Applying and Optimising a Multi-Scale Probit Model for Cross-Source Text Complexity Classification and Ranking in Swedish

Elsa Andersson, Johan Falkenjack, Arne Jönsson
Department of Computer and Information Science
Linköping Univsersity
Linköping, Sweden
elsa@anderssondito.se, {johan.falkenjack, arne.jonsson}@liu.se

Abstract

We present results from using Probit models to classify and rank texts of varying complexity from multiple sources. We use multiple linguistic sources including Swedish easy-to-read books and investigate data augmentation and feature regularisation as optimisation methods for text complexity assessment. Multi-Scale and Single Scale Probit models are implemented using different ratios of training data, and then compared. Overall, the findings suggest that the Multi-Scale Probit model is an effective method for classifying text complexity and ranking new texts and could be used to improve the performance on small datasets as well as normalise datasets labelled using different scales.

1 Introduction

Measuring or estimating text complexity is essential in various fields, including readability research and the adaptation and recommendation of texts for different audiences. In this paper, text complexity refers only to the linguistic characteristics that affect how easy or difficult a text is to read, without considering the interaction between the text and any particular reader.

Any comprehensive evaluation of text complexity must include three key components. First, linguistic features must be quantified, such as calculating the average sentence length. Second, relevant linguistic features need to be selected for evaluation. Third, the impact of each linguistic feature on text complexity must be assessed, for example, determining whether longer sentences increase or decrease complexity and to what extent. The distinction between effective and ineffective evaluations lies in the execution of these components. The selection of features and the methods employed to measure them significantly affect the quality of the evaluation (Bailin and Grafstein, 2001).

Moreover, text complexity is not defined by a single superficial quality; rather, it results from an interplay of various features, each influencing complexity in distinct ways (Santini and Jönsson, 2020). Understanding how and to what extent each linguistic feature contributes to overall text complexity poses an additional challenge. The approaches for identifying and selecting linguistic features vary, ranging from employing theoretical linguistic frameworks and reasoning about feature impacts (Ellis, 2020) to training machine learning models on specific features and assessing their performance (Falkenjack et al., 2013), or even employing a combination of these methods.

Another aspect of the assessment of text complexity is the type of output that is produced. Depending on the purpose of the evaluation, the results may be in the shape of a single binary classification of ”easy to read” or ”not easy to read”. This type of evaluation is traditionally realised through simple linear functions, or more recently using machine learning models like the Support Vector Machine (SVM) that splits texts into two classes (Benjamin, 2012). Another common evaluation method is to use one or a few linguistic features in a simple equation (often referred to as readability formulas) and computing a score to measure the complexity (e.g. the *Flesch Reading Ease formula* (Flesch, 1948)). These methods are beneficial in several ways, but all share a common downside. When using a few simple features or classifying texts in a binary manner, much nuance of text complexity is lost, and comparisons between texts are less informative (Bailin and Grafstein, 2001).

To solve these problems, we propose creating a model that uses many complex linguistic features

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 17–27

and classifies or ranks texts into non-binary levels. This approach would, however, usually require data that are already labelled according to class or rank. The more features used in the model to increase the complexity of the evaluation, the more data is required in each class or rank (Bengio et al., 2000).

One method that has the potential to resolve many of the issues mentioned above is the Multi-Scale Probit model, proposed and first implemented in Falkenjack (2018). The Probit model is a well established statistical model, introduced in the 1930s (Bliss, 1934b) and used primarily for classification. It is closely related to the younger but somewhat more well known *Logit model*, or *Logistic regression* as it is often called in psychometric contexts, but it has some properties which make it especially suitable for Bayesian modelling (McCulloch et al., 2000).

The Multi-Scale Probit model is a generalisation of the Bayesian Ordered Probit model and is capable of training on data labelled into ordered levels, such as how hard a text is to read, from multiple text sources. These sources may use completely different scales, meaning that the levels need not correspond in any sense between sources apart from indicating text complexity. There is no requirement for a minimum amount of texts per level, which enables the use of data that would have to be discarded in other approaches. The key idea behind the model is the presence of a *latent variable* that is shared among all labelling schemes. In this context, that latent variable is text complexity, with the assumption that the different labelling schemes used across different data sources all represent measures of that latent variable. Information about the latent variable is captured in the features, and the model learns how the latent variable is affected by the features, making it able to classify and even rank the text complexity of new texts.

We explore how the Multi-Scale Probit model performs when trained and evaluated on novel data, consisting of easy-to-read literature for children, teenagers, and adults[1].

2 Text complexity analysis

Text complexity generally refers to characteristics of a text that make it more or less cognitively engaging during reading (Vega et al., 2013). Quantitative and qualitative assessments of text complexity are of great value, as they can be used in many fields such as education (e.g. determining the appropriate material (Fitzgerald et al., 2015) or automatic essay grading (Valenti et al., 2003)), customisable text simplification (e.g. determining which texts to simplify (Štajner et al., 2012)), or customising texts based on cognitive requirements (e.g. for readers with dyslexia (Santini and Jönsson, 2020)).

