	All but definite articles	87.86
9 best single features	All but definite articles and subjunctions	86.96
7 best single features	NQ, Noun/Pron, PP, Vowels, Diff.words, Synt.depth, NP-attr.	86.87
8 best single features	All but definite articles, subjunctions and sent-len	86.57
6 best single features	NQ, Noun/Pron, PP, Vowels, Diff.words, Synt.depth	85.37
2 best/readability level	NQ, Sentence-length, PP, Vowels, NP-attr.	85.37
The best/readability level	NQ, Sentence-length, Noun/Pron	83.68
3 best single features	NQ, Noun/Pron, PP	82.69
2 best single features	NQ, Noun/Pron	69.85

Table 4.2: The results from the initial 40-folded cross validation.

With the results from the initial validation, some conclusions could be made. More features in the combination increases the recall of the model in all the cases in the initial validation but one; when increasing the number of features from 7 to 8, the recall decreases 0.3 percentage points. With that information the continued validation only considered models that were induced from as many features as possible, which is 10 or all features. All 10-feature combinations were validated in a 40-folded cross validation. The results of that validation can be seen in Table 4.3. The difference between the best performing and the worst performing model is 2.29 percentage points.

The only model from the initial validation that performs at a level comparable to the five best 10-feature combination models are the model induced from all features.

Feature combination	Recall
All – Phrase-length	**88.76**
All – NQ	88.06
All – PP	88.06
All – Definite articles	87.86
All – Difficult words	87.86
All – Nouns/Pron	87.86
All – Syntactic-depth	87.56
All – Vowels	87.36
All – Sentence-length	87.26
All – NP-attributes	86.57
All – Subjunctions	86.47

Table 4.3: The results of the models induced from 10 features.

The model induced from all features has the second highest recall.

No model induced from just a few features performed at a level high enough to be further evaluated. The initial plan was to pick three models for the final evaluation, but since the third place is tied between two models, both of them are selected.

The models that were selected for the final evaluation are:

- The model induced from all features but the phrase-length feature, this model is called the notPhrase-model. 88.76% recall.
- The model induced from all features, this model is called All-model. 88.26% recall.
- The model induced from all features but the prepositional phrases feature, this model is called notPP-model. 88.06% recall.
- The model induced from all features but the noun quotient feature, this model is called notNQ-model. 88.06% recall.

4.2.3 Test

For the final evaluation of the models, the test-set that was created and removed from the rest of the data, described in section 3.1.2, was used. The classification models used for testing are induced from the entire training data-set.

The results from the classification of the previously unseen data are presented in Table 4.4. To interpret the test results correctly, the size of the test-set has to be taken into consideration. The upper and lower limits of the 95% confidence interval show the reliability of the classification, based on the size of the data-set. This means that when classifying a previously unseen data-set of this size the model will, by 95% confidence, always perform within those limits. The upper limit of the confidence interval for the notPP-model is 92.78, and the lower limit is 85.14 (calculated using the Wilson procedure).

Model	Precision	Recall	F-score
notPP-model	**90.21**	**89.56**	**89.88**
notPhrase-model	88.93	88.35	88.64
All-model	88.90	88.35	88.62
notNQ-model	88.55	87.95	88.25
Baseline	69.84	69.88	69.86

Table 4.4: The results from the classification shown as overall recall, precision and F-score.

Three of the models performed better than they did during the validation, and the only model that performed worse is the notNQ-model. The best performing model was no longer the notPhrase-model, which was the best during the validation. The model that achieved the highest recall on the test-set was the notPP-model. There were no significant differences between the results from the different models, 1.63 percentage points between the best and the worst F-score in the test. At the .05 significance level the differences between the results from the evaluated models are not statistically significant. In fact, there is a 45% risk to get the same difference by chance (calculated using McNemar's test), when comparing the best and the worst performing model.

To see the performance of the models when classifying the different readability levels, the precision and recall score were calculated for each level. The scores are shown in Table 4.5.

