# Using the probability of readability to order Swedish texts

**Johan Falkenjack**[1], **Katarina Heimann Mühlenbock**[2]

(1) Santa Anna IT Research Institute AB, Linköping, Sweden
(2) Språkbanken, University of Gothenburg, Gothenburg
johsj@ida.liu.se, katarina.heimann.muhlenbock@gu.se

**Abstract**

In this study we present a new approach to rank readability in Swedish texts based on lexical, morpho-syntactic and syntactic analysis of text as well as machine learning. The basic premise and theory is presented as well as a small experiment testing the feasibility, but not actual performance, of the approach. The experiment shows that it is possible to implement a system based on the approach, however, the actual performance of such a system has not been evaluated as the necessary resources for such an evaluation does not yet exist for Swedish. The experiment also shows that a classifier based on the aforementioned linguistic analysis, on our limited test set, outperforms classifiers based on established metrics used to assess readability such as LIX, OVIX and Nominal Ratio.

## 1. Motivation

Studies have shown that as many as 25 % of the Swedish adult population can not read at the level expected of students in the 9th grade in the Swedish school system. For many of these people, access to information is dependent on the ability to find the most easy-to-read texts describing the subject.

To this purpose a search engine, Webblättlast, capable of finding not only the most relevant texts but also the most easy-to-read is being developed at the Department of Computer and Information Science at Linköping University. This search engine mainly uses the three established Swedish metrics, LIX (Läsbarhetsindex (Björnsson, 1968)) which is a readability metric based on surface structure and OVIX (Ordvariationsindex (Hultman and Westman, 1977)) and Nominal ratio (Hultman and Westman, 1977) which are complexity metrics which measure word variation and information density respectively.

However, research has shown that these established Swedish readability metrics are insufficient when used individually (Mühlenbock and Johansson Kokkinakis, 2009). Also, the same study showed that LIX and OVIX very well might result in different orderings when used to rank documents according to supposed degree of readability.

## 2. Background

The years since 2000 have seen quite a few developments in the field of readability assessment for English. Some new readability assessment systems have utilized tools such as grammar parsers and discourse analysis (Feng et al., 2009). Other more data intensive studies have applied statistical language models such as n-gram models to the field of readability assessment with good results (Collins-Thompson and Callan, 2004). Most of these approaches were based on access to a corpus of readability assessed texts, Weekly Reader.

All texts in the Weekly Reader corpus are tagged with a suitable grade level in the U.S. school system. These grade levels can been used both as a basis for regression (Pitler and Nenkova, 2008) and for creation of detectors, single-class classifiers (Petersen, 2007). Both a formula generated by regression and a set of detectors can be used for ranking documents according to degree of readability.

However, no equivalent corpora exist for Swedish so another approach must be devised.

## 3. A new approach

If the assumption is made that *the degree of readability of a text is proportionate to the probability that the text is classified as easy-to-read by a perfect classifier* the problem becomes one of constructing such a classifier and finding a way to extract probabilities from it. Of course, such a perfect classifier is a purely theoretical construct. However, a good enough linear classifier, tweaked to output class probabilities (soft classification) rather than just the most probable class (hard classification), should be able to calculate a reasonable approximation, at least within a limited span. While this metric might not be linear, and therefore perhaps not suitable for single document assessment without some kind of smoothing, it should provide a way to rank documents based on their degree of readability.

## 4. Feasibility of the approach

To test whether the approach is feasible we first have to test whether a traditional classifier, able to identify easy-to-read texts, can be constructed. To do this we used documents from the LäSBarT easy-to-read corpus (Mühlenbock, 2008) to represent easy-to-read texts and documents from the GP2007, a corpus made up of articles from the newspaper Göteborgsposten from 2007, to represent non-easy-to-read texts.

### 4.1 Hard classification

These documents were analysed using the Korp corpus import tool developed by Språkbanken and six different models were created. Three models were based on the established Swedish readability metrics LIX, OVIX and Nominal ratio (NR), the fourth model (COM) combined all three, the fifth model (NODEP) added further surface, lexical and morpho-syntactic features and the sixth and last

model (FULL) added features based on dependency parsing. These larger models are similar to the ones used by the Italian READ-IT project (Dell'Orletta et al., 2011).

Using the Waikato Environment for Knowledge Analysis, or WEKA, we tested the accuracy of a support vector machine (using the sequential minimal optimization training algorithm (Platt, 1998)) with the six different models. Each model was tested using 7-fold cross-validation over 1400 documents, 700 from each corpus. See Table 1 for the results.

| Model | Accuracy |
|---|---|
| LIX | 77.4 |
| OVIX | 84.5 |
| NR | 53.0 |
| COM | 89.3 |
| NODEP | 97.0 |
| FULL | 97.6 |

Table 1: The accuracy, percentage of correctly classified documents, for each model using hard classification.

Overall results, a maximum accuracy of 97.6 %, implies that it is possible to create a reasonably accurate classifier for easy-to-read Swedish documents.

### 4.2 Soft classification

If we are to order documents based on their probability of readability our classifier must be able to output a class probabilities. This can, for our classifier, be done by fitting logistic models to the output from the SVM. (Another approach would be to use logistic regression alone but as an SVM was the most accurate classifier in a related experiment we decided to use this hybrid approach.)

We must, however, make sure that this does not impair the accuracy of the classifier. We should also check the number of equivalence classes generated in the test, that is, how many documents are awarded the same probability of readability and therefore not sortable with regard to each other. As only the NODEP and the FULL models had accuracies $> 90$ % only these models were evaluated. All documents with an error smaller than 50 % were considered correct. The same 7-fold cross validation scheme was used again and the calculated percentages were registered to calculate the number of equivalence classes. See Table 2 for the results.

| Model | Accuracy | #Equivalence classes |
|---|---|---|
| NODEP | 97.7 | 1398 |
| FULL | 97.0 | 1398 |

Table 2: The accuracy and number of equivalence classes for each model using soft classification.

The result shows that the accuracy is still high and the number of equivalence classes shows that all but 3 documents are sortable.

## 5. Conclusion and future work

The experiment shows that a SVM classifier with a high accuracy on readability based classification also can order documents without a large risk of non-sortable pairs. This implies that a system for ranking documents based on this principle could be feasible. However, other classification algorithms, more suited for what we call soft classification, such as pure logistic regression, might be more effective as they might produce more normalized results.

However, until tested we can not be sure that the ability to accurately identify easy-to-read documents entails the ability to order documents according to degree of readability. Further research, using ordered sets of documents, is necessary.

## 6. References


Carl Hugo Björnsson. 1968. *Läsbarhet*. Liber, Stockholm.

Kevyn Collins-Thompson and Jamie Callan. 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*.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 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, July.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively Motivated Features for Readability Assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL*.

Tor G. Hultman and Margareta Westman. 1977. *Gymnasistsvenska*. LiberLäromedel, Lund.

Katarina Mühlenbock and Sofie Johansson Kokkinakis. 2009. LIX 68 revisited - An extended readability measure. In Michaela Mahlberg, Victorina González-Díaz, and Catherine Smith, editors, *Proceedings of the Corpus Linguistics Conference CL2009*, Liverpool, UK, July 20-23.

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

Sarah Petersen. 2007. *Natural language processing tools for reading level assessment and text simplification for bilingual education*. Ph.D. thesis, University of Washington, Seattle, WA.

Emily Pitler and Ani Nenkova. 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, October.

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