A multivariate model for classifying texts' readability

Katarina Heimann Mühlenbock, Sofie Johansson Kokkinakis
Department of Swedish, University of Gothenburg, Sweden
katarina.heimann.muhlenbock@gu.se, sofie.johansson.kokkinakis@gu.se
Caroline Liberg, Åsa af Geijerstam, Jenny Wiksten Folkeryd
Department of Education, Uppsala University, Sweden
Arne Jönsson, Erik Kanebrant, Johan Falkenjack
Department of Computer and Information Science, Linköping University, Sweden

Abstract

We report on results from using the multivariate readability model SVIT to classify texts into various levels. We investigate how the language features integrated in the SVIT model can be transformed to values on known criteria like vocabulary, grammatical fluency and propositional knowledge. Such text criteria, sensitive to content, readability and genre in combination with the profile of a student's reading ability form the base of individually adapted texts. The procedure of levelling texts into different stages of complexity is presented along with results from the first cycle of tests conducted on 8th grade students. The results show that SVIT can be used to classify texts into different complexity levels.

1 Introduction

Standardized international tests demonstrate a continuous deterioration for Swedish 15-year-olds when it comes to knowledge in mathematics, reading and science (OECD, 2014). The task of finding adequate texts to fit the individual student's reading level is by necessity one of the most challenging and important tasks for teachers today. To facilitate this, tools for teachers are needed that allow individual reading comprehension testing, presenting a reading profile for each student and suggestions of texts suitable with regard to genre, age and reading level. Essential to this endeavour is the ability to measure text complexity; which is what the SVIT model (Heimann Mühlenbock, 2013) is designed to do.

In this paper we will start in Section 2 by presenting how text complexity is measured, how the texts were selected and levelled, and how the tests were carried out. In Section 3 we present the first results from a subset of the tests carried out in the first year of the project. The correspondence between the text levelling and the students' results will be discussed in Section 4, and a final discussion on to which extent the results from the tests actually agree with the automatic selection and levelling of texts will follow in Section 5.

2 Method

The first task was to find adequate reading materials for each of the three school grades. Automated readability assessment has long rested upon very simplified heuristics for text complexity. Most measurements contain a factor that relates to a text's average sentence length and another factor related to the average word length (Flesch, 1948; Gunning, 1952; Senter and Smith, 1967; McLaughlin, 1969; Coleman and Liau, 1975; Kincaid et al., 1975). The underlying assumption is that the sentence length to some extent is correlated with the syntactic complexity and that the word length reflects the semantic load of a text. We used a more sophisticated method based on the SVIT model (Heimann Mühlenbock, 2013) for grading a bank of texts into appropriate levels. The first cycle of tests was devoted to investigating narrative texts, while texts concerning civics and science will follow in future studies.

After text selection, reading comprehension tests of narrative texts suiting students in the 4th, 6th and 8th grades in 74 Swedish schools were carried out on more than 4000 students. The schools were situated in three major Swedish municipalities, one for each grade. All the tests were given anonymously, and only the teacher was able to see the results from the individual students.

2.1 SVIT

Quantitative measures of readability are appealing since they are easy to perform computationally. The obvious drawback of measuring text phenomena at the surface level is that the results

are purely descriptive and not interpretive. This is why readability researchers long struggled to find an easy and cost efficient way to devise a link between the quantitative textual properties and the qualitative characteristics. Eventually, the most widely known and used methods for readability measurement built upon formulas containing both a semantic and a syntactic variable. The semantic component is expected to be expressed in the vocabulary use, and more precisely in the texts average word length. The syntactic properties are accordingly anticipated to be found in the sentence structure through calculation of the average sentence length. The Swedish LIX Readability Index is based on these principles (Björnsson, 1968).

A multilevel theoretical readability framework considers additional levels of language and discourse. Chall (1958) proposed four different elements to be significant for a readability criterion; namely vocabulary load, sentence structure, idea density and human interest. Graesser et al (2011) distinguish five levels, including words, syntax, text base, situation mode,l and genre and rhetorical structure. The two theories are consistent in that they, in addition to vocabulary and syntactical properties, also consider the message intended to be conveyed in a text, through analysis of the idea density or text base. The genre structure refers to the category of text, while the situation model is assumed to capture the subject matter content of the text and inferences that are activated by the explicit text. Finally, the human interest level proposed by Chall is evidently strongly tied to the readers experiences and thus the least prone to external inspection.