Model	Easy			Medium			Difficult		
	P	R	F	P	R	F	P	R	F
notPP-model	**82.47**	**96.39**	**88.89**	**93.42**	**85.54**	**89.30**	**94.74**	86.75	**90.57**
notPhrase-model	**82.47**	**96.39**	**88.89**	93.06	80.72	86.45	91.25	87.95	89.57
All-model	82.29	95.18	88.27	93.06	80.72	86.45	91.36	**89.16**	90.25
notNQ-model	81.44	95.18	87.77	92.10	84.34	88.05	92.10	84.34	88.05
Baseline	57.50	55.42	56.44	66.67	69.88	68.24	85.36	84.34	84.85

Table 4.5: The precision (P), recall (R) and F-score (F) for every readability level.

5 Discussion

The total F-score of the baseline is 69.86, to be compared to the F-score achieved during the final evaluation of the notPP-model, 89.88. The notPP-model has improved the baseline with as much as 19.75 percentage points or 27%. At the .01 significance level the difference between the results from the baseline and the PP-model is statistically significant calculated using a McNemar's test. The notPP-model gets a higher F-score on every readability level. The readability level that the baseline classifies the best is the Difficult-level. The F-score of the baseline for the Difficult-level is 84.85, to compare with the F-score of the notPP-model, 90.57.

Even though the most accurate classification is achieved by the notPP-model, the big risk that the differences between the models occurred by chance, makes the selection of the notPP-model as the best model unreliable. To select the best model, the validation has to be taken into consideration. During the validation, the All-model achieved the second highest recall, together with a good F-score on the test-set, it might make the All-model the best and most reliable model. Since the differences are so small, any of the best models can be used depending on the task of the model. The precision and recall scores for each readability level have to be considered. If the task of the model is to separate for example, the documents belonging to the Medium level, then the All-model is a bad selection since both the notPP-model and the notNQ-model achieve a higher F-score for that level. So when designing a model for a specific purpose there is a need for a full analysis to get the feature combination that best suits the purpose.

The results from each readability level show the same tendency for all models. When a document is classified into either the Difficult- or into the Medium-level, it is a correct classification in more than 90% of the cases. When the document is assigned to the Easy-level there is a bigger chance that it is a misclassification compared to when the document is assigned to the Difficult- or Medium-level. The best model finds more than 96% of the documents belonging to the Easy-level. Most misclassified documents belong to the Difficult- and the Medium-level, and they are classified as belonging to the Easy-level.

The Medium-level was considered as the less homogenous readability level in the project, but the results from the classification, made by the PP-model, show that the F-score for the Medium-level is the second best. Despite the varying writer skills of the students, the texts obviously have much in common by terms of the features that the classification is based on.

The fact that the All-model with a total F-score of 88.62 performs that high means that all the features in this study, has been considered by the algorithm to be good predictors of the readability levels. The same conclusion can be made when the results from the single feature classification is considered. The classification shows that all models induced from one feature perform at a level significantly higher than guess-

ing, if we assume that guessing would assign a third of the documents to the correct readability level.

The results from the single feature classification, show that NQ is the best single predictor and PP the is the third best single predictor. Despite of that, two of the best models, notNQ-model and notPP-model, are not using those features in the classification. This indicates that finding good combinations of features are as important as finding good features. This verifies the advantage of data-driven technique of establishing readability over a traditional readability formula. The features of traditional readability formulas are selected from how well each feature represent the texts, but feature combinations are never or seldom evaluated because of the workload to manually analyze it.

The results from previous research that is similar to this [Collins-Thompson and Callan, 2004] (presented in section 2.2.2), is difficult to compare with the results presented here. They provide two recall scores from 10-fold cross-validation, a recall score of 64%, obtained on a corpus assembled from the web with the grade levels 1-6, and a recall score of 79% on a corpus assembled from the web with the grade levels 1-12. The test-set is derived from another corpus than the corpus that the model is induced from, which makes the comparison uneven. Since the results that can be compared, only concern validation and the number of readability levels used in the previous study differ greatly from the number of readability levels used here, the results from this study are only compared to the baseline.