Pinpointing the properties of a text that tells us about its complexity has been proven to be a difficult and confusing task. The factors that make up the complexity of a text can themselves create a hyperplane that spans across a highly multidimensional space.

Classification is a simplified version of this with the purpose of assigning texts into one or more classes such as ”easy to read”. Classification approaches consist of machine learning algorithms, statistical methods, and other NLP techniques.

Such approaches need to be trained on different text features or combinations of features and then evaluated on their performance in classifying texts accurately. As model performance becomes an indirect measurement of the relevance of the feature(s) to text complexity analysis, the features used to train models with better performance are chosen over the features of models with poorer performance (Falkenjack, 2018). Another category of classification algorithms is logistic regression and its variants. Compared to SVMs and similar methods, cf. Schwarm and Ostendorf (2005); Pitler and Nenkova (2008); Falkenjack (2018), the binary outcome is modelled as a probability between 0 and 1. For instance, a book could be classified as ”easy to read” with a probability of 0.6, meaning that there is a 60% probability (according to the model) that the book is ”easy to read”. A common approach for such probability estimation is Logistic regression (Hosmer Jr et al., 2013), or the Logit model, and in this paper we apply a version of the closely related Probit model.

3 Text complexity features

The most commonly used method for the analysis of text complexity is automatic evaluation using quantifiable features of texts, which are then used to compute one or more ratings of text complexity. These features measure different aspects of the text and can be categorised into four ordered levels

[1] https://www.nyponochviljaforlag.se/om-oss/

of increasing analytical depth, as outlined below.

Shallow features: The features in the first category do not contain information about the content of the text. They simply consist of letter and word counts; very little or no knowledge of their meaning is necessary to measure or understand shallow features. Nevertheless, they have been proven to be useful for measuring text complexity and are very simple to extract. The text is processed through tokenisation to create tokens out of words (and other components, e.g. delimiters). The tokens can be counted either as they are or by tallying the characters they contain. Several traditional metrics are based on one or more of these features or variants thereof, cf. Flesch (1948); Björnsson (1968).

Lexical composition: The lexical composition of a text targets frequencies of words based on the lexical category they belong to. The categorisation process includes lemmatising all words using a large vocabulary. For Swedish text, a vocabulary called *SweVoc* was developed in 2012 for this purpose by Mühlenbock and Kokkinakis (2012). In *SweVoc*, each word is represented as a lemma with some additional information depending on how it is used, including which category (or categories) it belongs to. In this research, the following categories will be used: *SweVocD* (words related to every-day matters), *SweVocH* (high-frequency words), and *SweVocTotal* (the total ratio of words in the text that are part of *SweVoc*). Because the *SweVoc* vocabulary is a subset of the Swedish language which excludes some complex or specialised words, it could be assumed that easy-to-read texts have a higher ratio of *SweVoc* words than more complex texts.

Morpho-syntactic features: Morpho-syntactic features include tagging words and tokens according to their part-of-speech (POS). The POS tags can then be used in a number of text features. In this research, *UnigramPOS* features will be used. The *UnigramPOS* features are the probabilities of a unigram occurring in a text, expressed as the ratio of each POS tag per token. Calculating the unigram probabilities of a text is a type of language modelling that can be effective in measuring the readability of a text (Heilman et al., 2007).

Syntactic features: Although unigram language models are effective in capturing content information and variations in word usage, they lack the ability to capture syntactic information. The analysis of syntactic complexity requires parsing of the text, which involves mapping words and phrases and their dependencies based on grammatical structures of a sentence. For this research, the syntactic text features consist of a subset of features extracted through dependency parsing on each sentence. These features are: *UnigramDep* (probabilities for each dependency relation type), *RightDep* (ratio of total dependencies where the headword occurs after the dependent word), *UVA* (unigram probabilities for verbs with a specific number of dependants), and *Lexical density* (ratio of content words).

4 Text complexity as a latent variable

A key idea behind the use of statistical models for text complexity assessment is the assumption of a latent variable. As established, text complexity cannot be measured directly. Instead, it is estimated using one or more linguistic features. Furthermore, text complexity is assessed and labelled in various ways. For example, texts may be labelled as "easy to read" (with the implication that regular texts are less "easy to read"), rated on a scale of 1 to 7, or categorised into age groups, among other methods. Although texts from different sources may use varying labels and methods to measure readability, we assume that they share the underlying *latent variable* of text complexity. In other words, variations in text complexity may be expressed differently, but the concept is consistently modelled across all sources. If data can be processed appropriately, it enables the latent variable to be statistically modelled and subsequently used to classify or rank texts.