The SVIT language model (Heimann Mühlenbock, 2013) includes a combination of properties at the surface, vocabulary, syntactical and idea density levels. The surface level measurement includes simple word and sentence length counts, but also measures of extra-long words (>13 characters), and token iteration. At the vocabulary level we find the vocabulary properties analysed in terms of word lemma variation and the proportion of words belonging to a Swedish base vocabulary (Heimann Mühlenbock and Johansson Kokkinakis, 2012). The syntactic level is inspected through measurements of mean distance between items in the syntactically parsed trees, mean parse tree heights, and the proportions of subordinate clauses and nominal modifiers. The idea density is supposed to be revealed through calculations of average number of propositions, nominal ratio and noun/pronoun ratio. Finally, it is assumed that the personal interest to some extent might be captured through the proportion of proper nouns in a text. We will focus on the results achieved for the most prominent features, i.e. those who were expected to be least mutually correlated. Some of the features listed in Table 1 are quite straightforward, while others need an explanation.

The *Lemma variation index* is calculated with the formula:

$$LVIX = \frac{\log(N)}{\log(2 - \frac{\log(U)}{\log(N)})}$$

where N = Number of lemmas and U = Number of unique lemmas

Words considered as *Difficult words* are those not present in category (C), (D), or (H) in the SweVoc base vocabulary. In category (C) we find 2,200 word lemmas belonging to the core vocabulary. Category (D) contains word lemmas referring to everyday objects and actions. Category (H), finally, holds word lemmas highly frequent in written text. In all, 4,700 word lemmas are included in these categories. The values in Table 2 refer to the mean percentage of the lemmas complementary to the mentioned categories.

The syntactic features *MDD*, *UA*, *AT*, *ET* and *PT* refer to properties in the dependency parsed sentences in the texts.

The *Nominal ratio* is achieved by calculating the proportion of nouns, prepositions and participles in relation to verbs, pronouns and adverbs.

2.2 Text selection method

In all, 22 texts from the LäSBarT corpus (Heimann Mühlenbock, 2013) and 31 texts from a bank of National reading tests were checked with the goal of finding suitable portions of narrative texts for the intended group of readers in the 4th, 6th and 8th Swedish school grades, which corresponds to students aged from 10 to 15 years.

2.3 Levelling texts

After the first manual check, the texts were graded into 7 levels of difficulty after multivariate analysis based on the SVIT model. The texts were classified manually after inspection of the SVIT values.

Level	Feature	Abbrev
Surface	Mean sentence length	MSL
	Mean word length in characters	MWL
Vocabulary	Lemma variation index	LVIX
	Difficult words	DW
Syntactic	Mean dependency distance	MDD
	Subordinate clauses	UA
	Prenominal modifier	AT
	Postnominal modifier	ET
	Parse tree height	PT
Idea density	Nominal ratio	NR

Table 1: SVIT features

Earlier experiments showed that for the task of differentiating between ordinary and easy-to-read childrens fiction, language features at the vocabulary and idea density levels were found to have the highest impact on the discriminative power. The features mirroring vocabulary diversity and difficulty, and idea density were given precedence when the metrics did not unambiguously point towards significant differences at the syntactic level.

For the students who attended 8th grade, six texts were selected based on the SVIT-values, ranging between 527 and 1166 words in length. The texts were then split into two groups (Group 1 and Group 2) with similar internal distribution between easier and more difficult texts. We will here present here the properties of the hardest and the easiest of the six texts, both present in Group 1. The two texts were not of equal length, but the students were allowed to read the texts and answer the questions at their own pace.

The easiest text (Text 1) was a short story retrieved from the collection of National reading tests, entitled *Att fiska med morfar* 'Fishing with Grandpa' by Ulf Stark. The most difficult text (Text 2) was also picked from the National reading tests. It is entitled *Populärmusik från Vittula* 'Popular music from Vittula', written by Mikael Niemi. Some of their respective values are shown in Table 2.

Based on the SVIT-values, we can derive the following information about the two texts:

Text 1 shows low average values regarding word and sentence lengths. At the vocabulary level, the word lemma variation is at medium level for the six texts. The syntactic complexity is very low, both regarding prenominal modifiers and parse tree height. The idea density level is below average.

Text 2 is slightly above average regarding word length. The word lemma variation index and percentage of difficult words are both considerably above average. The syntactic complexity is slightly above average for all features. Finally, the idea density is above average.

2.4 Testing method

Items testing two overall reading processes were constructed for each of the three texts (Langer, 2011; Luke and Freebody, 1999; Mullis et al., 2009; OECD, 2009):

1. Retrieve explicitly stated information and make straightforward inferences

2. Interpret and integrate ideas and information and reflect on, examine and evaluate content, language, and textual elements

In assessing the vocabulary knowledge of students, we focused on the receptive knowledge of subject and domain neutral content words in Swedish novels as test items. We used a reliable approach to create vocabulary tests similar to the Vocabulary Levels Test (VLT) (Nation, 2001). The test items were extracted from a frequency based vocabulary list of compiled corpora representing each level of text difficulty. The test items were presented in context in an authentic sentence taken from the text book. Three alternative meanings of the test item and one correct meaning were presented to the student. The alternative meanings (distractors) were similar to the test item regarding phonology or orthography.