6 Conclusions

The purpose of the thesis was to create models for classification of texts into readability levels for Swedish. The problem has been solved by training a Support Vector Machine classifier on features that previously has been shown to represent readability. The features are derived from data that represent three different readability levels, Easy, Medium and Difficult. The results from the classification have been evaluated using standard techniques for evaluation of text classification, namely precision, recall and F-score.

This thesis has shown that a high quality classification into readability levels based on a small amount of data is possible. The usage of existing solutions to machine learning, PoS-tagging and parsing together with previously established measurements of readability has rendered a new solution for measuring the readability. The results from the classification are promising, almost 90% of the documents are classified and found.

The following are by-products that come with the creation of the classification model;
A number of features that predict readability have been established and summarized. These features provide a stable ground for further research in classification into readability levels, for Swedish and probably for other languages as well, since many of the features are based on research for English. Tools for the extraction of these features have been developed and can be used as a tool-kit. The base for a corpus where the texts are annotated with labels for reading difficulty on three levels has been created.

6.1 Future work

There are many interesting steps to take in the future, a few of them are listed below.

The use of the results of this study is mainly within information retrieval, for use in applications like search engines. A model for classification into readability levels could be used to classify the indexed documents of a search engine, where every document would be assigned to belong to a readability level. There are several ways to handle the information about the readability levels assigned to the documents;

Give the users a possibility to select the readability levels of the requested documents. This is probably the easiest and yet most powerful usage.

The readability levels could be given an importance in the ranking of the retrieved documents. This requires an analysis of what kind of readability levels that are the most requested in the search engine.

A separate model for classification into readability levels based on the users' query could be used to match the readability level of the user query to the appropriate documents [Liu and Oh, 2004]

A natural extension of this study would be to include more readability levels in the models and perform an evaluation of the models with more readability levels. For implementation of the model as a part of a search engine it is important that the model can handle more than the three readability levels that are defined at the moment. Another important step would be to perform a user evaluation of the classified documents, to see the correlation between human estimation of readability and the classification model.

The effect of the model used in a search engine is crucial for future research in the subject, since search engines are the biggest field of application for a model like this. A readability level classifier can be used for the traditional tasks of a readability formula, like helping school teachers selecting reading material or helping writers to adjust the texts readability to the appropriate level. It is not possible for every teacher or every writer to induce and tune their own model, but a solution would be to induce and tune a classification model at a central website, for example, a classification model for helping teachers to select reading material could be located at the web site of the Board of Education.

To further improve the existing features and to be able to find new features for classification, a dependency analysis of the constituents in the texts would be an important part. To be able to see the relations between words and which words or phrases that are connected, would give plenty of useful information. The function of words along with their positions in the sentence, e.g. the placing of the subjects and objects, would also provide useful information.

Within this study there has been no time for tuning the SVM, there are many options, including the use of different kernel-functions, available to further improve the classification. A thorough analysis of the results from the classification is needed, mainly an analysis of the documents that are misclassified by the model. There has been no time in this project for analyzing the reasons of a misclassification. By looking at the decision values and the hyperplanes used by the SVM, together with an analysis of the text in the misclassified documents, the model might be further improved.

A comparison of the classification from this classification model to the reading difficulty classification from the famous Swedish readability formula LIX (described in section 2.2.1) would be interesting, since LIX is the most well-known and thus most used way to estimate the readability of texts.

Bibliography

Robert A. Adams. *Calculus*. Addison-Weasly Longman Ltd, 4 edition, 1999. ISBN 0-201-39607-6.

Jarl Backman. *LIX - Mätning av läsbarhet eller trycksvärta?* Universitetet och lärarhögskolan i Umeå, 1976.

Douglas Biber. *Variation across speech and writing.* Cambridge University Press, New York, 1988. ISBN 0-521-42556-5.

Carl-Hugo Björnsson. *Läsbarhet*. GEC GAD, Köpenhamn, 1971. ISBN 8712739081.