4.1 The Probit model

The Multi-Scale Probit model we use is a generalisation of the Ordered Probit model which itself is a generalisation of the Probit model. The Probit model can be viewed as a linear binary classifier. It can also be considered a *Generalized Linear Model* with the inverse of the cumulative distribution function (CDF) of the Standard Normal distribution, the *Probit* function (Bliss, 1934a), as link function. In essence, the Probit model takes a vector of covariates $\mathbf{x}_i$ of the ith observation and uses it to predict the outcome, or label, y_i. It does so by estimating a coefficient vector $\boldsymbol{\beta}$ that represents the effects of $\mathbf{x}_i$ on the value of y_i. In simple terms, the model can be expressed as "what is the

probability that y_i is 1, given the information in $\mathbf{x}_i$?". Mathematically, the Probit model can be expressed as

$$P(y_i = 1|\mathbf{x}_i) = \Phi(\alpha + \mathbf{x}_i{}^T\boldsymbol{\beta}), \quad (1)$$

where Φ is the CDF of the Standard Normal distribution and α is the *intercept*, defined as a constant value that represents the baseline probability of y_i being 1 even if all covariates are 0.

In the context of text complexity, $\mathbf{x}_i$ would consist of measurements of linguistic features and y_i would represent a certain label, for example "easy to read". Furthermore, the Probit model can generally be conceptualised as a latent variable model, the latent variable y^* in our application being text complexity. By setting a threshold $\gamma = -\alpha$ and denoting the two binary outcomes as 1 and 2, the Probit model can instead be expressed as

$$y_i = \begin{cases} 2 & \text{if } y_i^* > \gamma \\ 1 & \text{otherwise} \end{cases} \quad \text{where} \quad y_i^* = \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon_i \quad (2)$$

where y_i^* represents the value of the latent variable and ϵ_i is the error term for the *i*th observation. Under this interpretation, we can view the Probit model as a linear regression over an unobserved, or latent, real-valued variable which underlies the assigned labels in the classification problem. If class 1 represents "easy to read" and class 2 'not easy to read', this can be expressed as "if the complexity of a text is above a certain threshold, it should be classified as 'not easy to read', otherwise it should be classified as 'easy to read'".

This latent variable formulation can be generalised to the case of an ordinal response variable with possible outcomes $C_1...C_m$ by introducing further thresholds $\gamma_1...\gamma_{m-1}$ giving rise to the Ordered Probit model:

$$y_i = \begin{cases} C_1 \text{ if } y_i^* \leq \gamma_1, \\ C_2 \text{ if } \gamma_1 < y_i^* \leq \gamma_2, \\ \vdots \\ C_m \text{ if } y_i^* > \gamma_{m-1} \end{cases} \quad (3)$$

where y_i^* is the same as in Equation 2.

The latent variable interpretation of Probit models lends itself especially well to a Bayesian approach. Essentially, a Bayesian approach entails declaring a prior belief, which is then updated using Bayes Theorem as new evidence is gathered, generating a posterior belief based on that evidence. These beliefs are commonly referred to as simply the prior and the posterior. Bayes' theorem can be applied for inference of the posterior probability distribution of the coefficients vector $\boldsymbol{\beta}$ and thresholds $\boldsymbol{\gamma}$ according to the following formulation

$$P(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X}) \propto P(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{X}) P(\boldsymbol{\beta}, \boldsymbol{\gamma}), \quad (4)$$

where $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$, $\mathbf{y} = (y_1, \ldots, y_n)^T$, $P(\boldsymbol{\beta}), \boldsymbol{\gamma}$ is the prior and $P(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{X})$ is the likelihood function. Although this posterior distribution is mathematically intractable, the *Markov Chain Monte Carlo* (MCMC) simulation can be used to estimate the posterior. Gibbs samplers for both the binary (Albert and Chib, 1993) and ordinal (Cowles, 1996) versions are well established.

The goal of the sampling process for the formulation in Equation 4 is to approximate the joint posterior distribution of $\boldsymbol{\beta}$ by estimating marginal distributions of individual variables.

4.2 The Multi-Scale Probit model

The formulation for the Probit model as a model for the latent variable in Equation 3 can be extended further to fit binary and non-binary data labelled on different scales. Let us take a practical example to demonstrate these characteristics. Say we have books sourced from two publishers, *A* and *B*. Publisher *A* labels its books on a scale from 'easy', 'medium' to 'hard' based on readability. Publisher *B* labels its books on a scale from 1 to 5, also based on readability. The publishers use unknown and possibly different methods for measuring readability, the difference in complexity between each level within either scale is unknown, and there is no known function translating between the scales. The only assumption we make is that the labels are ordered and that they constitute measures of the same phenomenon, i.e. text complexity. The Multi-Scale Probit model uses one set of thresholds to discriminate between levels for each scale such that $\gamma^{(s)}$ is the set of thresholds for scale *s*. Using our example, the two sets would be $\gamma^{(A)} = \{\gamma^{(A_{easy})}, \gamma^{(A_{medium})}, \gamma^{(A_{hard})}\}$ and $\gamma^{(B)} = \{\gamma^{(B_1)}, \gamma^{(B_2)}, \gamma^{(B_3)}, \gamma^{(B_4)}, \gamma^{(B_5)}\}$. Furthermore, the model fits a single latent variable y^* to all data, meaning that only a single coefficient vector $\boldsymbol{\beta}$ is estimated. The Multi-Scale Probit model

