Studies on automatic assessment of students' reading ability

Katarina Heimann Mühlenbock[2], Erik Kanebrant[1], Sofie Johansson Kokkinakis[2]
Arne Jönsson[1], Caroline Liberg[3], Åsa af Geijerstam[3], Johan Falkenjack[1]
Jenny Wiksten Folkeryd[3]

Department of Computer and Information Science, Linköping University[1],
Department of Swedish, Göteborg University[2], Department of Education, Uppsala University[3]
katarina.heimann.muhlenbock@gu.se, arnjo@ida.liu.se

Abstract

We report results from ongoing research on developing sophisticated measures for assessing a student's reading ability and a tool for the student and teacher to create a profile of this ability. In the project we will also investigate how these measures can be transformed to values on known criteria like vocabulary, grammatical fluency and so forth, and how these can be used to analyse texts. Such text criteria, sensitive to content, readability and genre in combination with the profile of a student's reading ability will form the base to individually adapted texts. Techniques and tools will be developed for selecting suitable texts, automatic summarisation of texts and automatic transformation to easy-to-read Swedish.

1. Introduction

It is shown in different studies that even a not so strong reader is able to read in a more advanced way if the text is adapted with respect to aspects such as the topic of the text and different linguistic features, e.g. Liberg (2010), Reichenberg (2000). Our focus is to support reading for ten to fifteen year old students. The means for this is to find appropriate texts that are individually suitable and adapted to each student's reading abilities.

2. Models of reading

Common to models of reading in an individual-psychological perspective is that reading consists of two components: comprehension and decoding, e.g. (Adams, 1990). Traditionally, the focus has been on decoding aspects, but in later years research with a focus on comprehension has increased rapidly.

The test of students' reading ability in this study will include, in accordance with a broad view, different text types of different degrees of linguistic difficulty, where the students are tested for various reading practices within different topic areas. Items testing the following reading practices will be constructed for each of these texts, cf. Mullis et al. (2009, p. 23–29), OECD (2009, p. 34–44):

1. Retrieve explicitly stated information and make straightforward inferences (cf. text-meaning practices of Luke and Freebody (1999) and first envisionment of Langer (2011)),
2. interpret and integrate ideas and information (cf. Luke's and Freebody's text-meaning practices and Langer's other envisonments), and
3. reflect on, examine and evaluate content, language, and textual elements (cf. Luke's and Freebody's pragmatic and critical practices).

Each of these practices also includes testing different aspects of vocabulary knowledge, c.f. Laufer and Nation (1995). These tests comprises everyday words originating from the same corpus as the readability texts.

3. Readability measures

We will consider global language measures built upon lexical, morpho-syntactic and syntactic features of a given text. The general readability of a text relates, however, not only to a combination of language properties making it easy or hard to grasp, but also to the specific reader (Mühlenbock and Johansson Kokkinakis, 2009). We will use the SVIT measures (Heimann Mühlenbock, 2013) which consider linguistic features appearing at the surface in terms of raw text, but also at deeper language levels. For the latter task we automatically process the text in four different steps: pre-processing, part-of-speech and lemma information annotation and finally parsing with dependency annotation.

The SVIT classification comprises four levels of linguistic complexity. The first level includes surface properties such as average word and sentence length and word token variation calculated with the OVIX formula. At the second level vocabulary properties are taken into account by measures of word lemma variation and the proportion of words belonging to a Swedish base vocabulary (Heimann Mühlenbock and Johansson Kokkinakis, 2012). The third, syntactic, level includes measuring the mean distance between items in the dependency trees, mean parse tree heights, the proportions of subordinate clauses, and pre- and post nominal modifiers. Finally, at the fourth level we considered the idea density present in the texts calculated in terms of average number of propositions, nominal ratio and noun/pronoun ratio.

A multivariate analysis revealed that for the task of discriminating between ordinary and easy-to-read childrens' fiction, feature values at the vocabulary and idea density levels had the highest impact upon the positive results in automatic classification (Heimann Mühlenbock, 2013). For the present purpose, we therefore gave values indicating higher vocabulary diversity and difficulty metrics priority when the results did not unambiguously demonstrated any difference between features at the syntactic level.