Thorsten Brants. TnT – a Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000)*, Seattle, WA, 2000.

Magnus Cedergren. Kvantitativa läsbarhetsanalyser som metod för datorstödd granskning. Technical report, Inst. för Numerisk Analys och Datologi, KTH, Stockholm, 1992.

Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: A Library for Support Vector Machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Kevyn Collins-Thompson and Jamie Callan. A language Modeling Approach to Predicting Reading Difficulty. *Proceedings of HLT / NAACL 2004, Boston, USA, May 2004*, May 2004.

Bengt Dahlqvist. The SCARRIE Swedish Newspaper Corpus. In *Working Papers in Computational Linguistics & Language Engineering*, 1999.

Dale and Chall. A Formula for Predicting Readability. *Educational Research Bulletin*, 27:37–53, 1948.

George R. Klare. *The Measurement of Readability*. The Iowa State University Press, 1963.

Croft Liu and Hart Oh. Automatic Recognition of Reading Levels from User queries. In *Annual ACM Conference on Research and Development in Information Retrieval*, number 27, pages 548–549, 2004.

Beata Megyesi. *Data-Driven Syntactic Analysis - Methods and Applications for Swedish.* PhD thesis, Department of Speech, Music and Hearing, KTH, Stockholm, 2002.

Lars Melin and Sven Lange. *Att analysera text*. Studentlitteratur, Lund, 1995. ISBN 91-44-23582-8.

John C. Platt. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. *Technical Report MSR-TR-98-14*, April 1998.

Kumar Tan, Steinbach. *Introduction to Data mining*. Pearson Education, Inc, 1 edition, 2006. ISBN 0-321-32136-7.

Vladimir Vapnik. *Statistical Learning Theory.* Wiley-Interscience, New York, 1998.

A Examples of the data-sets

A.1 SCARRIE data

Waitangifördraget följdes aldrig!

Historien börjar för drygt tusen år sedan. I söder ligger det jungfruliga landet, skapat av vulkaner och hav. Söderhavet sanna vikingar, båda krigsstammar från Polynesien, siktar från sina stora havsgående segelkanoter Aotearoa, “det långa vita molnet”, som är maoriernas namn på det vi i dag kallar Nya Zeeland.

Ett land utan ormar eller andra farliga djur men fyllt med stora vinglösa moafåglar, med orädda sälar och annat lättfångat byte i en välkomnande subtropisk natur. Från början levde nybyggarna som jagande nomader på en nivå som närmast motsvarar vår stenålder, okunniga om hjulet men också om pil och båge. Så småningom blev stammarna bofasta och började bruka jorden. Alltjämt krigiska och i ständig fejd med grannstammarna förskansade man sig i väl befästa byar, oftast belägna högt på en kulle invid havet eller intill någon av de många fiskrika floderna. Som mest tror man att mellan 100 000 och en kvarts miljon maorier befolkade öarna när de första européerna dök upp. Visserligen hade sjöfararna Abe Tasman 1642 och James Cook 1769 tillfälliga kontakter med landet, men först i slutet av 1700-talet kom valfångare och handelsmän att slå sig ner, följda av missionärer. På 1830-talet var invandringen organiserad från Europa, främst från England. Med de nya invandrarna kom gevär, alkohol och sjukdomar, vilket tillsammans med blodiga inbördesstrider reducerade maorifolket till som lägst ca 40 000 individer.