can therefore be expressed as

$$y_i = \begin{cases} C_1^{(s_i)} & \text{if } y_i^* \leq \gamma_1^{(s_i)}, \\ C_2^{(s_i)} & \text{if } \gamma_1^{(s_i)} < y_i^* \leq \gamma_2^{(s_i)}, \\ \vdots & \\ C_m^{(s_i)} & \text{if } \gamma_{m-1}^{(s_i)} < y_i^* \end{cases} \quad (5)$$

for observation $i = 1, \ldots, n$, where y_i^* is the same as in Equations 2 and 3, the response label y_i is measured on scale s_i, and $C_1^{(s_i)} \ldots (C_m^{(s_i)}$ denotes the labels for scale s_i. The complete posterior distribution of the joint is estimated using a variation of the Gibbs sampling algorithm proposed by Cowles (1996) for the Ordinal Probit model. The conditional posteriors for all sets of $\gamma^{(s)}$, $\boldsymbol{\beta}$ and the latent variable y^* can be sampled through the process described above. The latent variable estimated by the Multi-Scale Probit model can be used to order data samples, enabling total ranking of all data samples. Essentially, the Multi-Scale Probit allows us to, from some number of disjunct and partially ordered sets, estimate a total order on the union of all sets.

The applicability of the Multi-Scale Probit to our domain has previously been investigated in Falkenjack et al. (2018).

4.3 Measures for evaluation

Because the Multi-Scale Probit model can be used for both classification and ranking, we want to evaluate it using appropriate measures for each purpose.

As the data we use are not balanced, i.e. there is not a consistent number of observations per class, straight *accuracy* would not be a suitable metric if we consider the performance as equally important for all classes. In such cases, it is common to use the *macro-averaged* F_1-score (Murphy, 2012, p. 185). The F_1-score of a single class is the harmonic mean of the *precision* and the *recall* for that class. The macro-averaged F_1-score is the average of the F_1-scores for all classes. This value can be used as an overall measurement of how well the model performs in regards to classification.

The Multi-Scale Probit estimates a numeric latent variable and can thus be viewed as a model for ranking in addition to classification. We evaluate this performance by computing the Kendall rank correlation coefficient, τ, between the estimated latent variable and the known observed variable.

Kendall's τ assesses the ordinal association between two variables and gives a score between -1 and 1 depending on the correlation. Since the observed variable is an ordinal class, giving rise to a large number of ties, we use a modified version called Kendall's τ_B specifically made to handle such situations (Kendall, 1945).

Just as the F-measure uses the harmonic mean between *Precision* and *Recall*, we can combine the classification performance F_1 and ranking performance Kendall's τ_B using the harmonic mean. We use this as a combined performance metric for both classification and ranking in our figures in Section 7.

5 Data

The majority of data used in this research consisted of books from a corpus called Nypon-Vilja, consisting of books from *Nypon och Vilja*, the largest Swedish publisher of easy-to-read literature for children, teenagers, and adults. Swedish easy-to-read literature is catered to people with reading difficulties, beginner readers, or non-native readers learning Swedish.

Books from *Nypon* and *Vilja* are (generally) aimed at two different target groups; *Nypon* at 'children and young' and *Vilja* at 'young adults and adults'. The publisher uses separate scales (with their own naming schemes), each consisting of 6 levels, to indicate how easy or difficult a book is, where the first level (1 and X-Small) is the easiest and the last level (6 and XX-Large) the most difficult.

Before processing, all books were manually annotated based on their alignment with one of two narrativity dimensions: *informational* and *narrative* (McNamara, 2013). *Informational* text tends to be non-fictional, written to inform about or explain a specific topic. *Narrative* text on the other hand is typically fictional and story-driven. In order to minimise the effects of variations in language use that affect text complexity between dimensions of narrativity, only books classified as *narrative* are used. Finally, as level 6 from *Nypon* contained only 2 books, they were merged with the books in level 5. This resulted in a dataset of 356 books with 5 levels in *Nypon* and 6 levels in *Vilja*, summarised in Table 1.

