Features indicating readability in Swedish text

Johan Falkenjack[1], *Katarina Heimann Mühlenbock*[2], *Arne Jönsson*[3]

(1) Department of Information and Computer Science, Linköping University, Linköping, Sweden
(2) Språkbanken, University of Gothenburg, Gothenburg
(3) SICS East Swedish ICT AB

johan.falkenjack@liu.se, katarina.heimann.muhlenbock@gu.se, arne.jonsson@liu.se

ABSTRACT

Studies have shown that modern methods of readability assessment, using automated linguistic analysis and machine learning (ML), is a viable road forward for readability classification and ranking. In this paper we present a study of different levels of analysis and a large number of features and how they affect an ML-system's accuracy when it comes to readability assessment. We test a large number of features proposed for different languages (mainly English) and evaluate their usefulness for readability assessment for Swedish as well as comparing their performance to that of established metrics. We find that the best performing features are language models based on part-of-speech and dependency type.

KEYWORDS: Readability assessment, Machine learning, Dependency parsing, Weka.

1 Introduction

The problem of readability assessment is the problem of mapping from a text to some unit representing the text's degree of readability. Measures of readability are mostly used to inform a reader how difficult a text is to read, either to give them a hint that they may try to find an easier to read text on the same topic or simply to inform them that a text may take some time to comprehend. Readability measures are mainly used to inform persons with reading disabilities on the complexity of a text, but can also be used to, for instance, assist teachers with assessing the reading ability of a student. By measuring the reading abilities of a person, it might also be possible to automatically find texts that fits that persons reading ability. It has further been shown that readability is a useful measure for finding a corpus for training vector space models (Smith et al., 2012).

Readability gives rise to a number of problems. For instance, readability is not a function of text only but a function of both text and reader, as defined by Dale and Chall (1949): "[Readability is] the sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it at optimal speed, and find it interesting." However, in this study we make the assumption that a function of text only can be a useful approximation. This assumption is supported by and related to the practice of American researchers to normalize their metrics to the U.S. grade level. Resources for such a normalisation for Swedish are not yet readily available and until they are we focus on the problem of classifying texts as either easy-to-read or not.

Readability assessment has been a field of interest for linguists since the 1920s but intensive research begun in the U.S. in the late 1940s (Sjöholm, 2012). This research resulted in the introduction of the first version of the Flesch Reading Ease test (Flesch, 1948) and the Dale-Chall formula, versions of which are still used today.

A number of easily calculated readability metrics (consisting of a small number of easily counted features such as average word length, lexical variation and frequency of "simple words") were introduced for English during the following three decades. Some examples are the Coleman-Liau index, which was specifically designed for automated assessment of readability (Coleman and Liau, 1975), the SMOG formula (McLaughlin, 1969) and the Fry readability formula (Fry, 1968). All of these metrics were designed to output a score corresponding to the U.S. grade level thought necessary for full comprehension of a text. In 1975 the Flesch Reading Ease test was reinvented as the Flesch-Kincaid Grade Level with the same principle in mind (Kincaid et al., 1975).

This way of constructing readability metrics was widely accepted as good enough for a long time. However, in the 1980's research questioning the performance of these traditional metrics was being published (Davison and Kantor, 1982).

Readability assessment for Swedish has mostly been done using metrics similar to the ones constructed for English. The most utilized readability metric for Swedish is LIX, Läsbarhetsindex (Readability index) (Björnsson, 1968), which is formulated in a way similar to that of the Flesch metric. Today the LIX metric is basically the standard metric for readability in Swedish. However, in recent years new research has shown that the metric is not always sufficient (Mühlenbock and Johansson Kokkinakis, 2009; Heimann Mühlenbock, 2013).

The OVIX Ordvariationsindex (Word variation index) and Nominal Ratio metrics (Hultman

and Westman, 1977) have been used in research to complement LIX as they are assumed to correlate with degree of readability viewed from other linguistic levels.

Since the early 2000s the speed and accuracy of text analysis tools such as lemmatizers, part-of-speech taggers and syntax parsers have made new text features available for readability assessment. By using machine learning a number of researchers have devised innovative ways of assessing readability. For instance, phrase grammar parsing has been used to find the average number of sub-clauses, verb phrases, noun phrases and average tree depth (Schwarm and Ostendorf, 2005).