332 students in the 8th grade in 5 schools in Gothenburg participated. The teachers were instructed to allow the students to read the texts and answer the questions at their own pace, which usually corresponded to a total time of one lesson per test. The tests were administered as paper questionnaires and the texts were read on paper.

3 Results from students' testings

The results presented here concern 94 students who performed all three tests on texts in Group 1, which included the two texts with SVIT-values exemplified in Table 2, deemed as the easiest and

Text	Surface features		Vocabulary features		Syntactical features					Idea density
	MSL	MWL	LVIX	DW	MDD	UA	AT	ET	PT	NR
Text 1	9.7	4.2	54.1	23.0	2.2	0.3	0.1	0.3	4.9	0.44
Text 2	11.3	4.8	69.8	33.0	2.1	0.3	0.3	0.4	5.5	0.70

Table 2: SVIT values

the most difficult. For each text, the students answered 12 questions regarding the content and 15 questions regarding the understanding of isolated words not present in the texts.

3.1 Text reading results

On average, students' performance on Text 1 was 74.2% correctness on content questions, and 68.3% on Text 2. These results indicate, that the texts were well adapted for the age group, but also that Text 2 was perceived as more difficult by the normally performing students.

For the low-performing students, the correlation between the results, i.e. correctness on content questions, and text complexity was even more obvious. 26 students scored <1 S.D. of the students' results. On average, they had 55.2% correctness on content questions for Text 1, and 37.0% correctness for Text 2. Furthermore, the 10 students who presented results <2 S.D. below normal scored on average 48.2% and 36.7%, respectively.

3.2 Vocabulary results

As could be expected, we found that there was a strong correlation between reading comprehension and vocabulary knowledge. The correlation was significant, 0.68 resp. 0.63 with Pearson's correlation at level 0.01.

4 Correspondence between readers and texts

When looking at the overall reading processes, it was found that among the 26 low-performing students, 12 showed reading profiles indicating that they were only able to correctly answer a few of the text-based and the interpretive and evaluative questions. They were assumed to work only on individual parts of the text and were most likely not able to grasp the big picture in the texts and how different aspects of the text were related. They all performed better on Text 1 than Text 2. Five students were found to perform pretty well on the text-based questions but not as good when it came to the interpretive and evaluative questions on Text 1. None of these students were able to correctly answer more than a few of the text-based and the interpretive and evaluative questions on Text 2. For the remaining low-performing students the results were more mixed. Four of them had a reading profile which implied rather good results on the text-based, interpretive and evaluative questions on Text 1, but did not grasp the content of Text 2. Finally, 5 students performed somewhat better on Text 2 than on Text 1, but were still not able to entirely comprehend the content of any of the two texts.

5 Discussion

We have presented an approach that investigates the extent to which automated text levelling with a multivariate analysis based on the SVIT language model really proved to correspond to students' actual reading performance. We found that the participating students performed better on the text judged as the easiest as opposed to the most difficult, with a mean difference in correctness of 5.9%. Furthermore, the low-performing students showed a significant difference in correctness of 18.2% between the two texts. These findings support the hypothesis that the SVIT readability model based on language features derived computationally, and present at deeper levels than the purely superficial, can devise a link between quantitative and qualitative text characteristics. Further studies investigating the efficiency on levelling texts from other genres than fiction will follow.

References

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

Jeanne S. Chall. 1958. *Readability: An appraisal of research and application.* Ohio State University Press, Columbus, OH.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284.

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

Arthur C. Graesser, Danielle S. McNamara, and Jonna M. Kulikowich. 2011. Coh-metrix. *Educational Researcher*, 40(5):223–234.

Robert Gunning. 1952. *The technique of clear writing*. McGraw-Hill International Book Co., New York, NY.

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, Lund, October. Linköping University Electronic Press.

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

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

Judith A. Langer. 2011. *Envisioning Knowledge. Building Literacy in the Academic Disciplines.* New York: Teachers' College Press.

Allan Luke and Peter Freebody. 1999. Further notes on the four resources model. Reading Online. http://www.readingonline.org/research/lukefreebody.html.

G. Harry McLaughlin. 1969. SMOG grading – a New Readability Formula. *Journal of Reading*, 12(8):639–646.

Ina V.S. Mullis, Michael O. Martin, Ann M. Kennedy, Kathleen L. Trong, and Marian Sainsbury. 2009. *PIRLS 2011 Assessment Framework.* PIRLS 2011 Assessment Framework.

Paul Nation. 2001. *Learning vocabulary in another language*. Cambridge University Press.

OECD. 2009. Pisa 2009 Assessment Framework. Key Competencies in reading, mathematics and science. Paris: OECD.

OECD. 2014. Pisa 2012. Results in Focus. What 15-year-olds know and what they can do with what they know. Paris: OECD.

R.J. Senter and E. A. Smith. 1967. Automated readability index. Technical report, Cincinnati Univ. Ohio, Cincinnati Univ. Ohio.