The Stockholm-Umeå Corpus (Ejerhed et al., 2006) (*SUC*) is a large collection of annotated Swedish texts written in the 90's. It contains

Nypon		Vilja	
Level	**N samples**	**Level**	**N samples**
1	48	X-Small	4
2	59	Small	14
3	68	Medium	20
4	46	Large	42
5	13	X-Large	38
		XX-Large	4
Sum	234		122

Table 1: Number of data samples from each level.

texts in 10 categories including newspaper reportage, popular lore, and imaginative prose, written for different audiences and with varying writing styles. The annotations contain information about linguistic, structural, and functional information. In this research, we used a free-for-use bag-of-sentences version (*SUCX 3.0*) publicly available from *Språkbanken*[2]. Thus, no text features dependent on sentence order are included in our analysis. Furthermore, in order to minimise the effects of variations in language use that affect text complexity between genres (Štajner et al., 2012; Hiebert, 2012; Dell'Orletta et al., 2014), only texts from the category 'imaginative prose' were extracted, giving a total of 127 texts from *SUC*. This category was assumed to contain texts in a style the most similar to those extracted from *Nypon och Vilja*, being non-informational. The purpose of using *SUC* is to obtain a composition of data at a level of text complexity above all books from *Nypon och Vilja*. This is a key assumption and is based on the rationale that texts from *SUC* are written for typical adult readers and not with the express purpose of being especially easy to read, meaning text complexity can be assumed to be higher compared to the books from *Nypon och Vilja*.

To extract all necessary linguistic features, all texts were processed using the StilLett API Service (SAPIS) (Fahlborg and Rennes, 2016). The API service allows for the tokenization, lemmatisation, part-of-speech tagging, and dependency parsing of any text input. It also allows for text complexity analysis through the SCREAM module (Falkenjack et al., 2013) which computes related metrics.

[2]https://spraakbanken.gu.se/

6 Model implementation and evaluation

The Multi-Scale model was implemented using a modified version of the framework developed by Falkenjack (2018) and executed using R (version 3.6.3) with *RStudio* (RStudio Team, 2022). The model uses a set of covariates as input. These covariates are the values of metrics extracted through the data processing step resulting in a total of 47 features, c.f. (Falkenjack, 2018)

Data containing values for all covariates in the feature set were split into 5 classes for the *Nypon* scale and 6 classes for the *Vilja* scale ordered according to their levels. The data were then first split into training and test sets 500 times, using different ratios for training and test data, creating 500 models. The training data were used to estimate the full joint posterior distribution described in Section 4.2 through sampling according to the scheme described in Falkenjack (2018). This step was completed to evaluate the performance of the models. Then, instead of splitting the data into training and test sets, all data were used to run 20 chains of the Gibbs sampler resulting in a combined set of samples of a full posterior distribution of the entire dataset. The number of chains was based on the number of CPU cores available, using one core per chain to speed up the sampling process.

Furthermore, the Multi-Scale model by definition uses *multiple scales*, meaning the posterior distribution is sampled using data from both *Nypon* and *Vilja*. However, since the Multi-Scale Probit is a generalised version of an Ordered Probit model, which uses only one scale, its performance on either scale can be compared with a traditional Ordered Probit sampled using data from only that dataset. Such models used the same implementation as the Multi-Scale model, but using one scale at a time. These models will be referred to as *Single Scale* models.

After completing the sampling processes, parts of the resulting posterior distributions were visualised. The posterior was also used for classification and ranking, where the performance was evaluated by computing values for several evaluation metrics. For all evaluation metrics above, a higher positive value indicates better performance. For visualisation purposes and to enable easier comparison between model distributions, mode values are also plotted. The mode corresponds to the point with the highest probability density of a dis-

tribution, i.e. equivalent to the value that appears most frequently in a discrete probability distribution.

7 Model performance

The data was randomly separated into 500 different permutations of training and test data, and models were trained and evaluated for each such permutation. To assess whether one model generally outperforms the other we compute the difference between the posterior mean performance (F_1-scores and *Kendall's* τ_B correlations, respectively, as well as the harmonic mean of them) between the models for each permutation. Finally, to further examine how well the models perform given varying amounts of training data, two different split ratios were used: 2/3 training and 1/3 test data, and vice versa.

With 2/3 of the data used for training, the mean posterior modes of the performance metrics F_1 and *Kendall's* τ_B can be seen in Table 2.

	$F_1^{(M)}$	$F_1^{(S)}$	$\tau_B^{(M)}$	$\tau_B^{(S)}$
Nypon	0.35	0.35	0.45	0.45
Vilja	0.27	0.4	0.25	0.27

Table 2: Model performance using 2/3 of the data for training.

We can see that the choice of model makes little difference to the performance on the *Nypon* dataset but has a noticeable impact for the *Vilja* dataset.