The use of language models to assess the degree of readability was also introduced in the early 2000s (Collins-Thompson and Callan, 2004) and later combined with classification algorithms such as support vector machines to further increase accuracy (Petersen, 2007; Feng, 2010)

In this paper we present a study on the problem of finding and evaluating features relevant for classification. Such classifiers have previously been experimented with for Italian (Dell'Orletta et al., 2011). An extension of such a classifier has been proposed as an alternative to regression or detectors when it comes to ordering documents based on degree of readability (Falkenjack and Heimann Mühlenbock, 2012). The present approach is experimental in the sense that several feature models, simple as well as complex, are tested and compared. The models are based on text properties acting at various language levels, and the task is to identify the best-performing feature model for readability assessment viewed from one or several specific aspects of written language. An even more complex model where also the semantic aspect is taken into account would demand language resources supplied with information on concepts and meaning as for instance WordNet (Miller, 1995). Such an approach is presented in Heimann Mühlenbock (2013), where readability is regarded as the totality of features acting at five different levels of language representation, including the idea density level.

2 Study

In the study presented in this paper we evaluate a number of models for readability on a variety of corpora to assess the models' ability to classify a text as easy-to-read or not.

2.1 Corpora

To train and test our classifier we use one easy-to-read corpus and five corpora representing ordinary language in different text genres. The latter corpora will further on be labeled as non-easy-to-read. For each category we use 700 texts.

Our source of easy-to-read material is the LäSBarT corpus (Mühlenbock, 2008). LäSBarT consists of manually created easy-to-read texts from a variety of sources and genres.

The non-easy-to-read material comprise texts from a variety of corpora to make sure that what we are classifying is readability rather than genre. This material consists of 215 articles from GP2007 (news text), 34 whole issues of Forskning och Framsteg (popular science), 214 articles from Läkartidningen 05 (professional news), 214 public information notices from Smittskyddsinstitutet (government text) and 23 full novels from the Norstedts publishing house (fiction).

By using a corpus with such a variety of documents we will get texts that have different degree of readability which is important as we want to be able to use the same model on all types of text.

The texts are preprocessed using the Korp corpus import tool (Borin et al., 2012). Steps in the preprocessing chain relevant for this study are tokenization, lemmatisation, part-of-speech tagging and dependency grammar parsing. The Korp tool is publicly available for testing.

2.2 Classification

We use the Waikato Environment for Knowledge Analysis (Weka) suite and its implementation of the popular classification algorithm Support Vector Machine (SVM). Support Vector Machines has been increasingly popular in Computational Linguistics in recent years and have, among other uses, been used for readability assessment with good results (Petersen, 2007; Feng, 2010).

The SVM algorithm is an algebraic approach to the classification problem. Objects with a known class are represented as points in an n-dimensional space, where n is the number of attributes. An algorithm then attempts to find a maximum margin hyperplane separating the objects by their class (Witten et al., 2011). New objects are classified by calculating on which side of this hyperplane the object's corresponding point occurs.

The version of SVM-learning (finding the separating hyperplane) we use is the SMO, Sequential Minimal Optimization, algorithm (Platt, 1998). A Java implementation of a SMO-based SVM is included in the standard Weka toolkit.

We chose the SVM-approach as prior research has shown that it is one of the best performing algorithms for degree of readability classification using the full set, or subsets, of features we evaluate in this study (Sjöholm, 2012).

2.3 Models

We have constructed a total of 34 models. First we have three models representing the established Swedish metrics used to measure, or assumed to correlate with some aspect of readability, namely LIX, OVIX and Nominal ratio (NR).

We also use 21 single feature models. These models represent features proposed for readability assessment in prior research, mainly on English texts. As the primary aim of this study is to evaluate these feature models' ability to predict readability these models are the most important.

As many of the single feature models result from the same kind of preprocessing, we have also decided to create ten compound models. We divide the features into four levels similar to the four levels used by the READ-IT system for Italian (Dell'Orletta et al., 2011). These levels are Shallow (requires tokenization), Lexical (requires lemmatisation), morpho-syntactic (requires part-of-speech tagging) and Syntactic (requires parsing, in our case with a dependency grammar parser).