4. Tools and techniques

The final tool that is to be developed will:

1. Select texts and vocabulary appropriate to estimated student level in different text genres.
2. Conduct test, using texts from 1, to establish texts to be recommended as ”start text” for students of a certain age within the age span. A students reading profile will consist of a range of texts and vocabulary tests from a ”base text” read with high scores to a ”top text” read with low scores. In most cases the ”start text” will be the same as the ”base text”.
3. Use information from 2 in order to automatically select student-adapted texts for the subject area.
4. Simplify texts using summarization and transformation to easy-to-read Swedish for texts in the subject area.

The first iteration of this tool includes a tool for teachers where they can view individual students' test results.

5. Results from the first test series

We have developed a first series of reading tests with texts and questions measuring reading ability and vocabulary knowledge. The tests comprise fiction texts and are expected to match three different levels of reading proficiency in the 4th, 6th and 8th school grades respectively. (The same test is used for the hardest grade i test and the easiest grade $i + 1$ test giving seven levels in total.) The texts vary in length between 450 and 1166 words, and were selected based on the SVIT measures. The tests were carried out in a total of 74 schools and more than 4000 students. Each student did a series of three tests, with texts and vocabulary on three levels of difficulty. The tests were conducted in the grade order 6, 4 and 8. We will briefly present current findings, all statistical analyses are not yet finished.

In general, students perform better on simpler texts than on more difficult. For the test conducted in grade 6 many students acquired top scores and therefore two of the texts from grade 6 were also used for grade 4 providing a stronger correlation between the students' results and the text's difficulty. We saw an even stronger correlation between text complexity, indicated by SVIT, and response rates of the weakest students, i.e. those whose overall test results were < 2 SD below the average. This observation held for all three school levels. Given that the SVIT measures were used as benchmark in the initial levelling phase, we believe that our findings strongly support the hypothesis that these measures are able to grade a text's complexity and hence readability.

There is a statistically significant correlation between the students' results on the vocabulary tests and the reading tests for all seven levels. This shows that the tools and theories used to develop the tests are applicable. Note that the vocabulary test comprises domain neutral every day words from the same corpus as the readability texts. The purpose is to assess a general vocabulary competence.

A tool for teachers has been developed and distributed to all teachers with students that did the tests (Kanebrant, 2014). It allows teachers to get results on reading ability for each individual student. The tool is password protected to ensure that results only can be accessed by the teacher. The response texts intend to describe the readability competencies and vocabulary knowledge assessed. For the tests on reading ability we decided to group the categories 2 and 3 ending up with the two categories: 1) Retrieve explicitly stated information and make straightforward inferences and 2) Interpret and integrate ideas and information and reflect on, examine and evaluate content, language, and textual elements. We believe that it is easier for teachers to comprehend the results that way.

References

M.J. Adams. 1990. *Beginning to Read. Thinking and Learning about Print.* Cambridge, Massachusets & London, England: MIT Press.

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. 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.

Erik Kanebrant. 2014. Automaster: Design, implementation och utvärdering av ett läroverktyg. Bachelors thesis, Linköping University.

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

B. Laufer and P. Nation. 1995. Vocabulary size and use: Lexical richness in l2 written production. *Applied Linguistics*, 16:307–322.

Caroline Liberg. 2010. Texters, textuppgifters och undervisningens betydelse för elevers läsförståelse. fördjupad analys av pirls 2006. Report, Skolverket.

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

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.

I.V.S. Mullis, M.O. Martin, A.M. Kennedy, K.L. Trong, and M. Sainsbury. 2009. *PIRLS 2011 Assessment Framework.* PIRLS 2011 Assessment Framework.

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

M. Reichenberg. 2000. *Röst och kausalitet i lärobokstexter: en studie av elevers förståelse av olika textversioner.* Ph.D. thesis, Acta Universitatis Gothoburgensis, Göteborg.