Direct comparison of the models is done by computing the difference of the posterior mean F_1 and *Kendall's* τ_B for each model over the 500 data permutations. This shows that the Multi-Scale model outperforms the Single Scale model with respect to the F_1-score 54.4% of the time on the *Nypon* dataset and 73.4% of the time on the *Vilja* dataset. The same comparison of *Kendall's* τ_B show that the Multi-Scale model is better 62% of the time on the *Nypon* dataset and 89.2% of the time on the *Vilja* dataset. Figure 1 plots the distribution of differences in the harmonic mean of F_1-score and *Kendall's* τ_B between the models over all 500 data permutations, showing that the Multi-Scale model is better in 58.4% and 85.4% of cases for *Nypon* and *Vilja* respectively when both performance metrics are considered.

When the ratio of training to test data is reversed (i.e. 1/3 of the data used for training, the rest for testing), we see similar differences in overall performance on the *Vilja* dataset but now, the the difference in overall performance on the *Nypon* also shows a marked difference. Figure 2 illustrates this for the harmonic mean of F_1-score and *Kendall's* τ_B. However, as expected, the performance of both models is slightly lower with mean posterior modes, as seen in Table 3.

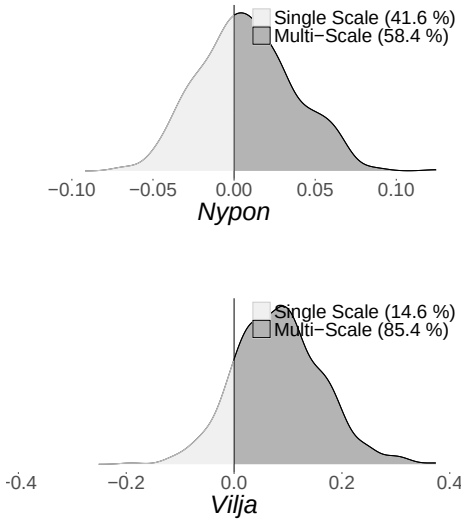

Figure 1: The posterior distribution for the difference in posterior harmonic mean of F_1-score and *Kendall's* τ_B between the Multi-Scale and Single Scale models. (2/3 of the data used for training.)

	$F_1^{(M)}$	$F_1^{(S)}$	$\tau_B^{(M)}$	$\tau_B^{(S)}$
Nypon	0.3	0.29	0.33	0.3
Vilja	0.25	0.24	0.3	0.24

Table 3: Model performance using 1/3 of the data for training.

This implies that the Multi-Scale model is especially useful when the availability of training data is limited.

Meanwhile, Figure 2 shows that the Multi-Scale model is better in 78.6% and 84.2% of cases for *Nypon* and *Vilja* respectively when both performance metrics are considered.

To summarise, the results show that the Multi-Scale model generally outperforms the Single Scale model on both datasets, particularly on the *Vilja* texts. Furthermore, this performance difference was greater when using a data split of 1/3 training data and 2/3 test data compared to a 2/3 training and 1/3 test data split. This implies that

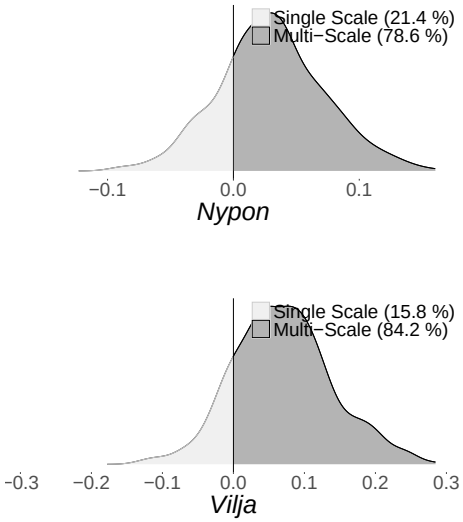

Figure 2: The posterior distribution for the difference in posterior harmonic mean of F_1-score and *Kendall's* τ_B between the Multi-Scale and Single Scale models. (1/3 of the data used for training.)

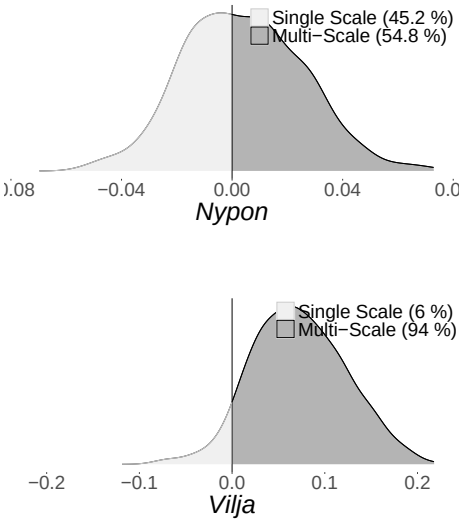

Figure 3: The posterior distribution for the difference in posterior harmonic mean of F_1-score and *Kendall's* τ_B between the Multi-Scale and Single Scale models. (2/3 of the *augmented* data used for training.)

the relative improvement of the Multi-Scale Probit over the single scale Ordered Probit decreases with the size of the dataset available for training a single scale model. In other words, the Multi-Scale Probit model is especially useful in contexts with sparse and diverse data for training.