Seven models based on these levels are constructed, four which covers only a single level each; Shallow, Lexical, Morpho and Syntactic. Three models incrementally add levels to the analysis; the LexicalInc model which consists of all features from the Lexical and Shallow models, the MorphoInc model which consists of all features from the LexicalInc and Morpho models and the SyntacticInc model which consists of all features from the MorphoInc and the Syntactic models. These models are used to evaluate to what degree each level of linguistic analysis improves our model's ability to predict readability.

We also create three models combining the established metrics, LIX, OVIX and NR. The first, called TradComb, comprise only the three established metrics. The other two combine TradComb

with SyntacticInc (Total) and MorphoInc (NoDep) respectively.

2.3.1 Shallow features

The shallow text features are the main features traditionally used for simple readability metrics. They occur in the "shallow" surface structure of the text and can be extracted after tokenization by simply counting words and characters. They include:

AvgWordLengthChars Average word length calculated as the average number of characters per word.

AvgWordLengthSylls Average word length calculated as the average number of syllables per word. The number of syllables is approximated by counting the number of vowels.

AvgSentLength Average sentence length calculated as the average number of words per sentence.

Longer sentences, as well as longer words, tend to predict a more difficult text as exemplified by the performance of the LIX metric and related metrics for English. These types of features have been used in a number of readability studies based on machine learning (Feng, 2010) and as baseline when evaluating new features (Pitler and Nenkova, 2008).

2.3.2 Lexical features

Our lexical features are based on categorical word frequencies. The word frequencies are extracted after lemmatization and are calculated using the basic Swedish vocabulary SweVoc (Heimann Mühlenbock, 2013). SweVoc is comparable to the list used in the classic Dale-Chall formula (Dale and Chall, 1949) for English and developed for similar purposes, however special sub-categories have been added (of which three are specifically considered). The following frequencies are calculated, based on different categories in SweVoc:

SweVocC SweVoc lemmas fundamental for communication (category C).

SweVocD SweVoc lemmas for everyday use (category D).

SweVocH SweVoc other highly frequent lemmas (category H).

SweVocTotal Unique, per lemma, SweVoc words (all categories, including some not mentioned above) per sentence.

A high ratio of SweVoc words should indicate a more easy-to-read text. The Dale-Chall metric (Chall and Dale, 1995) has been used as a similar feature in a number of machine learning based studies of text readability for English (Feng, 2010; Pitler and Nenkova, 2008). The SweVoc metrics are also related to the language model features used in a number of studies (Schwarm and Ostendorf, 2005; Heilman et al., 2008).

2.3.3 Morpho-syntactic features

The morpho-syntactic features concern a morphology based analysis of text. For the purposes of this study the analysis relies on previously part-of-speech annotated text, which is investigated with regard to the following features:

UnigramPOS Unigram probabilities for 26 different parts-of-speech in the document, that is, the ratio of each part-of-speech, on a per token basis, as individual attributes. Such a unigram language model based on part-of-speech, and similar metrics, has shown to be a relevant feature for readability assessment for English (Heilman et al., 2007; Petersen, 2007; Dell'Orletta et al., 2011).

RatioContent The ratio of content words (nouns, verbs, adjectives and adverbs), on a per token basis, in the text. Such a metric has been used in a number of related studies (Alusio et al., 2010).

2.3.4 Syntactic features

These features are estimable after syntactic parsing of the text. The syntactic feature set is extracted after dependency parsing using the Maltparser (Nivre et al., 2006). Such parsers has been used for preprocessing texts for readability assessment for Italian (Dell'Orletta et al., 2011). The dependency based features consist of:

AvgDepDistDep The average dependency distance in the document on a per dependent basis. A longer average dependency distance could indicate a more complex text (Liu, 2008).

AvgDepDistSent The average dependency distance in the document on a per sentence basis. A longer average total dependency distance per sentence could indicate a more complex text (Liu, 2008).

RightDeps The ratio of right dependencies to total number of dependencies in the document. A high ratio of right dependencies could indicate a more complex text.

SentenceDepth The average sentence depth. Sentences with deeper dependency trees could be indicative of a more complex text in the same way as phrase grammar trees has been shown to be (Petersen and Ostendorf, 2009).

UnigramDepType Unigram probabilities for the 63 dependency types resulting from the dependency parsing, on a per token basis. These features are comparable to the part-of-speech unigram probabilities and to the phrase type rate based on phrase grammar parsing used in earlier research (Nenkova et al., 2010).