Laglöshet och ständiga bråk om äganderätten till mark rådde dels inbördes mellan européerna, dels med maorierna. Redan 1833 hade James Busy, för att skydda brittisk handel i området, gjort en första överenskommelse om maoriskt oberoende 35 stamhövdingar på Nordön. Men när regelrätta strider ändå utbröt mellan nybyggarna och maorierna svseglade kapten William Hobson från Australien 1837 för att stävja oroligheterna. Han återvände sedan 1840, nu som av drottningen utsedd ställföreträdande guvernör. Med hjälp av lokala missonärer lyckades han samla ett stort antal hödvidingar hemma hos Busby i dennes residens i Waitangi. Hobson framlade förslaget att hövdingarna skulle acceptera den engelska drottningens suveränitet mot att Storbritannien tog landet under sitt beskydd och styrde landet i lag och ordning. (Det är viktigt att notera den brittiska regeringens villkor: maktövertagandet skulle ske på frivillig basis, genom förhandlingar,mellan parterna.) Oenigheten var stor bland de församlade hövdingarna, man drog sig tillbaka till andra stranden av floden för enskilda överläggningar som pågick hela natten. Nästa dag, 6 februari, undertecknade dock den mäktige hövdingen Hone Heke som första man traktaten och ytterligare 46 följde efter. Med hjälp av missionärer insamlades sedan under året påskrifter från ytterligare 500 hövdingar runt om på Nordön. Waitangifördraget

måste ses som ett för sin tid enastående dokument. I korthet kan sägas att de invandrade européernas jämställdhet och lika rätt. Drottningen och Kronan garanterade maorifolket äganderätten till deras land, skogar, fiskevatten och andra värden de kunde besitta. I gengäld underkastade sig hövdingarna drottningens överhöghet. De blev brittiska undersåtar, med alla de rättigheter och privilegier som tillkom sådnan, mot att Kronan fick ensamrätt på att köpa land av dem. Men dokument är en sak, verkligheten en annan. Dels kom väl avtalet aldrig att hållas till fullo av någondera parten, dels hade maorierna med sin relativt outvecklade samhällsstruktur dåliga förutsättningar att förstå innebörden av fördraget. När sedan européerna visade missnöjde med den nya tidens lag och ordning och försämrade villkor utbröt nya oroligheter. Trots fördragets uttryckliga försäkran om frivillighet tvingades ofta maorier att sälja landområden mot sin vilja och till oskäligt lågt pris. Maorierna tog under Hone Heke till vapen och gjorde uppror på Nordön. Sydön var då alltjämt glest befolkad och hade helt enkelt annekterats av England. Först under 1880-talets sista år kan man säga att fred inträdde på Nya Zeeland. Men som framgång av nyhetssändningarna har Waitangifördraget alltjämt hög politisk aktualitet. Det åberopas som en naturfolkets självklara rätt till landets rikedomar, oavsett hur de uppkommit. Likheten med samernas och de nordamerikanska indianernas krav är tydliga. Skillnaden är att maorierna har papper på sin rätt.

A.2 www.sesam.nu data

Försvaret gör reklam för att tjejer ska göra lumpen. Reklamen säger att lumpen är bättre än att dreja, åka på ridläger eller bli au pair. När man är au pair, arbetar man en tid hos en familj i ett annat land. Man kan studera språket i landet och lära sig hur man lever där. Många unga kvinnor vill bli au pair. Kan man jämföra lumpen och au pair? Kapten Staffan Slörner är ma rknadsansvarig på försvaret. Han säger:- När unga tjejer är au pair eller gör lumpen, är det ofta första gången de kommer hem ifrån. De får prova att klara sig själva.Staffan Slörner säger att det finns mycket humor i reklamkampanjen. Men han säger at t det är allvar också.- Det är coolt att göra lumpen. Det är fånigt att vara au pair. Om jag hade döttrar, skulle jag hellre vilja att de gjorde lumpen.Mårten Hedlund jobbar på reklamfirman Saatchi och Saatchi. Han var med och gjorde försvarets rekla mkampanj. Mårten Hedlund säger:- Reklam går fort. Folk ska se den. Man kan inte ge information i början. Vi vill bara säga at t tjejerna ska göra något kul. Sedan ska de gå in på lumpens hemsida.Mårten Hedlund och Staffan Slörner talar om att “göra nå got kul” och att det är “coolt” att göra lumpen. Ska tjejerna göra lumpen för att det är coolt?Åsa Carlman på Svenska Freds t ycker inte det. Hon säger:- Lumpen är till för att man ska bli militär. Försvaret måste vara desperata när de gör en sådan ka mpanj.Förut hade Sverige ett försvar mot angrepp från andra länder. Nu har vi en annan typ av försvar. Försvaret kallar det f ör insatsförsvar. Det ska försvara landet mot terror och hjälpa till i andra länder tillsammans med EU och FN.Alla killar mås te mönstra, alltså gå på en undersökning för att se om de ska göra lumpen. Men tjejer som vill mönstra måste ansöka.Lumpen är tre terminer. Alla som gör lumpen måste göra minst två terminer. För tjejer är lumpen frivillig, men när en tjej ha valt att göra lumpen måste hon ändå göra två terminer. Den tredje terminen handlar om att arbeta utomlands, i FN-styrkorna till exemp el.Ska man säga till tjejer att göra lumpen för att det är coolt? Staffan Slörner tycker att det finns en allvarligare mening med att göra lumpen. Men han tycker inte