7.1 Results of data augmentation

Augmenting both scales with the *SUC* corpus added an additional level to both scales above the other levels. The results of models using 500 augmented data sets with a data ratio of 2/3 training and 1/3 testing show that the modes of the posterior F_1-score and *Kendall's* τ_B show a marked improvement with an augmented dataset as shown in Table 4.

	$F_1^{(M)}$	$F_1^{(S)}$	$\tau_B^{(M)}$	$\tau_B^{(S)}$
Nypon	0.47	0.47	0.71	0.71
Vilja	0.4	0.32	0.7	0.6

Table 4: Model performance using 2/3 of the augmented data for training.

Figure 3 shows that when considering both F_1-score and *Kendall's* τ_B the Multi-Scale model outperforms the Single Scale model most of the time for both datasets.

Using 1/3 of the data for training and 2/3 for testing, as shown in Table 5, reinforces what we saw with the original data and with the 2/3 training ratio with augmented data. There is only a small improvement on the larger *Nypon* dataset but a more noticeable improvement on the smaller *Vilja* dataset.

	$F_1^{(M)}$	$F_1^{(S)}$	$\tau_B^{(M)}$	$\tau_B^{(S)}$
Nypon	0.4	0.39	0.68	0.67
Vilja	0.35	0.33	0.68	0.56

Table 5: Model performance using 1/3 of the augmented data for training.

Figure 4 again shows the Multi-Scale model outperforming the Single Scale model most of the time for both datasets.

The results show that, just as in the previous section, the Multi-Scale model generally outperforms the Single Scale model, particularly when tested on the smaller *Vilja* dataset, and that this improvement is greater when reducing the ratio of training data.

7.2 Results of regularisation

The regularisation process consisted of inspecting the marginal posteriors for each of the original 47 features, removing features with a positive or negative influence certainty % below specific thresholds (25%, 50% and 75%), and then resampling

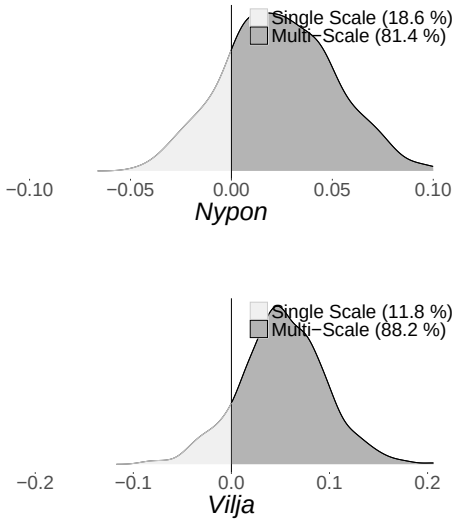

Figure 4: The posterior distribution for the difference in posterior harmonic mean of F_1-score and *Kendall's* τ_B between the Multi-Scale and Single Scale models. (1/3 of the *augmented* data used for training.)

the posterior distributions with each of the three reduced feature sets. The purpose of the regularisation process was to examine whether the change of feature set affects the predictive capabilities of the models.

Using 500 data sets with a ratio of 2/3 training data and 1/3 testing data, the modes of the posterior distribution of the harmonic mean align nearly perfectly for both models across all three feature sets when tested on the *Nypon* scale. On the *Vilja* scale, the Multi-Scale model slightly outperforms the Single Scale model across all feature sets. Furthermore, the modes for both models increase slightly between the first and second feature sets when tested on both scales. On the *Nypon* scale, there is an increase between the second and third feature sets, but no noticeable increase when tested on the *Vilja* scale.

The training/test split was again reversed (1/3 training, 2/3 test) on 500 reduced feature sets of data. The performance results of the models show that the modes of the posterior distribution of the harmonic mean are marginally higher for the Multi-Scale model compared to the Single Scale model across all feature sets when tested on the *Nypon* scale. On the *Vilja* scale, the difference in modes between the two models is greater, approximately 0.1 higher for the Multi-Scale model across all feature sets. Furthermore, the modes for both models increase slightly between the first and second feature set when tested on both scales, and a larger increase between the second and final feature set when tested on the *Nypon* scale, but not the *Vilja* scale.