VerbalRoots The ratio of sentences with a verbal root, that is, the ratio of sentences where the root word is a verb to the total number of sentences (Dell'Orletta et al., 2011).

AvgVerbArity The average arity of verbs in the document, calculated as the average number of dependents per verb (Dell'Orletta et al., 2011).

UnigramVerbArity The ratios of verbs with an arity of 0-7 as distinct features (Dell'Orletta et al., 2011).

We also propose the following four syntactic features:

TokensPerClause The average number of tokens per clause in the document. This is related to the shallow feature average number of tokens per sentence.

PreModifiers The average number of nominal pre-modifiers per sentence.

PostModifiers The average number of nominal post-modifiers per sentence.

PrepComp The average number of prepositional complements per sentence in the document.

2.4 Evaluation

We evaluated the 21 single feature models presented above, three traditional metric models and the ten compound models presented above. Some features, rendered in italics in Table 2, consist of more than one concrete attribute and a few attributes are considered both as features in themselves and as attributes in larger feature models.

To test our models we use 7-fold cross validation over a set of 1400 documents. Each chunk consists of 100 easy-to-read texts and 100 non-easy-to-read texts. The corpora used to make up the non-easy-to-read set are shuffled and might therefore not be evenly distributed among the chunks (if this is a problem it should show up as a generally high standard deviation for all accuracies).

The result of the evaluation represents each model's ability to correctly identify easy-to-read texts. The accuracy of a model represents the proportion of the documents which are correctly classified as either easy-to-read or non-easy-to-read. A higher accuracy implies that the model, and its underlying features, more strongly predict degree of readability. To complement the accuracy we also provide precision and recall for the sets of easy-to-read texts and non-easy-to-read texts respectively, this to better understand where low performing models might go wrong.

In addition we present the standard deviation for each percentage based on the 7 folds of the cross validation. A high standard deviation implies inconsistent performance.

3 Results and discussion

The results of our test runs are presented below. We present the average values for the 7-fold cross validation in percent as well as the standard deviation in percentage points.

3.1 Traditional metrics

Among the traditional metrics (see Table 1) OVIX actually seems to perform about as well as LIX. This is somewhat surprising as LIX is designed to directly measure readability while OVIX is only assumed to indirectly measure readability. As OVIX considers totally different features from LIX, it does, perhaps, strengthen the point that LIX, as the standard readability metric for Swedish, might be overly simplistic.

Nominal ratio, NR, is the worst performing of the traditional metrics. It seems that the NR model tend to over-classify documents as easy to read, demonstrated by high recall but low precision for LäSBarT. As NR and ratio of content words, RatioContent, (see Table 2) both

Table 1: Performance of the three traditional metrics. The accuracy represents the average percentage of texts classified correctly, with the standard deviation within parentheses. Precision and Recall are also provided, with standard deviations within parenthesis, for both easy-to-read (LäSBarT) and non-easy-to-read (Other) sets.

		LäSBarT		Other	
Model	Accuracy	Precision	Recall	Precision	Recall
LIX	84.6 (1.9)	87.9 (2.9)	80.4 (2.8)	82.0 (2.1)	88.9 (3.0)
OVIX	85.6 (2.3)	86.8 (4.3)	84.4 (3.1)	84.9 (2.4)	86.9 (5.0)
NR	55.3 (9.1)	53.5 (6.8)	99.1 (1.9)	96.0 (7.7)	11.4 (20.1)

perform badly, it seems that only a more complex part-of-speech based feature, such as the multi-attribute feature consisting of unigrams for all POS-types is sufficient. Further analysis of single POS-type models might yield interesting results though.

3.2 Single feature models

Looking at Table 2 we see that most single feature models provide some indication on degree of text readability. There are however some models which perform a lot worse than anticipated.

It seems that the average dependency distance per sentence, AvgDepDistSent, is more or less useless, it might be that this is nothing more than a convoluted way to talk about sentence length, AvgSentLength, which in itself appears to be a highly inconsistent feature. Both these metrics over-classify documents as easy-to-read to a very high degree.