att man ska säga det i reklamen.- Man får upptäcka det när man har valt att göra lu mpen.Stefan Sandborg är major i Livgardet vid Kungsängen. Han tycker att man ska tänka på om man vill göra internationell tjä nst, innan man gör lumpen.- Internationell tjänst bidrar till fred och säkerhet i världen. Det är också meningen med att göra lumpen.Han tycker ändå att reklamkampanjen är okej. Han tycker att den borde ha haft ett djupare budskap om den hade varit r iktad till lite äldre personer. Men den passar för unga tjejer, säger han.- Det borde funka. De kan söka mer information själ va.Försvaret vill att 40 procent av alla som gör lumpen ska vara tjejer. Nu är 5-6 procent av alla som gör lumpen tjejer. Sta ffan Slörner berättar om hur de tänkte när de gjorde reklam-kampanjen. De skulle inte säga något dåligt om kvinnor. Ingenting skulle handla om sex eller religion.Men reklamen säger att det är dåligt att bli au pair och bra att göra lumpen, som mest k illar gör. Man kan tycka att den säger något dåligt om tjejer.Reklamkampanjen har fått ett pris av tidningen Resumé. Juryn va r personer från olika reklam-firmor. Men de tyckte olika om kampanjen. Josephine Wallin från reklambyrån King säger:- Om jag gick i skolan och såg reklamen, skulle jag tänka att jag hellre ville göra lumpen än bli au pair.Malin von Werder från reklam byrån Garbergs tycker att kampanjen säger att allt som tjejer brukar göra är fånigt. Hon tycker att det är hemskt och sorglig t.- Om man tycker om att dreja, då? Är det sämre än att spränga broar?

A.3 High school exams

För Barnboken!

Redan som liten började mina föräldrar att läsa böcker för mig. Jag växte upp med barnsagor, och det var en självklarhet med godnattsagor innan jag skulle sova. Resultatet av läsningen visade sig snabbt, och sedan var även jag en inbiten läsare. Barnböckerna jag läste var lätta och ofta mycket bilder i. Allt för att intressera ett barn som hellre springer ut och leker i sandlådan. Efter ett antal böcker fyllda av skratt och kanske lite snyft, började jag att avancera i lässvårighet. Nu gick jag över till deckare och kriminalare, vilket visade sig vara positivt för min läsning. Nu tuggade jag böcker som aldrig förr, och när jag väl hade kommit in i boken var jag okontaktbar. Jag började få ett bra grepp på det Svenska språket och dess formuleringar. Sedan började jag skolan med allt vad det innebar. Läxläsning, studera till prov och skaffa mig ett socialt umgänge. Det gick bra! Jag hade lätt för att skaffa kompisar och att kommunicera med andra människor. Även om jag som alla andra barn var ett litet busfrö, kunde jag koncentrera mig på lektioner och hemuppgifter. Jag började också att intressera mig för sport, och på fritiden sysslade jag med flera olika idrotter kombinerat med läsning och läxor. Hela uppväxten underlättades av min kärlek till böcker.