8 Conclusion

The purpose of this research was to utilise the Multi-Scale Probit model in order to enable a standardised ranking and classification of text complexity, while also exploring how the model can be optimised. The assessment of text complexity can be used for a wide range of purposes, making its development pivotal in the field of natural language processing. The results from applying the Multi-Scale Probit on easy-to-read Swedish books have indicated that the model outperforms the Single Scale model in nearly all cases of classification and ranking, measured by F_1-scores and Kendall τ_B correlations. Furthermore, the results accentuate how the output from the Multi-Scale Probit model can be used in a simple manner to classify and rank new texts in the same domain, or adapted to other domains by creating new models. Through data augmentation and feature regularisation, the model can be optimised in terms of computational complexity and performance in specific contexts. The ability of the Multi-Scale Probit model to utilise data from different sources, without the necessity of large data quantities per category, enables assessments of text complexity that have previously not been possible. This research has contributed to the goal of developing methods for classifying and ranking text complexity, with the broader aim of creating a more accessible society for readers with varying needs.

Acknowledgments

This research is part of the project Text Adaptation for Increased Reading Comprehension, funded by The Swedish Research Council.

References

James H Albert and Siddhartha Chib. 1993. Bayesian Analysis of Binary and Polychotomous Response Data. *Source Journal of the American Statistical Association*, 88(422):669–679.

Alan Bailin and Ann Grafstein. 2001. The linguistic assumptions underlying readability formulae: A critique. *Language & communication*, 21(3):285–301.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in neural information processing systems*, 13.

Rebekah George Benjamin. 2012. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24:63–88.

CH Björnsson. 1968. Läsbarhet [readability] stockholm. *Sweden: Liber*.

C. I. Bliss. 1934a. The method of probits. *Science*, 79(2037):38–39.

Chester I. Bliss. 1934b. The Method of Probits. *Science*, 79(2037):38–39.

Mary Kathryn Cowles. 1996. Accelerating monte carlo markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing*, 6:101–111.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2014. Assessing document and sentence readability in less resourced languages and across textual genres. *ITL-International Journal of Applied Linguistics*, 165(2):163–193.

Eva Ejerhed, Gunnel Källgren, and Benny Brodda. 2006. Stockholm umeå corpus version 3.0.

Nick C. Ellis. 2020. Theoretical frameworks in l2 acquisition. In Patrick Rebuschat, editor, *The Cambridge Handbook of Language Learning*, chapter 4, pages 143–188. Cambridge University Press.

Daniel Fahlborg and Evelina Rennes. 2016. Introducing sapis-an api service for text analysis and simplification. In *The second national Swe-Clarin workshop: Research collaborations for the digital age, Umeå, Sweden*.

Johan Falkenjack. 2018. *Towards a model of general text complexity for swedish*. Ph.D. thesis, Linköping University Electronic Press.

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

Johan Falkenjack, Mattias Villani, and Arne Jönsson. 2018. Modeling text complexity using a multi-scale probit. *arXiv preprint arXiv:1811.04653*.

Jill Fitzgerald, Elfrieda H Hiebert, Kimberly Bowen, E Jackie Relyea-Kim, Melody Kung, and Jeff Elmore. 2015. Text complexity: Primary teachers' views. *Literacy Research and Instruction*, 54(1):19–44.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human language technologies 2007: the conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*, pages 460–467.

Elfrieda H Hiebert. 2012. Readability and the common core's staircase of text complexity. *Text Matters*, 1.

David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression*. John Wiley & Sons.

M. G. Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.

Robert E. McCulloch, Nicholas G. Polson, and Peter E. Rossi. 2000. A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics*, 99(1):173–193.

Danielle S McNamara. 2013. The epistemic stance between the author and reader: A driving force in the cohesion of text and writing. *Discourse Studies*, 15(5):579–595.

Katarina Heimann Mühlenbock and Sofie Johansson Kokkinakis. 2012. Swevoc-a swedish vocabulary resource for call. In *Proceedings of the SLTC 2012 workshop on NLP for CALL*, pages 28–34. Citeseer.

Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*. MIT press.

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.

RStudio Team. 2022. *RStudio: Integrated Development Environment for R*. RStudio, PBC, Boston, MA.

Marina Santini and Arne Jönsson. 2020. Pinning down text complexity: An exploratory study on the registers of the stockholm-umeå corpus (suc). *Register Studies*, 2(2):306–349.

Sarah E Schwarm and Mari Ostendorf. 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 (ACL'05)*, pages 523–530.

Sanja Štajner, Richard Evans, Constantin Orasan, and Ruslan Mitkov. 2012. What can readability measures really tell us about text complexity. In *Proceedings of workshop on natural language processing for improving textual accessibility*, pages 14–22. Citeseer.

Salvatore Valenti, Francesca Neri, and Alessandro Cucchiarelli. 2003. An overview of current research on automated essay grading. *Journal of Information Technology Education: Research*, 2(1):319–330.

Benjamin Vega, Shi Feng, Blair Lehman, Art Graesser, and Sidney D'Mello. 2013. Reading into the text: Investigating the influence of text complexity on cognitive engagement. In *Educational Data Mining 2013*.