Also surprising is that the ratio of content words, RatioContent, does not seem to be a good indicator of readability. However, this does not seem to be a problem of over-classification, rather the model seems to be equally bad at classifying both sets, based on precision and recall close to 50 % for both sets. It might be that a high ratio of content words indicate a higher information density and therefore a more complex text while at the same time a low ratio might instead indicate a syntactically complex text. In such cases a simple SVM classification is not sufficient. Also, for an inflecting language like Swedish, the ratio of content words might yield different results than for languages with a more modest morphology, as for instance English. As with the nominal ratio metric, a closer inspection of single POS-type ratios might yield some further clues.

The average number of tokens per clause, TokensPerClause, and the ratio of nominal post-modifiers also seem to have a tendency to over-classify documents as easy-to-read having high LäSBarT recall but relatively low precision. Nominal pre-modifiers, however, while still suffering from slight easy-to-read over-classification, perform almost as well as LIX or OVIX when only accuracy is considered.

Best performing of the single feature models are the unigram models for part-of-speech, UnigramPOS, and dependency type, UnigramDepType. This is not surprising as these features represent simple language models and language models are often very powerful when compared to single attribute features.

It is only the unigram language models, UnigramPOS and UnigramDepType, and the average

Table 2: Performance of the single feature models, italicised models consist of more than one concrete attribute. The accuracy represents the average percentage of texts classified correctly, with the standard deviation within parentheses. Precision and Recall are also provided, with standard deviations within parenthesis, for both easy-to-read (LäSBarT) and non-easy-to-read (Other) sets.

		LäSBarT		Other	
Model	Accuracy	Precision	Recall	Precision	Recall
AvgWordLengthChars	79.6 (2.6)	82.3 (5.0)	75.7 (1.4)	77.4 (1.3)	83.4 (5.5)
AvgWordLengthSylls	75.6 (2.6)	78.7 (4.0)	70.3 (2.8)	73.1 (2.1)	80.9 (4.4)
AvgSentLength	62.4 (8.1)	58.0 (7.5)	98.7 (3.0)	97.8 (4.0)	26.1 (19.2)
SweVocC	79.3 (0.8)	84.3 (1.1)	72.0 (2.1)	75.6 (1.2)	86.6 (1.3)
SweVocD	57.6 (3.8)	63.1 (7.4)	37.9 (5.2)	55.5 (2.7)	77.4 (6.3)
SweVocH	63.1 (4.5)	63.1 (4.6)	63.4 (5.1)	63.2 (4.5)	62.9 (5.4)
SweVocTotal	75.2 (1.4)	80.6 (3.4)	66.7 (2.3)	71.6 (0.8)	83.7 (4.2)
UnigramPOS	*96.8 (1.6)*	*96.9 (2.5)*	*96.7 (1.1)*	*96.7 (1.1)*	*96.9 (2.6)*
RatioContent	50.4 (1.8)	50.4 (1.7)	52.7 (3.1)	50.4 (1.9)	48.1 (3.6)
AvgDepDistDep	88.5 (2.0)	88.5 (2.3)	88.6 (2.2)	88.6 (2.1)	88.4 (2.4)
AvgDepDistSent	53.9 (10.2)	52.8 (7.2)	99.7 (0.8)	28.1 (48.0)	8.1 (21.1)
RightDeps	68.9 (2.1)	70.6 (3.2)	65.1 (4.0)	67.7 (2.1)	72.7 (4.6)
SentenceDepth	75.1 (3.5)	79.1 (4.3)	68.4 (4.6)	72.2 (3.4)	81.9 (4.2)
UnigramDepType	*97.9 (0.8)*	*97.7 (1.1)*	*98.0 (1.3)*	*98.0 (1.3)*	*97.7 (1.1)*
VerbalRoots	72.6 (2.0)	77.0 (3.4)	64.6 (3.3)	69.5 (1.7)	80.6 (4.3)
AvgVerbArity	63.4 (3.0)	64.9 (3.2)	58.4 (4.9)	62.3 (3.0)	68.4 (3.2)
UnigramVerbArity	68.6 (1.7)	70.2 (2.6)	65.0 (2.8)	67.4 (1.5)	72.3 (4.0)
TokensPerClause	71.4 (4.7)	64.2 (4.4)	98.6 (1.0)	97.0 (1.8)	44.3 (10.0)
PreModifiers	83.4 (2.9)	78.1 (3.1)	93.0 (2.2)	91.3 (2.6)	73.9 (4.5)
PostModifiers	57.4 (4.3)	54.1 (2.7)	99.9 (0.4)	98.4 (4.2)	15.0 (8.5)
PrepComp	83.5 (3.5)	80.1 (2.4)	89.1 (5.9)	88.1 (5.8)	77.9 (2.7)

dependency distance per dependency, AvgDepDistDep, that outperform the traditional metrics OVIX and LIX. However, the average number of prepositional complements per sentence, PrepComp, and nominal pre-modifiers per sentence, PreModifiers, respectively do come close.