Nu efter ett antal års erfarenhet av läsning kände jag att det börjar bli ett aktuellt ämne att skriva en debattartikel om. Efter att ha läst skolverkets häfte “Läge för läsning” samt en artikel om “Barns läsning” av Leni Filipson, Barnbarometern 98/99-1999 blev jag uppriktigt sagt orolig. I artikeln har Leni framfört ett diagram som klart visar den nedåtgående trenden av barns läsning. Hon skriver att “Andelen bokläsare bland de yngsta minskade kontinuerligt under 80-talet, stabiliserades sedan under första hälften av 90-talet men har nu åter minskat”. Hon framställer även en teori om att “Minskningen i andel “läsare” är nu mest påtaglig hos de allra yngsta

och bland de barn med högutbildade föräldrar” Är det verkligen så att dagens industrisamhälle har framhävt en sådan stress, att högutbildade personer inte längre har tid för barnen och deras läsning? Ja uppenbarligen är det så! (Vi ser också en nedåtgående trend vad gäller barnafödsel. Detta är ett stort problem som vi måste göra något åt) Föräldrar måste ta ansvar för att deras barn får den läsning som de behöver. Hela barnens personliga utveckling är uppbyggd kring läsandet. När ett barn läser en bok så utvecklas deras språk på ett mycket positivt sätt. De får andvända sin fantasi, bygga upp egna tankegångar samt tänka strategiskt. Barnen får även en inblick i det verkliga livet gällande etik och moral. Ett barns utveckling är väldigt viktig redan vid ett tidigt stadium i deras liv. Deras framtida kompetens och allmänbildning kommer att bli hämmnad, om de inte blir uppmanade av föräldrarna att läsa böcker. Man behöver inte lägga ribban på en alltför svår nivå. Det räcker med enkla bildeböcker för att skapa stimulans. Jag tror att den all kriminalitet som finns i samhället idag bygger mycket på okunskap. Man har haft en dålig uppväxt. Föräldrarna kanske struntade i en. Böcker var inte alls något som man intresserade sig för, utan man började att “hänga” på stan. Sedan kom man i dåligt umgänge vilket resulterade i kriminalitet. Tänk istället att de här barnen hade intresserade föräldrar som uppmanade dom att läsa böcker. Då tror jag inte att de hade hamnat så snett i samhället. Utan de hade fått kunskap och kunnat reflektera böcker och sin fantasi på ett helt annat sätt. Jag förväntar mig inte att alla barn kommer att läsa böcker nu direkt. Utan man får nu ha ett mer långsiktligt mål. En bra början kan vara att ta upp detta ämne i TV samt att debattera ämnet i Tidningar. Politiker har också en stor del i det här ämnet. De måste på något sätt uppmärksamma ämnet i kommande kampanjer och valrörelser, och göra alla medvetna om detta problem. Sverige är som bekant inte det enda typexemplet, utan det finns många andra länder som måste framföra betydelsen av läsning och språkutveckling. Om vi ska fortsätta att utvecklas i samma raska takt som samhället gör idag, är det viktigt att barnen inte glömms bort. Det är faktiskt barnen som är vår kommande generation, och om de ska fortsätta utvecklingen och saknar underlag i Svenska språket, kommer detta problem att bli ofantligt stort. Klassklyftorna kommer att öka mellan hög och lågutbildade och vi kommer att få ett samhälle som är ännu mer dominerat av kapitalisterna. Även språkutvecklingen går framåt, och nuförtiden talar man nästan en helt annan Svenska än vad man gjorde förr. Det har till och med gått så långt att man i vissa invandrartäta områden talar annan Svenska. Ta Rinkeby, där pratar man något som de kallar Rinkebysvenska. Behåll det Svenska språket och dess auktoritet. En bra början kan vara att läsa mycket böcker som liten? Vem minns inte “Sune” och “Bert” böckerna, som man så ofta kämpade sig igenom som liten. Jag tyckte detta var ett viktigt ämne att ta upp med tanke på den kommande Världsboksdagen. Det är en viktig dag som borde uppmärksammas mer runt hela världen. Allt för att skapa mer intresse runt läsning och dess nödvändiga utveckling. Tack för ordet!