3.3 Compound models

When we look at the compound models, Table 3 we can see highly improved performance. Not surprisingly, we get the best performance from the Total model consisting of all features that the system is able to extract.

All compound metrics except for the Shallow and Lexical models outperform the traditional metrics. However, combining these two, the LexicalInc model, does outperform the traditional metrics LIX, OVIX and NR.

Interestingly the UnigramPOS feature model seems to perform slightly better than the Morpho model (which actually consists of UnigramPOS and RatioContent). The bad performance of the ratio of content words model, RatioContent, might introduce some performance-decreasing

Table 3: Performance of the ten compound models. The accuracy represents the average percentage of texts classified correctly, with the standard deviation within parentheses. Precision and Recall are also provided, with standard deviations within parenthesis, for both easy-to-read (LäSBarT) and non-easy-to-read (Other) sets.

		LäSBarT		Other	
Model	Accuracy	Precision	Recall	Precision	Recall
TradComb	91.4 (3.0)	92.0 (4.6)	91.0 (2.1)	91.1 (2.2)	91.9 (4.9)
Shallow	81.6 (2.7)	83.3 (4.4)	79.4 (3.1)	80.3 (2.5)	83.9 (4.9)
Lexical	78.4 (2.2)	81.8 (2.9)	73.0 (2.9)	75.6 (2.1)	83.7 (3.0)
Morpho	96.7 (1.6)	96.8 (2.6)	96.7 (1.4)	96.7 (1.3)	96.7 (2.7)
Syntactic	98.0 (1.1)	97.9 (1.7)	98.1 (1.2)	98.1 (1.2)	97.9 (1.8)
LexicalInc	90.1 (2.9)	87.1 (4.1)	94.3 (2.6)	93.8 (2.7)	85.9 (4.9)
MorphoInc	97.3 (0.8)	96.9 (1.6)	97.7 (1.6)	97.7 (1.5)	96.9 (1.7)
SyntacticInc	98.4 (0.9)	98.3 (1.4)	98.6 (1.0)	98.6 (1.0)	98.3 (1.4)
NoDep	98.3 (1.0)	97.4 (1.9)	99.3 (1.3)	99.3 (1.2)	97.3 (2.0)
Total	98.9 (1.0)	98.9 (1.1)	98.9 (1.1)	98.9 (1.1)	98.9 (1.1)

confusion though.

The Morpho and Syntactic models both more or less equal the UnigramPOS and UnigramDepType models respectively implying that these are by far the most important features in the respective models.

4 Conclusions

In this study we have presented a large number of feature models proposed for readability assessment. Most of these models have previously been shown to be useful for assessing readability of English texts. Our results show that many of them are also relevant for Swedish, however some models are less relevant, most notably the ratio of content words, RatioContent, for which we have no simple explanation. Contrary to, for instance, NR which erroneously classify many texts as readable and consequentially achieves avery low accuracy, RatioContent does not have a high recall on any category in the corpora.

The best performing features seem to be part-of-speech or dependency type based language models, especially the compound models that require parsing using a dependency parser; Syntactic, SyntacticInc and Total. These models all have high Accuracy, more than 98% and a fairly low standard deviation showing a stable performance.

We also show that a combination of the three established metrics outperform the standard LIX metric but also that the use of the raw data necessary to calculate those metrics (the data in the MorphoInc model) might possibly be put to even better use. Dependency grammar parsing seems to provide very useful data for identifying easy-to-read texts but in an environment where such heavy calculations are infeasible a very good result might be found without it as demonstrated by the NoDep model.

We propose that future research look further into the language models represented by the UnigramPOS and UnigramDepType models. It might be possible to construct relatively simple

metrics based on only a few of the attributes in these models. It might also be possible to construct even better performing models by looking at bigrams or trigrams instead of just unigrams.

Acknowledgments

We would like to thank Santa Anna IT Research Institute AB for funding this research as well as the staff members at Språkbanken who created and let us use the Korp corpus import tool.

References

Alusio, S., Specia, L., Gasperin, C., and Scarton, C. (2010). Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9.

Björnsson, C. H. (1968). *Läsbarhet*. Liber, Stockholm.

Borin, L., Forsberg, M., and Roxendal, J. (2012). Korp – the corpus infrastructure of språkbanken. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*.

Chall, J. S. and Dale, E. (1995). *Readability revisited: The new Dale–Chall readability formula*. Brookline Books, Cambride, MA.

Coleman, M. and Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284.

Collins-Thompson, K. and Callan, J. (2004). A Language Modeling Approach to Predicting Reading Difficulty. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Dale, E. and Chall, J. S. (1949). The concept of readability. *Elementary English*, 26(23).

Davison, A. and Kantor, R. N. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly*, 17(2):187–209.

Dell'Orletta, F., Montemagni, S., and Venturi, G. (2011). READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification. In *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83.

Falkenjack, J. and Heimann Mühlenbock, K. (2012). Readability as probability. In *Proceedings of The Fourth Swedish Language Technology Conference*, pages 27–28.

Feng, L. (2010). *Automatic Readability Assessment*. PhD thesis, City University of New York.

Flesch, R. (1948). A new readibility yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Fry, E. B. (1968). A readability formula that saves time. *Journal of Reading*, 11:513–516.

Heilman, M. J., Collins-Thompson, K., Callan, J., and Eskenazi, M. (2007). Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. In *Proceedings of NAACL HLT 2007*, pages 460–467.

Heilman, M. J., Collins-Thompson, K., and Eskenazi, M. (2008). An Analysis of Statistical Models and Features for Reading Difficulty Prediction. In *Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications*, pages 71–79.

Heimann Mühlenbock, K. (2013). *I see what you mean. Assessing readability for specific target groups*. Dissertation, Språkbanken, Dept of Swedish, University of Gothenburg.

Hultman, T. G. and Westman, M. (1977). *Gymnasistsvenska*. LiberLäromedel, Lund.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy enlisted personnel. Technical report, U.S. Naval Air Station, Millington, TN.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):169–191.

McLaughlin, G. H. (1969). SMOG grading - a new readability formula. *Journal of Reading*, 22:639–646.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Mühlenbock, K. (2008). Readable, Legible or Plain Words – Presentation of an easy-to-read Swedish corpus. In Saxena, A. and Viberg, Å., editors, *Multilingualism: Proceedings of the 23rd Scandinavian Conference of Linguistics*, volume 8 of *Acta Universitatis Upsaliensis*, pages 327–329, Uppsala, Sweden. Acta Universitatis Upsaliensis.

Mühlenbock, K. and Johansson Kokkinakis, S. (2009). LIX 68 revisited - An extended readability measure. In Mahlberg, M., González-Díaz, V., and Smith, C., editors, *Proceedings of the Corpus Linguistics Conference CL2009*, Liverpool, UK.

Nenkova, A., Chae, J., Louis, A., and Pitler, E. (2010). *Structural Features for Predicting the Linguistic Quality of Text Applications to Machine Translation, Automatic Summarization and Human–Authored Text.*, pages 222–241. Empirical Methods in NLG. Springer-Verlag.

Nivre, J., Hall, J., and Nilsson, J. (2006). MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)*, pages 2216–2219.

Petersen, S. (2007). *Natural language processing tools for reading level assessment and text simplification for bilingual education*. PhD thesis, University of Washington, Seattle, WA.

Petersen, S. and Ostendorf, M. (2009). A machine learning approach toreading level assessment. *Computer Speech and Language*, 23:89–106.

Pitler, E. and Nenkova, A. (2008). Revisiting Readability: A Unified Framework for Predicting Text Quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, HI.

Platt, J. C. (1998). Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Technical Report MSR-TR-98-14, Microsoft Research.

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

Sjöholm, J. (2012). Probability as readability: A new machine learning approach to readability assessment for written Swedish. Master's thesis, Linköping University.

Smith, C., Danielsson, H., and Jönsson, A. (2012). A good space: Lexical predictors in vector space evaluation. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC), Istanbul, Turkey*.

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. The Morgan Kaufmann series in data management system. Morgan Kaufmann Publishers, third edition.